Deinterleaving Finite Memory Processes via Penalized Maximum Likelihood

Gadiel Seroussi, *Fellow, IEEE,* Wojciech Szpankowski, *Fellow, IEEE,* and Marcelo J. Weinberger *Fellow, IEEE*

***Abstract*—We study the problem of deinterleaving a set of finite-memory (Markov) processes over disjoint finite alphabets, which have been randomly interleaved by a finite-memory switch. The deinterleaver has access to a sample of the resulting interleaved process, but no knowledge of the number or structure of the component Markov processes, or of the switch. We study conditions for uniqueness of the interleaved representation of a process, showing that certain switch configurations, as well as memoryless component processes, can cause ambiguities in the representation. We show that a deinterleaving scheme based on minimizing a penalized maximum-likelihood cost function is strongly consistent, in the sense of reconstructing, almost surely as the observed sequence length tends to infinity, a set of component and switch Markov processes compatible with the original interleaved process. Furthermore, under certain conditions on the structure of the switch (including the special case of a memoryless switch), we show that the scheme recovers *all* possible interleaved representations of the original process. Experimental results are presented demonstrating that the proposed scheme performs well in practice, even for relatively short input samples.**

***Index terms*— Markov process, finite memory process, interleaved Markov process, FSM source, penalized maximum likelihood**

I. Introduction

Problems in applications such as data mining, computer forensics, finance, and genomics, often require the identification of streams of data from different sources, which may be intermingled or hidden (sometimes purposely) among other unrelated streams, in large interleaved record files. In this haystack of records can lie buried valuable information whose extraction would be easier if we were able to separate the contributing streams. The deinterleaving problem studied in this paper is motivated by these applications (more detailed accounts of which can be found, for example, in [**?**], [**?**], [**?**]).

In our setting, the data streams, as well as the interleaving agent, will be modeled as sequences generated by discrete-time random processes over finite alphabets. Specifically, let $A_1, A_2, \ldots, A_m$ be finite, nonempty, disjoint alphabets, let $\mathcal{A} = A_1 \cup A_2 \cup \cdots A_m$, and $\Pi = \{A_1, A_2, \ldots, A_m\}$. We refer to the A_i as *subalphabets*, and to Π as a *partition*, of $\mathcal{A}$. Consider m independent, *component random processes* $P_1, P_2, \ldots, P_m$, defined, respectively, over $A_1, A_2, \ldots, A_m$, and a random *switch process* $P_{\rm w}$ over the alphabet Π, independent of the component processes. The *interleaved process* $P \triangleq \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$ is generated as follows: At each time instant, a subalphabet $A_i \in \Pi$ is selected according to $P_{\rm w}$, and the next output sample for P is selected from A_i according to the corresponding process P_i (we say, loosely, that the switch "selects" P_i at that instant). The component processes P_i are idle when not selected, i.e., if P_i is selected at time t, and next selected at time $t+T$, then the samples emitted by P at times t and $t+T$ are *consecutive* emissions from P_i, regardless of the length of the intervening interval T.

Given a sample z^n from P, and without prior knowledge of the number or the composition of the subalphabets A_i, the *deinterleaving* problem of interest is to reconstruct the original sequences emitted by the component processes, and the sequence of switch selections.

So far, we have made two basic assumptions on the structure of the interleaved system: the independence of the component and switch processes, and the disjointness of the subalphabets. The latter assumption implies that, given an interleaved input stream, identifying the partition Π is equivalent to identifying the component substreams and the sequence of switch selections. Thus, identifying the partition Π is sufficient to solve the deinterleaving problem. Identifying the substreams when the subalphabets are not disjoint is also a problem of interest, but it appears more challenging [**?**], and is outside the scope of this paper. Even with these assumptions, it is clear that without further restrictions on the component and switch processes, the problem defined would be either ill-posed or trivial, since two obvious hypotheses would always be available: the interleaved process P could be interpreted as having a single component $P_1 = P$, or as an interleaving of constant processes over singleton alphabets interleaved by a switch $P_{\rm w}$ essentially identical to P. Therefore, for the problem to be meaningful, some additional constraints must be posed on the structure of the component and switch processes.

In this paper, we study the case where the components and switch are ergodic *finite memory* (Markov) processes, i.e., for each $i \in \{1, 2, \ldots, m, {\rm w}\}$, there is an integer $k_i \geq 0$ such that for any sufficiently long sequence u^t over the appropriate alphabet, we have $P_i(u_t|u^{t-1}) = P_i(u_t|u^{t-1}_{t-k_i})$. We refer to P in this case as an *interleaved Markov process* (IMP), and to $\mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$ as an *IMP representation* of P

W. Szpankowski's work was partially done while visiting HP Labs, Palo Alto, CA, and also supported by NSF Science and Technology Center Grants CCF-0939370 and CCF-0830140. Parts of this paper were presented at the 2009 IEEE International Symposium on Information Theory and the 2011 IEEE International Symposium on Information Theory.

G. Seroussi is with Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA (e-mail:gseroussi@ieee.org).

W. Szpankowski is with the Department of Computer Science, Purdue University, IN 47907, USA (e-mail: spa@cs.purdue.edu).

M. J. Weinberger is with Hewlett-Packard Laboratories, Palo Alto, CA 94304, USA (e-mail:marcelo.weinberger@hp.com).

(we also say that Π is *compatible* with P). We assume no knowledge or bound on the process orders k_i. Except for some degenerate cases (e.g., when all the component processes are memoryless), the IMP P is generally *not* a finite memory process, since the interval between consecutive selections of a component process is unbounded. Hence, in general, the two obvious hypotheses mentioned above are not available, and the deinterleaving problem for IMPs is well-posed, non-trivial, and, as we shall show, solvable.

The main result of the paper is a deinterleaving scheme that, when presented with a sample from an IMP P, identifies, eventually almost surely, an IMP representation of P. The solution is based on finding a partition Π of A and an *order vector* $\mathbf{k} = (k_1, k_2, \ldots, k_m, k_{\mathrm{w}})$ that minimize a *penalized maximum-likelihood* (penalized ML) cost function of the form

$$C_{\Pi,\mathbf{k}}(z^n) = n\hat{H}_{\Pi,\mathbf{k}}(z^n) + \beta\kappa\log n\,,$$

where $\hat{H}_{\Pi,\mathbf{k}}(z^n)$ is the empirical entropy of the observed sequence z^n under an IMP model induced by Π and $\mathbf{k}$, κ is the total number of free statistical parameters in the model, and β is a nonnegative constant. Penalized ML estimators of Markov process order are well known (cf. [?], [?], [?]). Here, we use them to estimate the partition Π, and also the Markov order of the processes P_i and the switch P_{w}.

Notice that, given an IMP P, any partition Π' of $\mathcal{A}$ induces a set of deinterleaved component and switch processes P'_i, P'_w. If Π' is the "wrong" partition (i.e., it is incompatible with P), then either some of the P'_i or P'_{w} will not be of finite order, or some of the independence assumptions will be violated. There could, however, be more than one "right" partition: IMP representations need not be unique, and we may have partitions $\Pi{\neq}\Pi'$ such that both Π and Π' are compatible with P. This situation is referred to as an *ambiguity* in the IMP representation of P.[1] We study these ambiguities, and identify two phenomena that may lead to them: the presence of memoryless component processes, and a so-called *alphabet domination* phenomenon which may arise from certain transition probabilities in the switch being set to zero (and which, therefore, does not arise in the case of memoryless switches). We derive a set of sufficient conditions for uniqueness, whose validity ensures that the IMP representation identified by our deinterleaving scheme is, almost surely, identical to the one that generated the process. We also show that in cases where ambiguities are due solely to memoryless components (the so-called *domination-free* case, which includes all cases with memoryless switches), our scheme yields *all* the IMP representations of P.

The deinterleaving problem for the special case where all processes involved are of order at most one has been previously studied in [?], where an approach was proposed that could identify an IMP representation of P with high probability as $n{\to}\infty$ (the approach as described cannot identify multiple solutions when they exist; instead, all cases leading to possible ambiguities are excluded using rather coarse conditions). The idea is to run a greedy sequence of tests, checking equalities and inequalities between various event probabilities (e.g., $P(ab){\neq}P(a)P(b)$, $P(abc) = P(a)P(b)P(c)$, $a, b, c \in \mathcal{A}$), and permanently clustering symbols into subalphabets sequentially, according to the test results (sequentiality here is with respect to the alphabet processing, not the input sequence, which has to be read in full before clustering begins). Empirical distributions are used as proxies for the true ones. Clearly, equalities between probabilities translate only to "approximate equalities" subject to statistical fluctuations in the corresponding empirical quantities, and an appropriate choice of the tolerances used to determine equality, as functions of the input length n, is crucial to turn the conceptual scheme into an effective algorithm. Specific choices for tolerances are not discussed in [?]. The attractive feature of the approach in [?] is its low complexity; equipped with a reasonable choice of tolerance thresholds, an efficient algorithm for the special case of processes of order one can be implemented. However, as we shall see in the sequel, the convergence of the algorithm is rather slow in practice, and very long samples are necessary to achieve good deinterleaving performance, compared to the schemes proposed here. The problem of deinterleaving hidden-Markov processes was also studied, mostly experimentally, in [?]. Another variant of the problem, where all the component processes are assumed to be identical (over the same alphabet), of order one, and interleaved by a memoryless switch, was studied in [?].

We note that IMPs are a special case of the broader class of *switching discrete sources* studied in [?], with variants dating back as early as [?]. However, the emphasis in [?] is on universally compressing the output of a switched source of known structure, and not on the problem studied here, which is precisely to identify the source's structure.

The rest of the paper is organized as follows. In Section II we present some additional definitions and notation, and give a more formal and detailed definition of an IMP, which will be useful in the subsequent derivations. We also show that an IMP can be represented as a unifilar *finite-state machine* (FSM) source (see, e.g., [?]), whose parameters satisfy certain constraints induced by the IMP structure. In Section III we study ambiguities and conditions for uniqueness of an IMP representation. Most of the derivations and proofs for the results of this section are presented in Appendix **??**. In Section IV we present our deinterleaving scheme, establish its strong consistency, and show that in the domination-free case, it can identify all valid IMP representations of the interleaved process. The derivations and proofs for these results are presented in Appendix **??**. Finally, in Section V we present some experimental results for practical implementations of deinterleaving schemes. We compare the performance of our scheme with that of an implementation of the scheme of [?] (with optimized tolerances) for the case of IMPs with memoryless switches, showing that the ML-based deinterleaver achieves high accuracy rates in identifying the correct alphabet partition for much shorter sequences than those required by the scheme of [?]. Our ideal scheme calls for finding the optimal partition through an exhaustive search, which is computationally expensive. Consequently, we show results for a randomized gradient descent heuristic that searches for the same optimal

[1] Since P and Π uniquely determine the component and switch processes, two different IMP representations of the same process P *must* be based on different partitions.

partition. Although in principle this approach sacrifices the optimality guarantees of the ideal scheme, in practice, we obtain the same results as with exhaustive search, but with a much faster and practical scheme. We also present results for IMPs with switches of order one. We show, again, that the ML-based schemes exhibit high deinterleaving success rates for sequences as short as a few hundred symbols long, and perfect deinterleaving, for the samples tested, for sequences a few thousand symbols long.

II. Preliminaries

A. Definitions

All Markov processes are assumed to be time-homogeneous and ergodic, and, consequently, to define limiting stationary distributions [**?**]. We denote the (minimal) order of P_i by $k_i \overset{\Delta}{=} \operatorname{ord}(P_i)$, refer to reachable strings u^{k_i} as *states* of P_i, and denote the set of such states by $\mathcal{S}(P_i)$, $i \in \{1, 2, \ldots, m, \mathrm{w}\}$. Some conditional probabilities may be zero, and some k_i-tuples may be non-reachable, but all states are assumed to be reachable and recurrent. We further assume that all symbols $a \in \mathcal{A}$ (and subalphabets $A \in \Pi$) occur infinitely often, and their stationary marginal probabilities are positive. We make no assumptions on the initial conditions of each process, and, in our characterization of ambiguities, we distinguish processes only up to their stationary distributions, i.e., we write $P = P'$ if and only if P and P' admit the same stationary distribution. Aside from simplifying some notations, this notion of process equivalence makes our results on uniqueness of IMP representations slightly stronger than if we had adopted a stricter notion (e.g., actual process identity). All probability expressions related to stochastic processes will be interpreted as (sometimes marginal) stationary probabilities, e.g., $P_i(u)$, or $P_i(a|u) = P_i(ua)/P_i(u)$ when u is not long enough to define a state of P_i.

A string over Π is referred to as a *switch sequence*. In particular, any string $u^t = u_1 u_2 \ldots u_t \in \mathcal{A}^t$ defines a switch sequence $\mathbf{A}_\Pi(u^t) = A_{i_1} A_{i_2} \ldots A_{i_t} \in \Pi^t$ with i_j being the unique index such that $u_j \in A_{i_j}$, $1 \le j \le t$. Also, for $\mathcal{A}' \subseteq \mathcal{A}$, and a string u over $\mathcal{A}$, we let $u[\mathcal{A}']$ denote the string over $\mathcal{A}'$ obtained by deleting from u all symbols that are not in $\mathcal{A}'$.

The IMP $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_\mathrm{w})$ is now formally defined as follows: Given $z^t \in \mathcal{A}^t$, $t \ge 1$, and assuming $z_t \in A_i$, we have

$$P(z_t|z^{t-1}) = P_\mathrm{w}(A_i|\mathbf{A}_\Pi(z^{t-1}))P_i(z_t|z^{t-1}[A_i])\,. \qquad (1)$$

It is readily verified that (1) completely defines the process P, which inherits whatever initial conditions hold for the component and switch processes, so that (1) holds for any conditioning string z^{t-1}, $t \ge 1$ (including $z^{t-1} = \lambda$). Also, by recursive application of (1), after rearranging factors, we obtain, for any sequence $z^n \in \mathcal{A}^n$,

$$P(z^n) = P_\mathrm{w}(\mathbf{A}_\Pi(z^n)) \prod_{i=1}^{m} P_i(z^n[A_i])\,. \qquad (2)$$

When initial conditions are such that the probabilities on the right-hand side of (2) are stationary, the equation defines a stationary distribution for P. (We adopt the convention that $P_i(\lambda) = 1$, $i \in \{1, 2, \ldots, m, \mathrm{w}\}$, and, consequently, $P(\lambda) = 1$.)

For conciseness, in the sequel, we will sometimes omit the arguments from the notations $\mathcal{I}_\Pi$ or $\mathcal{I}_{\Pi'}$, assuming that the respective sets of associated subalphabets and processes (resp. $\{A_i\}$, $\{P_i\}$ or $\{A'_i\}$, $\{P'_i\}$) are clear from the context. For IMP representations $\mathcal{I}_\Pi$ and $\mathcal{I}_{\Pi'}$, we write $\mathcal{I}_\Pi \equiv \mathcal{I}_{\Pi'}$ if the representations are identical, i.e., $\Pi = \Pi'$ and $P_i = P'_i$, $i \in \{1, 2, \ldots, m, \mathrm{w}\}$ (in contrast with the relation $\mathcal{I}_\Pi = \mathcal{I}_{\Pi'}$, which is interpreted to mean that $\mathcal{I}_\Pi$ and $\mathcal{I}_{\Pi'}$ generate the same process).

We will generally denote sequences (or strings) over $\mathcal{A}$ with lower case letters, e.g., $u \in \mathcal{A}^*$, and (switch) sequences over Π with upper case letters, e.g., $U \in \Pi^*$. We say that $u^n \in \mathcal{A}^n$ and $U^n \in \Pi^n$ are *consistent* if $P(u^n) > 0$ and $U^n = \mathbf{A}_\Pi(u^n)$. Clearly, for every sequence u^n with $P(u^n) > 0$ there exists a sequence $U^n = \mathbf{A}_\Pi(u^n)$, with $P_\mathrm{w}(U^n) > 0$, that is consistent with u^n; conversely, if $P_\mathrm{w}(U^n) > 0$, it is straightforward to construct sequences u^n consistent with U^n. Unless specified otherwise, we assume that an upper case-denoted switch sequence is consistent with the corresponding lower case-denoted string, e.g., when we write $UV = \mathbf{A}_\Pi(uv)$, we also imply that $U = \mathbf{A}_\Pi(u)$ and $V = \mathbf{A}_\Pi(v)$.

B. IMPs and FSM sources

A *finite state machine* (FSM) over an alphabet $\mathcal{A}$ is defined by a triplet $F = (S, s_0, f)$, where S is a set of *states*, $s_0 \in S$ is a (possibly random) *initial state*, and $f : S \times \mathcal{A} \to S$ is a *next-state function* (see, e.g., [**?**]). A (unifilar) *FSM source*, in turn, is defined by associating a conditional probability distribution $P_F(\cdot|s)$ with each state s of F, and a probability distribution $P_F^{\mathrm{init}}(\cdot)$ on the initial state s_0. To generate a random sequence x^n, the source draws s_0 according to $P_F^{\mathrm{init}}(\cdot)$ and then draws, for each i, $1 \le i \le n$, a symbol $x_i \in \mathcal{A}$ distributed according to $P_F(\cdot|s_{i-1})$, and transitions to the state $s_i = f(s_{i-1}, x_i)$. Markov sources of order k over $\mathcal{A}$ are special cases of FSM sources with $S \subseteq \mathcal{A}^k$. For convenience, we will assume that FSM sources have arbitrary but *fixed* initial states. In particular, we will assume that a fixed initial state $s_0^{(j)} \in \mathcal{S}(P_j)$ is defined for the component/switch processes P_j, $j \in \{1, 2, \ldots, m, \mathrm{w}\}$, where we recall that $\mathcal{S}(P_j)$ denotes the state set of P_j. The results are easily generalized to arbitrary initial state conditions, since any initial state distribution can be written as a convex combination of fixed initial state conditions.

We next observe that an IMP is an FSM source. We refer to the vector $\mathbf{k} = (k_1, k_2, \ldots, k_m, k_\mathrm{w})$, where $k_j = \operatorname{ord}(P_j)$, $j \in \{1, 2, \ldots, m, \mathrm{w}\}$, as the *order vector* of the IMP $\mathcal{I}_\Pi$. Denote by f_j the next-state function of the FSM associated with P_j, $j \in \{1, 2, \ldots, m, \mathrm{w}\}$, and consider the FSM $\mathcal{F}_{\Pi,\mathbf{k}} = (S, \mathbf{s}_0, f)$, with state set $S = \mathcal{S}(P_1) \times \mathcal{S}(P_2) \times \cdots \mathcal{S}(P_m) \times \mathcal{S}(P_\mathrm{w})$, initial state $\mathbf{s}_0 = (s_0^{(1)}, s_0^{(2)}, \ldots, s_0^{(m)}, s_0^{(\mathrm{w})}) \in S$, and next-state function f defined as follows:

Given a state $\mathbf{s} = (s^{(1)}, s^{(2)}, \ldots, s^{(m)}, s^{(\mathrm{w})}) \in S$, and $a \in \mathcal{A}$ such that $\mathbf{A}_\Pi(a) = A_i$, we have $f(\mathbf{s}, a) = \mathbf{s}' = (s'^{(1)}, s'^{(2)}, \ldots, s'^{(m)}, s'^{(\mathrm{w})})$ where $s'^{(j)} = s^{(j)}$ for $j \in \{1, 2, \ldots, m\} \setminus \{i\}$, $s'^{(i)} = f_i(s^{(i)}, a)$, and $s'^{(\mathrm{w})} = f_\mathrm{w}(s^{(\mathrm{w})}, A_i)$.

To complete the definition of an FSM source, we define, for each state $\mathbf{s} \in S$ and symbol $a \in A$, with $\mathbf{A}_\Pi(a) = A_i \in \Pi$, the conditional probability distribution

$$P_{\Pi,\mathbf{k}}(a \,|\, \mathbf{s}) = P_{\mathrm{w}}(A_i | s^{(\mathrm{w})}) P_i(a \,|\, s^{(i)}) \,. \tag{3}$$

The following proposition is readily verified.

Proposition 1: $\mathcal{F}_{\Pi,\mathbf{k}}$, with transition probabilities $P_{\Pi,\mathbf{k}}$, generates $P{=}\mathcal{I}_\Pi(P_1, P_2, \ldots, P_m, P_{\mathrm{w}})$.

Results analogous to Proposition 1 for switching discrete sources are given in [**?**]. The class of finite state sources considered in [**?**], however, is broader, as unifilarity is not assumed.

It follows from the ergodicity and independence assumptions for IMP components and switch that P is an ergodic FSM source, and every state $\mathbf{s} \in S$ has a positive stationary probability. Let $\alpha_i = |A_i|$, $1 \le i \le m$, and $\alpha = |\mathcal{A}| = \sum_{i=1}^m \alpha_i$. By the definition of the state set S, we have $|S| \le m^{k_{\mathrm{w}}} \prod_{i=1}^m \alpha_i^{k_i}$ (equality holding when all k_j-tuples over the appropriate alphabet are reachable states of P_j, $j \in \{1, 2, \ldots, m, \mathrm{w}\}$). Hence, the class of *arbitrary* FSM sources over $\mathcal{A}$, with underlying FSM $\mathcal{F}_{\Pi,\mathbf{k}}$, would have, in general, up to

$$\mathcal{K}(\Pi, \mathbf{k}) = (\alpha - 1) m^{k_{\mathrm{w}}} \prod_{i=1}^m \alpha_i^{k_i} \tag{4}$$

free statistical parameters. The conditional probability distributions in (3), however, are highly constrained, as the parameters $P_{\Pi,\mathbf{k}}(a|\mathbf{s})$ satisfy relations of the form

$$P_{\mathrm{w}}(A_i | s'^{(\mathrm{w})}) P_{\Pi,\mathbf{k}}(a|\mathbf{s}) = P_{\mathrm{w}}(A_i | s^{(\mathrm{w})}) P_{\Pi,\mathbf{k}}(a|\mathbf{s}'),$$

where $A_i = \mathbf{A}_\Pi(a)$, for all states $\mathbf{s}'$ such that $s^{(i)} = s'^{(i)}$. In particular, it follows directly from (3) that $P_{\Pi,\mathbf{k}}(a|\mathbf{s}) = P_{\Pi,\mathbf{k}}(a|\mathbf{s}')$ if $s^{(i)} = s'^{(i)}$ and $s^{(\mathrm{w})} = s'^{(\mathrm{w})}$. Overall, the number of free parameters remains, of course, that of the original component Markov processes and switch, i.e., up to

$$\kappa(\Pi, \mathbf{k}) = \sum_{i=1}^m \alpha_i^{k_i} (\alpha_i - 1) + (m-1) m^{k_{\mathrm{w}}}, \tag{5}$$

which is generally (much) smaller than $\mathcal{K}(\Pi, \mathbf{k})$.

We refer to an FSM source satisfying the constraints implicit in (3) as an *IMP-constrained* FSM source.

III. Uniqueness of IMP representations

In this section, we study conditions under which the IMP representation of a process is unique, and, for IMPs that are free from certain "pathologies" that will be discussed in the sequel, characterize all IMP representations of a process when multiple ones exist. Notice that although, as shown in Section II, IMPs can be represented as constrained FSM sources, the study of ambiguities of IMP representations differs from the problem of characterizing different FSM representations of a source [**?**], or more generally of representations of hidden Markov processes [**?**]. It is known [**?**] that all FSMs that can generate a given FSM source P are *refinements*[2] of a so-called *minimal FSM representation* of the source. In particular, this applies to the FSM corresponding to any IMP representation. However, the minimal FSM representation is not required to satisfy the IMP constraints, so it needs not coincide with a minimal (or unique) IMP representation. Notice also that, when defining IMPs and their FSM representations, we have assumed that the orders k_i of all the Markov processes involved are minimal, thus excluding obvious FSM refinements resulting from refining some of the individual Markov processes.

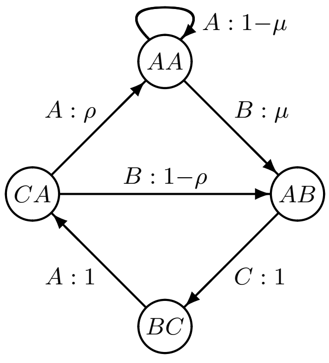

Fig. 1. A switch P_{w} of order two over $\Pi = \{A, B, C\}$. Arcs are labeled $X : \xi$, where X is the emitted symbol and ξ the corresponding transition probability. Transitions not drawn are assumed to have probability zero.

A. Alphabet domination

Let A, B be arbitrary subalphabets in Π. We say that A *dominates* B (relative to P_{w}) if there exists a positive integer M such that if P_{w} has emitted M occurrences of B without emitting one of A, then *with probability one* P_{w} will emit an occurrence of A before it emits another occurrence of B. In other words, if $P_{\mathrm{w}}(U) > 0$, then $U[\{A, B\}]$ does not contain any run of more than M consecutive occurrences of B. We denote the domination relation of A over B as $A \sqsupset B$, dependence on P_{w} being understood from the context; when A does not dominate B, we write $A \not\sqsupset B$ (thus, for example, $A \not\sqsupset A$). We say that A is *dominant* (in Π, relative to P_{w}) if either $m = 1$ (i.e., $\Pi = \{A\}$) or $A \sqsupset B$ for some $B \in \Pi$, and that A is *totally dominant* if either $m = 1$ or $A \sqsupset B$ for *all* $B \in \Pi \setminus \{A\}$. If $A \sqsupset B$ and $B \sqsupset A$, we say that A and B are in *mutual domination.* It is readily verified that domination is an irreflexive transitive relation. When no two subalphabets are in mutual domination, the relation defines a strict partial order (see, e.g., [**?**]) on the finite set Π. We shall make use of the properties of this strict partial order in the sequel.

Domination can occur only if some transition probabilities in P_{w} are zero; therefore, it never occurs when P_{w} is memoryless. The approach for a switch of order one in [**?**] assumes

[2] A refinement [**?**] of an FSM $F = (S, s_0, f)$ is an FSM $F^+ = (S^+, s_0^+, f^+)$ such that for some fixed function $g : S^+ \to S$ and any sequence x^n, the respective state sequences $\{s_i\}$ and $\{s_i^+\}$ satisfy $s_i = g(s_i^+)$, $0 \le i \le n$ (for example, the FSM underlying a Markov process of order $k+1$ is a refinement of the FSM underlying one of order k). By suitable choices of conditional probabilities, a refinement of F can generate any process that F can generate.

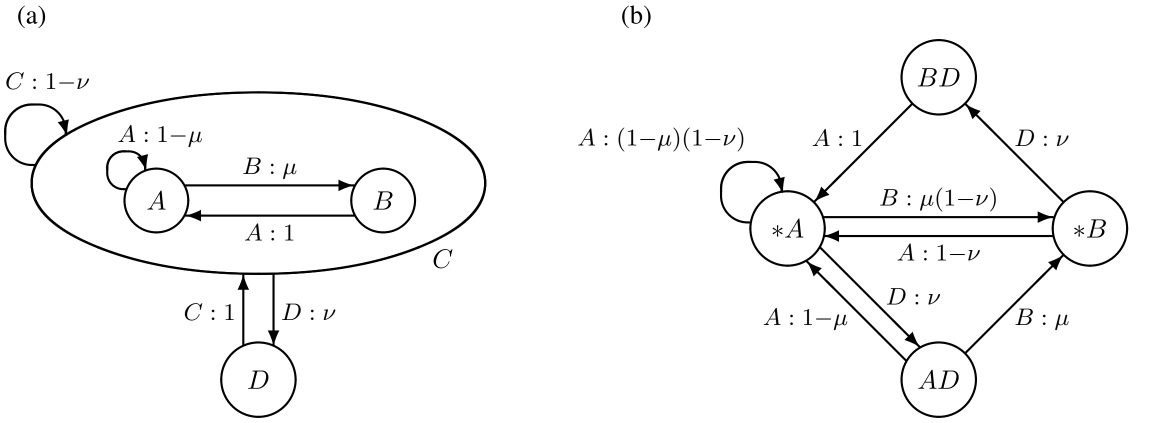

Fig. 2. Switches for ambiguous IMP representation: (a) $P_{\rm w}$ over $\{C, D\}$, $\mathrm{ord}(P_{\rm w}) = 1$ ($C = A \cup B$, and the internal structure of P_C is also shown), (b) $P'_{\rm w}$ over $\{A, B, D\}$, $\mathrm{ord}(P'_{\rm w}) = 2$. Arcs are labeled with their corresponding emitted symbols and transition probabilities; transitions not shown have probability zero.

that $P_{\rm w}(A|A) > 0$ for all $A \in \Pi$. Clearly, this precludes alphabet domination. However, the condition is too stringent to do so, or as a condition for uniqueness.

Example 1: Consider an IMP $P = \mathcal{I}_\Pi(P_1, P_2, P_3; P_{\rm w})$ with $\Pi = \{A, B, C\}$, and $P_{\rm w}$ as defined by Figure 1, where $\mathrm{ord}(P_{\rm w}) = 2$, and transitions are labeled with their respective emitted symbols and probabilities. We assume that $\mu \in (0, 1]$ and $\rho \in (0, 1)$. For this switch, we have $A \sqsupset B$, $A \sqsupset C$, $B \sqsupset C$ and $C \sqsupset B$. Subalphabet A is totally dominant, and, if $\mu < 1$, it is not dominated. Subalphabets B and C are in mutual domination for all $\mu \in (0, 1]$; if $\mu = 1$, every pair of subalphabets is in mutual domination. In all cases, $P_{\rm w}$ is aperiodic.

B. Conditions for uniqueness

We derive sufficient conditions for the uniqueness of IMP representations, and show how ambiguities may arise when the conditions are not satisfied. The main result of the subsection is given in the following theorem, whose derivation and proof are deferred to Appendix **??**.

Theorem 1: Consider an IMP $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$ satisfying the following conditions:

i) no two subalphabets in Π are in mutual domination,
ii) no subalphabet in Π is totally dominant, and
iii) none of the processes P_i is memoryless.

Then, if $P = \mathcal{I}_{\Pi'}(P'_1, P'_2, \ldots, P_{m'}; P'_{\rm w})$ for some partition Π' and finite memory processes $P'_1, P'_2, \ldots, P_{m'}, P'_{\rm w}$, we must have $\mathcal{I}_\Pi \equiv \mathcal{I}_{\Pi'}$.

Example 2: We consider alphabets A, B, D, and $C = A \cup B$, and respective associated processes P_A, P_B, P_D, P_C. Part (a) of Fig. 2 shows a switch $P_{\rm w}$ of order 1 over $\Pi = \{C, D\}$. Here, P_C is in itself an interleaved process $P_C = \mathcal{I}_{\{A,B\}}(P_A, P_B; P^C_{\rm w})$ with P_B chosen as a memoryless process so that P_C has finite memory (specifically, $\mathrm{ord}(P_C) \leq 2\,\mathrm{ord}(P_A)$); P_D is not memoryless, and we have $\nu, \mu \in (0, 1)$. Part (b) shows a switch $P'_{\rm w}$ of order two over $\Pi' = \{A, B, D\}$. The node labeled $*A$ (resp. $*B$) represents all the states that end in A (resp. B). It is readily verified that $P = \mathcal{I}_\Pi(P_C, P_D; P_{\rm w}) = \mathcal{I}_{\Pi'}(P_A, P_B, P_D; P'_{\rm w})$, so P is an ambiguous IMP. It is also readily verified that both $\mathcal{I}_\Pi$ and $\mathcal{I}_{\Pi'}$ violate Condition (ii) of Theorem 1: C is totally dominant in $\mathcal{I}_\Pi$, and A is totally dominant in $\mathcal{I}_{\Pi'}$. In fact, the figure exemplifies a more detailed variant of Theorem 1, presented as Theorem 2 below, which characterizes ambiguities when Condition (ii) of the original theorem is removed.

Given partitions Π and Π' of $\mathcal{A}$, we say that $A_i \in \Pi$ *splits* in Π' if A_i is partitioned into subalphabets in Π', i.e. $A'_j \subseteq A_i$ for all $A'_j \in \Pi'$ such that $A'_j \cap A_i \neq \phi$.

Theorem 2: Consider an IMP $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\rm w})$ such that no two subalphabets are in mutual domination, and none of the processes P_i is memoryless. Then, if $P = \mathcal{I}_{\Pi'}(P'_1, P'_2, \ldots, P'_{m'}; P'_{\rm w})$ for some partition Π', we must have $A_i \in \Pi'$ for all subalphabets $A_i \in \Pi$ except possibly for one subalphabet $A_{i_0} \in \Pi$, which must be totally dominant and split in Π'.

The proof of Theorem 2 is also deferred to Appendix **??**. The theorem covers the special case of $m = 1$, which is excluded by Condition (ii) in Theorem 1. In this case, the IMP is actually a finite-memory process, which admits the two “obvious” IMP representations (with $m = 1$ and $m = |\mathcal{A}| = |A_1|$, respectively) mentioned in the introduction.

C. Ambiguities due to memoryless components in the domination-free case

In this subsection, we eliminate Condition (iii) of Theorem 1, while strengthening Conditions (i) and (ii) by excluding all forms of alphabet domination. We characterize all the representations of an IMP when ambiguities, if any, are due solely to memoryless components.

We say that a partition Π' is a *refinement* of Π if every subalphabet $A_i \in \Pi$ splits in Π'. When Π' is a refinement of Π, we denote by $\Psi_{\Pi,\Pi'} : \Pi' \to \Pi$ the function mapping a subalphabet $A'_j \in \Pi'$ to the subalphabet $A_i \in \Pi$ that contains it. The notation and map extend in the natural way to arbitrary strings, namely $\Psi_{\Pi,\Pi'} : (\Pi')^k \to \Pi^k$ for all $k \geq 0$. We will omit the indices Π, Π' from Ψ when clear from the context.

Lemma 1: Consider a partition $\Pi = \{A_1, A_2, \ldots, A_m\}$, together with a refinement $\Pi' = \{B_1, B_2, A_2, \ldots, A_m\}$ of Π

(i.e., $A_1 = B_1 \cup B_2$). Let $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_{\mathrm{w}})$, where P_1 is memoryless, and let $P' = \mathcal{I}_{\Pi'}(P_1^{(1)}, P_1^{(2)}, P_2, \ldots, P_m; P'_{\mathrm{w}})$, where both $P_1^{(1)}$ and $P_1^{(2)}$ are memoryless. Then, $P = P'$ if and only if the following conditions hold:

$$P_1^{(j)}(b) = \frac{P_1(b)}{P_1(B_j)}, \;\; b \in B_j\,, \quad j \in \{1, 2\}, \tag{6}$$

$$\mathcal{S}(P'_{\mathrm{w}}) = \{\, S' \in (\Pi')^{k_{\mathrm{w}}} \,\big|\, \Psi(S') \in \mathcal{S}(P_{\mathrm{w}})\}, \tag{7}$$

and for all $A \in \Pi'$ and $S' \in \mathcal{S}(P'_{\mathrm{w}})$, with $S = \Psi(S')$,

$$P'_{\mathrm{w}}(A|S') = \begin{cases} P_{\mathrm{w}}(A|S), & A = A_i,\, i \geq 2, \\ P_{\mathrm{w}}(A_1|S)P_1(B_j), & A = B_j,\, j = 1, 2\,. \end{cases} \tag{8}$$

Remark. The proof of Lemma 1 is deferred to Appendix **??**. The lemma is interpreted as follows: since, given $\mathcal{I}_\Pi$, processes $P_1^{(1)}, P_1^{(2)}$, and P'_{w} can always be defined to satisfy (6)–(8), an IMP P with a nontrivial memoryless component always admits alternative representations where the alphabet associated with the memoryless process has been split into disjoint parts (the split may be into more than two parts, if the lemma is applied repeatedly). We refer to such representations as *memoryless refinements* of the original representation $\mathcal{I}_\Pi$. Using the lemma repeatedly, we conclude that P admits a refinement where all the memoryless components are defined over singleton alphabets. On the other hand, the memoryless components $P_1^{(1)}$ and $P_1^{(2)}$ of P' can be merged if and only if P'_{w} satisfies the constraint

$$P'_{\mathrm{w}}(B_2|S') = \gamma P'_{\mathrm{w}}(B_1|S') \tag{9}$$

for a constant γ independent of $S' \in \mathcal{S}(P'_{\mathrm{w}})$. Indeed, when (9) holds, we set $P_1(B_1) = 1/(1+\gamma)$ and $P_1(B_2) = \gamma/(1+\gamma)$, and P_1, P_{w} are defined implicitly by (6)–(8). Notice that the constraint (9) is trivially satisfied when the switch P'_{w} is memoryless (and so is also the resulting P_{w}). Thus, in this case, memoryless component processes can be split or merged arbitrarily to produce alternative IMP representations. When the switch has memory, splitting is always possible, but merging is conditioned on (9). We refer to a representation where no more mergers of memoryless processes are possible, as well as to the corresponding partition Π, as *canonical* (clearly, the canonicity of Π is relative to the given IMP).[3]

We denote the canonical representation associated with an IMP $P = \mathcal{I}_\Pi$ by $(\mathcal{I}_\Pi)^*$, and the corresponding canonical partition by $(\Pi)^*_P$. Also, we say that P is *domination-free* if there is no alphabet domination in *any* IMP representation of P. The main result of the subsection is given in the theorem below, whose proof is presented in Appendix **??**.

Theorem 3: Let $P = \mathcal{I}_\Pi$ and $P' = \mathcal{I}_{\Pi'}$ be domination-free IMPs over $\mathcal{A}$. Then, $P = P'$ if and only if $(\mathcal{I}_\Pi)^* \equiv (\mathcal{I}_{\Pi'})^*$.

Theorem 3 implies that, in the domination-free case, all the IMP representations of a process are those constructible by sequences of the splits and mergers allowed by Lemma 1. In particular, this always applies to the case of memoryless switches, where domination does not arise.

Corollary 1: Let $P = \mathcal{I}_\Pi$ and $P' = \mathcal{I}_{\Pi'}$ be IMPs over $\mathcal{A}$, where the switches P_{w} and P'_{w} are memoryless. Then, $P = P'$ if and only if $(\mathcal{I}_\Pi)^* \equiv (\mathcal{I}_{\Pi'})^*$.

IV. The deinterleaving scheme

Given any finite alphabet A, a sequence $u^t \in A^t$, and a nonnegative integer k, denote by $\hat{H}_k(u^t)$ the kth order (unnormalized) empirical entropy of u^t, namely, $\hat{H}_k(u^t) = -\log \hat{P}_k(u^t)$, where $\hat{P}_k(u^t)$ is the ML (or empirical) probability of u^t under a kth order Markov model with a fixed initial state. Let z^n be a sequence over $\mathcal{A}$. An arbitrary partition Π of $\mathcal{A}$ naturally defines a deinterleaving of z^n into sub-sequences $\mathbf{z}_i = z^n[A_i]$, $\;1 \leq i \leq m$, with a switch sequence $\mathbf{Z}_{\mathrm{w}} = \mathbf{A}_\Pi(z^n)$. Given, additionally, an order vector $\mathbf{k} = (k_1, k_2, \ldots, k_m, k_{\mathrm{w}})$, we define

$$\hat{H}_{\Pi,\mathbf{k}}(z^n) = \sum_{i=1}^{m} \hat{H}_{k_i}(\mathbf{z}_i) + \hat{H}_{k_{\mathrm{w}}}(\mathbf{Z}_{\mathrm{w}})\,.$$

This quantity can be regarded as the (unnormalized) empirical entropy of z^n with respect to $F = \mathcal{F}_{\Pi,\mathbf{k}}$ *for an IMP-constrained* FSM source (as discussed in Subsection II-B). Indeed, let $\hat{P}_{\Pi,\mathbf{k}}(z^n)$ denote the ML probability of z^n with respect to F under IMP constraints, i.e., denoting by $\mathcal{P}_\mathcal{I}(\mathcal{F}_{\Pi,\mathbf{k}})$ the class of all IMPs generated by F (i.e., all FSM sources based on F with parameter vectors satisfying the IMP constraints), we have

$$\hat{P}_{\Pi,\mathbf{k}}(z^n) = \max_{P \in \mathcal{P}_\mathcal{I}(\mathcal{F}_{\Pi,\mathbf{k}})} P(z^n)\,. \tag{10}$$

Clearly, by (2), $\hat{P}_{\Pi,\mathbf{k}}(z^n)$ is obtained by maximizing, independently, the probabilities of the component and switch sequences derived from z^n, and, thus, we have $\hat{H}_{\Pi,\mathbf{k}}(z^n) = -\log \hat{P}_{\Pi,\mathbf{k}}(z^n)$. Notice that $\hat{P}_{\Pi,\mathbf{k}}(z^n)$ is generally different from (and upper-bounded by) the ML probability with respect to F for an *unconstrained* FSM source; this ML probability will be denoted $\hat{P}^*_F(z^n)$. The *penalized cost* of z^n relative to Π and $\mathbf{k}$ is defined as

$$C_{\Pi,\mathbf{k}}(z^n) \;\;=\;\; \hat{H}_{\Pi,\mathbf{k}}(z^n) + \beta\kappa\log(n+1)\,, \tag{11}$$

where $\kappa = \kappa(\Pi, \mathbf{k})$, as given in (5), is the number of free statistical parameters in a generic IMP-constrained FSM source based on F, and β is a nonnegative (penalization) constant.[4]

Given a sample z^n from an IMP P, our deinterleaving scheme estimates a partition $\hat{\Pi}(z^n)$, and an order vector $\hat{\mathbf{k}}(z^n)$, for the estimated IMP representation of P. The desired estimates are obtained by the following rule:

$$\left(\hat{\Pi}(z^n), \hat{\mathbf{k}}(z^n)\right) = \arg\min_{(\Pi',\mathbf{k}')} C_{\Pi',\mathbf{k}'}(z^n), \tag{12}$$

where $(\Pi', \mathbf{k}')$ ranges over all pairs of partitions of $\mathcal{A}$ and order vectors $\mathbf{k}'$. In the minimization, if $C_{\Pi',\mathbf{k}'}(z^n) =$

[3] The special case of this result for IMPs with memoryless switches discussed in [**?**] uses a slightly different definition of canonicity.

[4] For convenience, we set the penalty terms in (11) all proportional to $\log(n+1)$, rather than the term corresponding to $\mathbf{z}_i$ being proportional to $\log|\mathbf{z}_i|$. Given our basic assumptions on switch processes, if z^n is a sample from an IMP, $|\mathbf{z}_i|$ will, almost surely, be proportional to n. Therefore, the simpler definition adopted has no effect on the main asymptotic results. Clearly, using $\log(n+1)$ in lieu of $\log n$, which will be convenient in some derivations, is also of negligible effect.

$C_{\Pi'',\mathbf{k}''}(z^n)$, for different pairs $(\Pi',\mathbf{k}')$ and $(\Pi'',\mathbf{k}'')$, the tie is broken first in favor of the partition with the smallest number of alphabets. Notice that although the search space in (12) is defined as a Cartesian product, once a partition Π' is chosen, the optimal process orders k'_j are determined independently for each $j \in \{1,2,\ldots,m,\mathrm{w}\}$, in a conventional penalized ML Markov order estimation procedure (see, e.g., [**?**]). Also, it is easy to verify that the optimal orders $\hat{k}_j$ must be $O(\log n)$, reducing the search space for $\mathbf{k}'$ in (12).

Our main result is given by the following theorem, whose derivation and proof are presented in Appendix **??**. Recall that $(\Pi)^*_P$ denotes the canonical partition of P (Subsection III-C).

Theorem 4: Let $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_\mathrm{w})$, and let z^n be a sample from P. Then, for suitable choices of the penalization constant β, $\hat{\Pi}(z^n)$ is compatible with P, and $\hat{\mathbf{k}}(z^n)$ reproduces the order vector of the corresponding IMP representation $\mathcal{I}_{\hat{\Pi}}$, almost surely as $n \to \infty$. Furthermore, if P is domination-free, we have

$$\hat{\Pi}(z^n) = (\Pi)^*_P \quad \text{a.s. as } \ n \to \infty\,.$$

Remarks. 1) Theorem 4 states that our scheme, when presented with a sample from an interleaved process, will almost surely recover an alphabet partition compatible with the process. If the interleaved process is domination-free, the scheme will recover the canonical partition of the process, from which *all* compatible partitions can be generated via repeated applications of Lemma 1. The difficulty in establishing the first claim of the theorem resides in the size of the class of models that participate in the optimization (12). The fact that a compatible partition will prevail over any specific incompatible one eventually almost surely, for any penalization coefficient $\beta \ge 0$, will be readily established through a large deviations argument. However, the class contains models whose size is not bounded with n. In fact, it is well known (see, e.g., [**?**]) that the stationary distribution of the ergodic process P can be approximated arbitrarily (in the entropy sense) by finite memory processes of unbounded order. Thus, without appropriately penalizing the model size, a sequence of "single stream" hypotheses of unbounded order can get arbitrarily close in cost to the partitions compatible with P. We will prove that an appropriate positive value of β suffices to rule out these large models that asymptotically approach P. To establish the second claim of the theorem, we will take advantage of the observation that the canonical representation of a domination-free IMP, is, in a sense, also the most "economical". Indeed, comparing the number of free statistical parameters in the two IMP representations considered in Lemma 1, we obtain, using (5),

$$\kappa(\Pi',\mathbf{k}') - \kappa(\Pi,\mathbf{k}) = m(m+1)^{k_\mathrm{w}} - (m-1)m^{k_\mathrm{w}} - 1\,. \quad (13)$$

It is readily verified that the expression on the right hand side of (13) vanishes for $k_\mathrm{w} = 0$, and is strictly positive when $k_\mathrm{w} > 0$ (since $m \ge 1$). Therefore, splitting a memoryless component as allowed by Lemma 1, in general, can only increase the number of parameters. Thus, the canonical partition minimizes the model size, and with an appropriate choice of $\beta > 0$, our penalized ML scheme will correctly identify this minimal model.

2) If a bound is known on the orders of the component and switch processes, then it will follow from the proof in Appendix **??** that the first claim of Theorem 4 can be established with any $\beta \ge 0$. However, an appropriate positive value of β is still needed, even in this case, to recover the canonical partition in the second claim of the theorem. Our deinterleaving scheme assumes that IMPs based on $\mathcal{F}_{\Pi,\mathbf{k}}$ are fully parametrized, i.e., the class has κ free statistical parameters. If the actual IMP being estimated is less than fully parametrized (i.e., it does have some transition probabilities set to zero), the effect of penalizing with the full κ is equivalent to that of using a larger penalization coefficient β.

V. Experimental results

We report on experiments showing the performance of practical implementations of the proposed deinterleaver. The experiments were based on test sets consisting of 200 interleaved sequences each. Each sequence was generated by an IMP with m=3, subalphabet sizes α_1=4, α_2=5, α_3=6, component Markov processes of order $k_i \le 1$ with randomly chosen parameters, and a switch of order $k_\mathrm{w} \le 1$ as described below. In all cases, the switches were domination-free. Deinterleaving experiments were run on prefixes of various lengths of each sequence, and, for each prefix length, the fraction of sequences correctly deinterleaved was recorded.

In the first set of experiments, the component Markov processes, all of order one, were interleaved by uniformly distributed memoryless switches (i.e., $\mathbf{k} = (1,1,1,0)$). We compared the deinterleaving performance of the ML-based scheme proposed here with that of an implementation of the scheme of [**?**], with tolerances for the latter optimized (with knowledge of the correct partition) to obtain the best performance for each sequence length. Two variants of the ML-based scheme were tested: Variant (a) implements (12) via exhaustive search over all partitions.[5] Since this is rather slow, a heuristic Variant (b) was developed, based on a randomized gradient descent-like search. This variant, which is briefly described next, is much faster, and achieves virtually the same deinterleaving performance as the full search.

The *neighborhood* of radius t of a partition Π, denoted $\mathcal{N}_t(\Pi)$, is the set of all partitions Π' obtained from Π by switching up to t symbols of $\mathcal{A}$ from their original subalphabets in Π to other subalphabets (including possibly new subalphabets not present in Π). The main component of the heuristic starts from an input sequence z^n and a random partition Π_0 of $\mathcal{A}$, and exhaustively searches for the partition Π' that minimizes the cost $C_{\Pi'}(z^n)$ within the neighborhood $\mathcal{N}_t(\Pi_0)$, for some small fixed value of t. The minimizing partition then becomes the center for a new exhaustive neighborhood search. This "greedy" deterministic process continues until no improvements in the cost function can be obtained. At this point, the best partition Π observed so far is perturbed by picking a random partition $\Pi'_0 \in \mathcal{N}_r(\Pi)$, for a fixed radius

[5] We recall that given a sequence z^n and a partition Π, the order vector $\mathbf{k}$ minimizing the cost $C_{\Pi,\mathbf{k}}(z^n)$ is determined through conventional penalized-ML order estimators for the various sub-sequences induced by Π. We assume that this minimizing order vector is used in all cost computations, and omit further mention of it.

TABLE I

FRACTION OF CORRECTLY DEINTERLEAVED SEQUENCES (OUT OF 200) VS. SEQUENCE LENGTH, FOR TWO VARIANTS OF THE PROPOSED SCHEME (ML(a) AND ML(b)), AND FOR THE SCHEME OF [?]. A PENALIZATION CONSTANT $\beta = \frac{1}{2}$ WAS USED IN ALL CASES FOR THE ML-BASED SCHEMES.

	memoryless switch			switch with memory			
	$\mathbf{k} = (1,1,1,0)$			$\mathbf{k} = (1,1,1,1)$		$\mathbf{k} = (0,1,1,1)$	
n	ML (a)	ML (b)	[?]	ML (a)	ML (b)	ML (b) canonical	ML (b) compatible
250	0.010	0.010	0.000	0.310	0.300	0.215	0.225
500	0.135	0.130	0.000	0.635	0.620	0.600	0.625
1000	0.440	0.420	0.000	0.915	0.915	0.880	0.900
2500	0.820	0.815	0.000	0.995	0.995	0.990	0.990
5000	0.960	0.960	0.005	1.000	1.000	1.000	1.000
10000	0.990	0.990	0.030	1.000	1.000	1.000	1.000
15000	1.000	1.000	0.080	1.000	1.000	1.000	1.000
20000	1.000	1.000	0.135	1.000	1.000	1.000	1.000
50000		1.000	0.460		1.000	1.000	1.000
100000		1.000	0.770		1.000	1.000	1.000
500000		1.000	0.965		1.000	1.000	1.000
1000000		1.000	0.980		1.000	1.000	1.000

$r > t$, and the deterministic search is repeated using Π_0' in lieu of Π_0 as the starting point. The routine stops if a given number N of consecutive rounds of such perturbations do not yield further cost reductions, at which point the best partition Π observed so far is returned as a candidate solution. To improve deinterleaving reliability, this basic scheme can be run for several independent starting random partitions Π_0, noting the overall cost minimum. The number R of such outer iterations, the maximum count N of consecutive perturbations without improvement, and the neighborhood radii t and r, are parameters controlling the complexity vs. deinterleaving performance trade-off of the heuristic. For our experiments, we found that $R = 5$, $N = 15$, $t = 1$, and $r = 2$, yielded performance virtually identical to a full exhaustive partition search, with orders of magnitude reduction in complexity.[6]

The results of the experiments with memoryless switches are summarized in columns 2–4 of Table **??**. The table shows that the proposed ML-based scheme (in either variant) achieves better than 80% deinterleaving accuracy for sequences as short as $n = 2500$, with perfect deinterleaving for $n \geq 15000$, whereas the scheme of [**?**], although fast, requires much longer sequences, correctly deinterleaving just one sequence in 200 for $n = 5000$, and achieving 98% accuracy for $n = 10^6$ (the maximum length tested in the experiments). This comparison is illustrated by the curves labeled ① and ② in Figure **??**.

In the second set of experiments, we used, for each sequence, the same component processes as in the first set, but with a switch P_w of order one (i.e., $\mathbf{k} = (1,1,1,1)$), with random parameters and uniform marginal subalphabet probabilities. The results are presented in columns 5–6 of Table **??**, and plotted in the curve labeled ③ in Figure **??**. We observe that the additional structure resulting from the switch memory allows for improved deinterleaving performance for shorter sequences: better than 60% accuracy is obtained for sequences as short as $n = 500$, while perfect deinterleaving is obtained for $n \geq 5000$. A comparison with the scheme of [**?**] is omitted in this case, as the determination of appropriate statistics thresholds (not discussed in [**?**]) appears more involved than in the memoryless switch case, and is beyond the scope of this paper.

Finally, in a third set of experiments, we maintained switches of order one, but let the component process P_1 in each case be memoryless (i.e., $\mathbf{k} = (0,1,1,1)$). Recall that, by Lemma 1, the resulting IMPs in this case have ambiguous representations. Results for the heuristic ML-based scheme are presented in columns 7–8 of Table **??**, which list the fraction of sequences of each length for which the deinterleaver picked the canonical partition, or any compatible partition, respectively. We observe that, except for minor deviations for the shorter sequence lengths, the deinterleaver consistently picks the canonical partition, as expected from Theorem 4. The fraction of sequences for which the canonical partition is chosen is plotted in the curve labeled ④ in Figure **??**. Memoryless components are excluded in [**?**], so a comparison is not possible in this case.

Recalling the second remark at the end of Section IV, we note that any nonnegative value of the penalization constant β would have sufficed for the ML schemes in the first two sets of experiments, since the IMPs considered have unique representations, and the order of all the processes tested was bounded by 1. However, a positive value of β is required to recover the canonical partition (and from it, all compatible partitions) in the case of the third set. For shorter sequences, a value of β as small as possible is preferred to exclude non-compatible partitions, while a value of β as large as possible is preferred to recover the canonical partition. Overall, a value $\beta = \frac{1}{2}$ worked well in practice in all cases, providing the best trade-off for shorter sequence lengths (clearly, the choice becomes less critical as the sequence length increases). This value of β is smaller than the value employed in the proof of Theorem 4. In general, the question of determining the minimal penalty that guarantees consistent deinterleaving remains open. The situation bears some similarity to the one encountered with Markov order estimators: while it is known that $\beta = \frac{1}{2}$ guarantees strong consistency in all cases, it

[6] In fact, to keep running times reasonable, the exhaustive search was given the benefit of limiting the search space to partitions Π with $|\Pi| \leq 4$. No such limitation was assumed for the heuristic scheme, whose search space included, in principle, partitions of any size $|\Pi| \leq |\mathcal{A}|$.

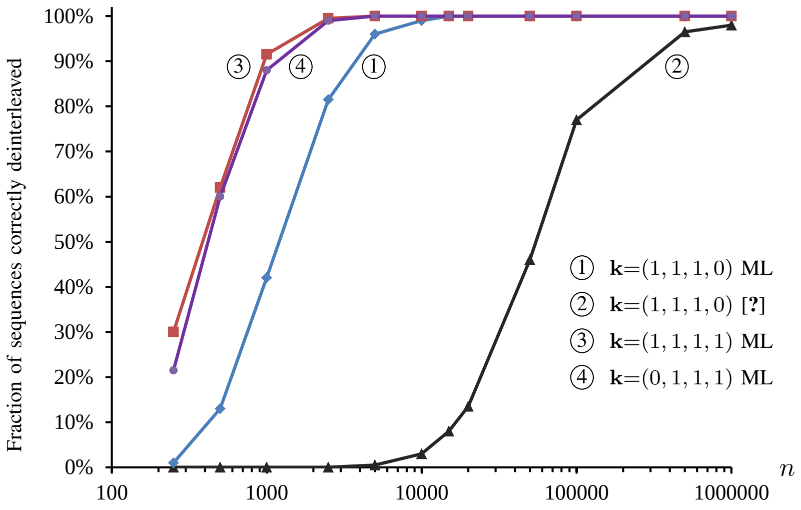

Fig. 3. Deinterleaving success rate vs. sequence length for various IMPs and deinterleavers.

is also known that much smaller penalization constants (or even penalization functions $o(\log n)$) may suffice when the process order is bounded [?]. The general question of the minimal penalization that guarantees consistent unbounded order estimation is, also in this case, open [?].

Appendix A
Uniqueness of IMP representations: derivations

A. Derivation of Theorems 1 and 2

Theorems 1 and 2 will be established through a series of lemmas. The first one (Lemma **??** below) captures some essential properties of the interleaved process $P{=}\mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_\mathrm{w})$ and of the domination relation, which we will draw upon repeatedly in the sequel. These properties follow immediately from our ergodicity and independence assumptions. Intuitively, the key point is that if $A_1 \not\sqsupset A_2$, the interleaved system can always take a trajectory (of positive probability) where it reaches an arbitrary state s of P_1, and then, without returning to A_1, visits any desired part of A_2 any desired number of times (while the state of P_1 remains, of course, unchanged). The last segment of the trajectory, with an unbounded number of occurrences of A_2, can be chosen independently of s. For ease of reference, these observations are formally stated in the lemma below, where $N_a(z)$ denotes the number of occurrences of a symbol a in a string z.

Lemma 2: Consider the subalphabets $A_1, A_2 \in \Pi$, and assume $A_1 \not\sqsupset A_2$.

i) Let M_1 and M be arbitrary integers. There exist strings $U, V \in \Pi^*$ such that $P_\mathrm{w}(UV) > 0$, $N_{A_1}(U) \geq M_1$, $N_{A_1}(V) = 0$, $N_{A_2}(V) \geq M$, and $P_\mathrm{w}(A_1 \,|\, UV) > 0$.

ii) Let M_2 be an arbitrary integer, let s be an arbitrary state of P_1, and consider an arbitrary subset $B_2 \subseteq A_2$ and an integer $M_1 \geq k_1$. There exists an integer $M \geq M_2$, and strings $u, v \in \mathcal{A}^*$ such that uv is consistent with UV (with $|u| = |U|$), where U and V are the strings obtained from Part **??**) for these values of M_1 and M, $u[A_1] = u's$ for some $u' \in A_1^*$, $\big|v[B_2]\big| \geq M_2$, and the choice of v does not depend on s (in particular, the same v can be chosen for any $s \in \mathcal{S}(P_1)$).

Proof: Part **??**) follows from the ergodicity of P_w, the positivity of both $P_\mathrm{w}(A_1)$ and $P_\mathrm{w}(A_2)$, and the definition of domination. The existence of the desired string u in Part **??**) follows further from the independence of the component and switch processes, and from the ergodicity of P_1 (in particular, the fact that $P_1(s) > 0$). Relying also on the ergodicity of P_2, we obtain the string v. The value of M is determined by how many times v must visit A_2 to obtain M_2 occurrences of symbols in the subset B_2. The independence of v from s follows from (2), which allows us to substitute any string over A_1, of positive probability, for $u[A_1]$ in uv, resulting in a string $\tilde{u}v$, with $P(\tilde{u}v) > 0$, $\tilde{u}$ compatible with U, and $\tilde{u}[A_1]$ ending in any desired state of P_1. ■

For succinctness, in the series of lemmas and corollaries that follows, we assume *throughout* that we are given an ambiguous IMP, $P = \mathcal{I}_\Pi(P_1, P_2, \ldots, P_m; P_\mathrm{w}) = \mathcal{I}_{\Pi'}(P_1', P_2', \ldots, P_{m'}'; P_\mathrm{w}')$, where $\Pi = \{A_1, A_2, \ldots, A_m\}$ and $\Pi' = \{A_1', A_2', \ldots, A_{m'}'\}$ are partitions of $\mathcal{A}$, with $\Pi \neq \Pi'$. Clearly, for at least one alphabet A_i we must have $A_i \notin \Pi'$, so we assume, without loss of generality, that $A_1 \notin \Pi'$, and, furthermore, that $A_1 \cap A_1' \neq \phi$. Also, we say that two subalphabets $A_i, A_j \in \Pi$ *share* a subalphabet $A_\ell' \in \Pi'$ if A_ℓ' intersects both A_i and A_j.

Lemma 3: Assume that A_2 shares A_1' with A_1, and $A_1 \not\sqsupset A_2$. Then, for all $a \in A_1 \cap A_1'$, $P_1(a \,|\, s)$ is independent of $s \in \mathcal{S}(P_1)$.

Proof: Let $a \in A_1 \cap A_1'$, and $s \in \mathcal{S}(P_1)$. Let $U, V \in \Pi^*$ and $u, v \in \mathcal{A}^*$ be the strings guaranteed by Lemma **??** for the given state s, $M_2 = \operatorname{ord}(P_1')$, and $B_2 = A_2 \cap A_1'$. Recall that v can be chosen independently of s, and $\big|v[B_2]\big| \geq M_2 = \operatorname{ord}(P_1')$. Let $\hat{v} = v[A_1']$, and let $U'V' = \mathbf{A}_{\Pi'}(uv)$. Then, applying (1) separately to each of the two given IMP representations of P, and noting that $|\hat{v}| \geq \big|v[B_2]\big| \geq \operatorname{ord}(P_1')$, we have

$$P(a|uv) = P_1(a|s)P_{\mathrm{w}}(A_1|UV) = P'_1(a|\hat{v})P'_{\mathrm{w}}(A'_1|U'V').$$

Now, recalling that $P_{\mathrm{w}}(A_1|UV) > 0$ by Lemma **??**(**??**), we obtain

$$P_1(a|s) = \frac{P'_1(a|\hat{v})P'_{\mathrm{w}}(A'_1|U'V')}{P_{\mathrm{w}}(A_1|UV)},$$

which is independent of s. ■

Lemma 4: Assume that $A'_1 \subseteq A_1$, $A'_2 \cap A_1 \neq \phi$, and $A'_1 \not\sqsupset A'_2$. Then, P'_1 is memoryless.

Proof: The lemma follows by applying Lemma **??** with the roles of Π and Π' reversed, and observing that $A'_1 \cap A_1 = A'_1$. ■

Lemma 5: Assume that $A_1 \not\sqsupset A_2$ and $A'_1 \subseteq A_1$. If $A'_2 \in \Pi'$, and $A'_2 \cap A_2 \neq \phi$, then $A'_1 \not\sqsupset A'_2$.

Proof: We apply Lemma **??**, referring only to the strings V and v guaranteed by the lemma, and with $B_2 = A_2 \cap A'_2$. Thus, for any integer M_2, there exists a string $V \in \Pi^*$ and a string v consistent with V such that $M_2 \leq \big|v[B_2]\big| \leq \big|v[A'_2]\big|$, while $N_{A_1}(V) = 0$ and, consequently, $\big|v[A'_1]\big| = 0$. Letting $V' = \mathbf{A}_{\Pi'}(v)$, we then have $N_{A'_1}(V') = 0$ and $N_{A'_2}(V') \geq M_2$ for arbitrarily large M_2. Thus, $A'_1 \not\sqsupset A'_2$. ■

Lemma 6: Assume that A_1 is not totally dominant, $A'_1 \subseteq A_1$, and P'_1 is memoryless. Then, for all $a \in A'_1$, $P_1(a|s)$ is independent of $s \in \mathcal{S}(P_1)$.

Proof: Since $m > 1$ and A_1 is not totally dominant, there exists a subalphabet, say $A_2 \in \Pi$, such that $A_1 \not\sqsupset A_2$. Consider a symbol $a \in A'_1$. Let s be an arbitrary state of P_1, and let U, V, u, and v be the strings guaranteed by Lemma **??** for the state s, with $M_2 = \max\{\mathrm{ord}(P_{\mathrm{w}}), \mathrm{ord}(P'_{\mathrm{w}})\}$. Then, applying (1) to the two IMP representations under consideration, we have

$$P(a|uv) = P_1(a|s)P_{\mathrm{w}}(A_1|UV) = P'_1(a)P'_{\mathrm{w}}(A'_1|U'V'), \tag{14}$$

where $U'V' = \mathbf{A}_{\Pi'}(uv)$, and we have relied on the fact that P'_1 is memoryless. Recall from Lemma **??**(**??**) that $P_{\mathrm{w}}(A_1|UV) > 0$. By our choice of M_2, it follows from (**??**) that

$$P_1(a|s) = P'_1(a)P'_{\mathrm{w}}(A'_1|V')/P_{\mathrm{w}}(A_1|V)\,,$$

which is independent of s. ■

Lemma 7: Assume that A_1 does not dominate any subalphabet A_j, $j > 1$, that shares some $A'_\ell \in \Pi'$ with A_1. Then, either P_1 is memoryless, or A_1 splits into subalphabets in Π'.

Proof: Assume that A_1 does not split into subalphabets in Π'. Then, there exists a subalphabet $A'_\ell \in \Pi'$ that intersects A_1 but is not contained in it, so A_1 shares A'_ℓ with some A_j, $j > 1$. By the lemma's assumptions, we have $A_1 \not\sqsupset A_j$. Therefore, by Lemma **??**, $P_1(a|s)$ is independent of $s \in \mathcal{S}(P_1)$ for all $a \in A_1 \cap A'_\ell$. Assume now that there is also a subalphabet $A'_i \in \Pi'$ such that $A'_i \subseteq A_1$. By Lemma **??**, we have $A'_i \not\sqsupset A'_\ell$, and, therefore, by Lemma **??**, P'_i is memoryless. Thus, by Lemma **??**, $P_1(a|s)$ is independent of s also when $a \in A'_i \subseteq A_1$. Consequently, if A_1 does not split in Π', since every $a \in A_1$ must belong to some $A'_h \in \Pi'$, and $P_1(a|s)$ is independent of $s \in \mathcal{S}(P_1)$ whether A'_h is contained in A_1 or not, P_1 must be memoryless. ■

Lemma 8: Assume that A_1 is not totally dominant, and that A_1 does not dominate any subalphabet A_j, $j > 1$, that shares some A'_ℓ with A_1. Then, P_1 is memoryless.

Proof: If P_1 is not memoryless, then by Lemma **??**, A_1 splits into subalphabets in Π'. Thus, up to re-labeling of subalphabets, we have $A_1 = A'_1 \cup A'_2 \cup \cdots \cup A'_r$, where $A'_i \in \Pi'$, $1 \leq i \leq r \leq m'$, with $r > 1$. Furthermore, by Lemma **??**, at least one of the A'_i, say A'_1, is not memoryless (for, otherwise, P_1 would be memoryless). By Lemma **??**, A'_1 must dominate all A'_i, $2 \leq i \leq r$, and in particular, $A'_1 \sqsupset A'_2$. It follows from this domination relation that there exists a string $U' \in (\Pi')^*$ such that $P'_{\mathrm{w}}(A'_2|U') = 0$, and $P'_{\mathrm{w}}(A'_1|U') > 0$. By the ergodicity of P'_{w}, we can assume without loss of generality that the number of occurrences of subalphabets $A'_1, A'_2, \ldots, A'_r$ in U' is at least $k_1 = \mathrm{ord}(P_1)$. Let u be a string consistent with U'. We have $\big|u[A_1]\big| \geq k_1$; let $t \in \mathcal{S}(P_1)$ be the suffix of length k_1 of $u[A_1]$. Consider a symbol $b \in A'_2$, and let $U'' = \mathbf{A}_\Pi(u)$. Applying (1) separately to the two available IMP representations of P, we have

$$\begin{aligned} P(b|u) &= P_1(b|t)P_{\mathrm{w}}(A_1|U'') \\ &= P'_2(b|u[A'_2])P'_{\mathrm{w}}(A'_2|U') = 0, \end{aligned} \tag{15}$$

where the last equality follows from our choice of U'. On the other hand, since we also have $P'_{\mathrm{w}}(A'_1|U') > 0$, we must have $P(a|u) > 0$ for some $a \in A'_1 \subseteq A_1$, and, therefore, $P_{\mathrm{w}}(A_1|U'') > 0$. Thus, it follows from (**??**) that $P_1(b|t) = 0$. By our assumptions on component processes, there must also be a state $s \in \mathcal{S}(P_1)$ such that $P_1(b|s) > 0$. Since A_1 is not totally dominant, there exists a subalphabet, say A_2, such that $A_1 \not\sqsupset A_2$. Let $B_2 = A_2$ and $M_2 = \max\{\mathrm{ord}(P_{\mathrm{w}}), \mathrm{ord}(P'_{\mathrm{w}})\}$. We apply Lemma **??**(**??**), separately to the states s and t, choosing the same string v for both as allowed by the lemma. Specifically, let U and V be the strings over Π obtained from the lemma, and let $u^{(t)}$, $u^{(s)}$, and v be strings such that $u^{(t)}[A_1] = u't$, $u^{(s)}[A_1] = u''s$ for some u', u'', both $u^{(s)}v$ and $u^{(t)}v$ are consistent with UV, and $|v[A_2]| \geq M_2$. Let $V' = \mathbf{A}_{\Pi'}(v)$. Clearly, $|V| = |V'| \geq M_2$, so V and V' determine states in the respective switches. Applying (1) again, we obtain

$$\begin{aligned} P(b|u^{(s)}v) &= P'_2(b|u^{(s)}[A'_2])P'_{\mathrm{w}}(A'_2|V') \\ &= P_1(b|s)P_{\mathrm{w}}(A_1|V) > 0\,, \end{aligned} \tag{16}$$

where the last inequality follows from our choice of s, and the fact that $P_{\mathrm{w}}(A_1|V) = P_{\mathrm{w}}(A_1|UV) > 0$ by our choice of M_2 and by Lemma **??**(**??**). Thus, we must have $P'_{\mathrm{w}}(A'_2|V') > 0$. On the other hand, we can also write

$$\begin{aligned} P(b|u^{(t)}v) &= P'_2(b|u^{(t)}[A'_2])P'_{\mathrm{w}}(A'_2|V') \\ &= P_1(b|t)P_{\mathrm{w}}(A_1|V) = 0\,, \end{aligned} \tag{17}$$

where the last equality follows from our choice of t. Since, as previously claimed, $P'_{\mathrm{w}}(A'_2|V'){>}0$, it follows from (**??**) that $P'_2(b|u^{(t)}[A'_2]) = 0$, which must hold for all $b \in A'_2$, a contradiction, since every state of P'_2 must have at least one symbol with positive probability (the argument holds even if $|u^{(t)}[A'_2]| < \mathrm{ord}(P'_2)$, reasoning with marginal probabilities). We conclude that P_1 must be memoryless. ■

The following corollary is an immediate consequence of Lemma **??**.

Corollary 2: Assume that A_1 is not dominant. Then, P_1 is memoryless.

Assume now that P_w is such that no two alphabets in Π are in mutual domination. As discussed in Section III-A, this ensures that $\sqsupset$ defines a strict partial order on Π. We classify alphabets in Π into disjoint *layers* L_i, $i \geq 0$, as follows: Given $L_0, L_1, \ldots, L_{i-1}$, and assuming that these layers do not exhaust Π, we let L_i consist of the alphabets that have not been previously assigned to layers, and that only dominate alphabets contained in layers $L_{i'}$, $0 \leq i' < i$ (e.g., L_0 consists of the non-dominant alphabets in Π). Since Π is finite, and every finite set endowed with a strict partial order has minima, L_i is well defined and non-empty. Thus, for some $r \geq 0$, we can write

$$\Pi = L_0 \cup L_1 \cup \cdots \cup L_r \tag{18}$$

where the layers $L_0, L_1, \ldots, L_r$ are all disjoint and non-empty.[7]

We are now ready to present the proofs of Theorems 1 and 2, which rely on the foregoing lemmas and corollaries, and on the classification of alphabets into layers L_i.

Proof of Theorem 1: For the layers in (**??**) we prove, by induction on i, that $L_i \subseteq \Pi'$ for $0 \leq i \leq r$. By the definition of L_0, alphabets $A_j \in L_0$ are not dominant. Thus, by Corollary **??**, we must have $A_j \in \Pi'$, since, by assumption (iii), A_j is not memoryless. Hence, $L_0 \subseteq \Pi'$. Assume now that the induction claim has been proven for $L_0, L_1, \ldots, L_{i-1}$, $1 \leq i \leq r$. Let A_j be any alphabet in L_i. By definition of L_i, A_j only dominates alphabets in layers $L_{i'}$, $i' < i$. But, by our induction hypothesis, alphabets in these layers are elements of Π', and, thus, they do not share with other alphabets from Π. Thus, A_j does not dominate any alphabet A_h with which it shares any A'_ℓ. By Lemma **??**, we must have $A_j \in \Pi'$, since A_j is neither totally dominant nor memoryless by the assumptions of the theorem. Hence, $L_i \subseteq \Pi'$, and our claim is proven. Now, it follows from (**??**) that $\Pi \subseteq \Pi'$, and, since both Π and Π' are partitions of the same alphabet $\mathcal{A}$, we must have $\Pi = \Pi'$. ■

Proof of Theorem 2: Examining the proof of Theorem 1, we observe that when Condition (ii) is removed, any totally dominant alphabet must reside in L_r, the last layer in (**??**). Furthermore, if there is such an alphabet A_{i_0}, it must be unique, for otherwise there would be alphabets in mutual domination. Thus, we have $L_r = \{A_{i_0}\}$, and $A_i \in \Pi'$ for all $i \neq i_0$, and, therefore, A_{i_0} splits into the remaining alphabets in Π' that are not equal to any A_i. ■

B. Derivation of Theorem 3

We start by proving Lemma 1 of Subsection III-C, and then proceed to present an additional auxiliary lemma, and the proof of Theorem 3.

Proof of Lemma 1: Assume $P_1^{(j)}$, $j \in \{1,2\}$, and P'_w satisfy (6)–(8). We prove that $P(u^n) = P'(u^n)$ for all lengths n and sequences $u^n \in \mathcal{A}^n$ by induction on n. For $n = 0$, the claim is trivially true due to the convention $P(\lambda) = P'(\lambda) = 1$. Assume that $P(u^{n-1}) = P'(u^{n-1})$ for $n > 0$ and all $u^{n-1} \in \mathcal{A}^{n-1}$, and consider a sequence $u^n = u^{n-1}u_n$. Let $U^n = \mathbf{A}_\Pi(u^n)$ and $(U')^n = \mathbf{A}_{\Pi'}(u^n)$, and let $S \in \mathcal{S}(P_\mathrm{w})$ and $S' \in \mathcal{S}(P'_\mathrm{w})$ be the states selected by U^{n-1} and $(U')^{n-1}$, respectively. Clearly, we have $S = \Psi(S')$. By the definition of Π', if $U_n = A_i$, $i \in \{2, 3, \ldots, m\}$, then $U'_n = U_n$, and we have

$$\begin{aligned} P'(u^n) &= P'(u^{n-1})P'(u_n|u^{n-1}) \\ &= P'(u^{n-1})P'_\mathrm{w}(A_i|S')P_i(u_n \left| u^{n-1}[A_i]) \right) \\ &= P(u^{n-1})P_\mathrm{w}(A_i|S)P_i(u_n \left| u^{n-1}[A_i]) \right) = P(u^n)\,, \end{aligned} \tag{19}$$

where the second and last equalities follow from the definitions of the respective IMPs, and the third equality follows from the induction hypothesis and (8). On the other hand, if $U_n = A_1$, then $U'_n = B_j$ for some $j \in \{1,2\}$, and we have

$$\begin{aligned} P'(u^n) &= P'(u^{n-1})P'(u_n|u^{n-1}) \\ &= P'(u^{n-1})P'_\mathrm{w}(B_j|S')P_1^{(j)}(u_n) \\ &= P(u^{n-1})P_\mathrm{w}(A_1|S)P_1(B_j)\frac{P_1(u_n)}{P_1(B_j)} = P(u^n)\,, \end{aligned} \tag{20}$$

where, this time, the third equality follows from the induction hypothesis, (8), and (6) (we recall that P_1, $P_1^{(1)}$, and $P_1^{(2)}$ are memoryless). This completes the induction proof and establishes that $P' = P$.

To prove the "only if" part of the lemma, we assume that $P' = P$, and consider a sufficiently long, arbitrary string u^n such that $P(u^n) > 0$. Let $U' = \mathbf{A}_{\Pi'}(u^{n-1})$, and assume first that $u_n \in A_i$ for some $i \geq 2$. Then, similarly to (**??**) (but proceeding from the inside out), and noting that $\mathbf{A}_\Pi(u^{n-1}) = \Psi(U')$, we can write

$$\begin{aligned} & P'(u^{n-1})P'_\mathrm{w}(A_i|U')P_i(u_n \left| u^{n-1}[A_i]) \right) \\ &\qquad = P'(u^n) = P(u^n) \\ &\qquad = P(u^{n-1})P_\mathrm{w}(A_i|\Psi(U'))P_i(u_n \left| u^{n-1}[A_i]) \right)\,. \end{aligned} \tag{21}$$

Since $P' = P$, and $P(u^n) > 0$, (**??**) can be simplified to

$$P'_\mathrm{w}(A_i|U') = P_\mathrm{w}(A_i|\Psi(U')), \quad i \in \{2, 3, \ldots, m\}\,, \tag{22}$$

for arbitrary $U' \in (\Pi')^{n-1}$ of positive probability. Consider now the case $u_n = b \in B_j$, $j \in \{1,2\}$. Then, in analogy with (**??**), we write

$$\begin{aligned} & P'(u^{n-1})P'_\mathrm{w}(B_j|U')P_1^{(j)}(b) \\ &\qquad = P'(u^n) = P(u^n) \\ &\qquad = P(u^{n-1})P_\mathrm{w}(A_1|\Psi(U'))P_1(b)\,. \end{aligned} \tag{23}$$

Adding over all $b \in B_j$ and simplifying, we obtain

$$P'_\mathrm{w}(B_j|U') = P_\mathrm{w}(A_1|\Psi(U'))P_1(B_j), \quad j \in \{1,2\}\,, \tag{24}$$

again for arbitrary U'. Conditions (7)–(8) now follow readily from (**??**) and (**??**) (which imply, in particular, that $k_\mathrm{w} = k'_\mathrm{w}$),

[7] The layers L_i correspond to height levels in the directed acyclic graph associated with the transitive reduction of the partial order $\sqsupset$.

and Condition (6) follows by substituting the right-hand side of (**??**) for $P'_{\mathrm{w}}(B_j|U')$ in (**??**) and solving for $P_1^{(j)}(b)$. ■

We say that the representations $\mathcal{I}_{\Pi}$ and $\mathcal{I}_{\Pi'}$ of an IMP P *coincide up to memoryless components* if the set of component processes of positive order is the same in both representations. The following lemma establishes the uniqueness of canonical partitions.

Lemma 9: Let $\mathcal{I}_{\Pi}$ and $\mathcal{I}_{\Pi'}$ be IMP representations of a process P that coincide up to memoryless components, and such that both are canonical. Then, $\Pi = \Pi'$.

Proof: Assume that $\Pi \neq \Pi'$, and let Π'' be the smallest common refinement of Π and Π' (i.e., $\Pi'' = \big\{A_i \cap A'_j \,\big|\, A_i \in \Pi,\ A'_j \in \Pi',\ A_i \cap A'_j \neq \phi\big\}$). By repeated application of Lemma 1, there exists an IMP representation $\mathcal{I}_{\Pi''}(P''_1, P''_2, \ldots, P''_{m''}; P''_{\mathrm{w}})$ of P. This representation is a memoryless refinement of both $\mathcal{I}_{\Pi}$ and $\mathcal{I}_{\Pi'}$. Since $\Pi \neq \Pi'$, there exists an alphabet, say $A'_1 \in \Pi'$ such that $A'_1 \notin \Pi$, P'_1 is memoryless, and we can assume without loss of generality that A'_1 intersects at least two alphabets, A_1 and A_2, in Π (otherwise, we can switch the roles of Π and Π'). Let $B_1 = A'_1 \cap A_1$ and $B_2 = A'_1 \cap A_2$, so that $B_1, B_2 \in \Pi''$. Applying Lemma 1 separately to $\mathcal{I}_{\Pi}$ and to $\mathcal{I}_{\Pi'}$ with respect to the refinement $\mathcal{I}_{\Pi''}$, we can write, for any $S'' \in \mathcal{S}(P''_{\mathrm{w}})$, and denoting $S = \Psi_{\Pi,\Pi''}(S'')$ and $S' = \Psi_{\Pi',\Pi''}(S'')$,

$$P''_{\mathrm{w}}(B_1|S'') = P_{\mathrm{w}}\left(A_1|\, S\right) P_1(B_1) = P'_{\mathrm{w}}\left(A'_1|\, S'\right) P'_1(B_1),$$

where $P_1(B_1)$ and $P'_1(B_1)$ are nonzero. (Notice that the equation holds also when $B_1 = A_1$, i.e., when A_1 is not actually refined in Π''.) Therefore, we can write

$$P_{\mathrm{w}}\left(A_1|\, S\right) = \frac{P'_{\mathrm{w}}\left(A'_1|\, S'\right) P'_1(B_1)}{P_1(B_1)}\,. \tag{25}$$

Using a similar argument for B_2 and A_2, we obtain

$$P_{\mathrm{w}}\left(A_2|\, S\right) = \frac{P'_{\mathrm{w}}\left(A'_1|\, S'\right) P'_1(B_2)}{P_1(B_2)}\,. \tag{26}$$

It follows from (**??**) and (**??**) that if $P'_{\mathrm{w}}\left(A'_1|\, S'\right) = 0$, then $P_{\mathrm{w}}(A_1|S) = P_{\mathrm{w}}(A_2|S) = 0$, and, otherwise,

$$\frac{P_{\mathrm{w}}\left(A_2|S\right)}{P_{\mathrm{w}}\left(A_1|S\right)} = \frac{P_1(B_1)P'_1(B_2)}{P'_1(B_1)P_1(B_2)} \overset{\Delta}{=} \gamma,$$

where $\gamma > 0$ is independent of S'' (and of S). Observing that S can assume any value in $\mathcal{S}(P_{\mathrm{w}})$, we conclude, by Lemma 1 and the remarks following its statement, that A_1 could be merged with A_2, contradicting the assumption that $\mathcal{I}_{\Pi}$ is canonical. Thus, we must have $\Pi = \Pi'$. ■

Proof of Theorem 3: Assume $P = P'$. Since there are no dominant alphabets in either representation, it follows from Corollary **??** that the representations must coincide up to memoryless components. It then follows from Lemma **??** that the canonical partitions of $\mathcal{I}_{\Pi}$ and $\mathcal{I}_{\Pi'}$ must be identical, and, thus, since they generate the same process, we must have $(\mathcal{I}_{\Pi})^* \equiv (\mathcal{I}_{\Pi'})^*$. The "if" part is straightforward, since $(\mathcal{I}_{\Pi})^*$ generates P, and $(\mathcal{I}_{\Pi'})^*$ generates P'. ■

Appendix B
The deinterleaving scheme: derivations

We will prove Theorem 4 through the auxiliary Lemmas **??** and **??** below, for which we need some additional definitions.

Let $F = (S, s_0, f)$ be an FSM, and let P and Q be processes generated by F, such that P is ergodic. The divergence (relative to F) between P and Q is defined as

$$D(P||Q) = \sum_{s\in S} P(s) D\big(P(\cdot|s)\big|\big|Q(\cdot|s)\big)\,, \tag{27}$$

where $P(s)$ denotes the stationary probability of the state $s \in S$, and $D\big(P(\cdot|s)\big|\big|Q(\cdot|s)\big)$ denotes the Kullback-Leibler divergence between the conditional distributions $P(\cdot|s)$ and $Q(\cdot|s)$. It is well known (see, e.g., [**?**]) that $D(P||Q)$ as defined in (**??**) is equal to the asymptotic normalized Kullbak-Liebler divergence between the processes P and Q, namely,

$$D(P||Q) = \lim_{n\to\infty} \frac{1}{n} \sum_{z^n\in\mathcal{A}^n} P(z^n) \log \frac{P(z^n)}{Q(z^n)}\,.$$

Let $\mathcal{V}(\mathcal{F}_{\Pi,\mathbf{k}})$ denote the set of parameter vectors corresponding to ergodic *unconstrained* FSM sources based on $\mathcal{F}_{\Pi,\mathbf{k}}$, and let $\overline{\mathcal{V}}(\mathcal{F}_{\Pi,\mathbf{k}})$ denote its topological closure. Assuming full parametrization, this set is a convex polytope in $\mathcal{K}$-dimensional Euclidean space. The boundary of $\overline{\mathcal{V}}(\mathcal{F}_{\Pi,\mathbf{k}})$ consists of parameter vectors with certain transition probabilities set to zero or one. Some of these vectors do not correspond to ergodic FSM sources, namely, those that make some of the marginal probabilities of states in S vanish (e.g., parameter vectors where the probabilities of all the transitions leading to a state vanish). Let $\mathcal{V}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$, in turn, denote the set of parameter vectors of IMP-constrained FSM sources based on $\mathcal{F}_{\Pi,\mathbf{k}}$, and $\overline{\mathcal{V}}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$ its topological closure. The set $\overline{\mathcal{V}}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$ is a closed κ-dimensional hypersurface within $\overline{\mathcal{V}}(\mathcal{F}_{\Pi,\mathbf{k}})$, determined by the parameter relations implicit in (3). As before, boundary points in $\overline{\mathcal{V}}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$ are either in $\mathcal{V}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$, or do not correspond to valid IMPs. We shall make use of these relations in the sequel.

The following lemma will be useful in proving the first claim of Theorem 4.

Lemma 10: Let $P = \mathcal{I}_{\Pi}(P_1, P_2, \ldots, P_m; P_{\mathrm{w}})$, and let $\mathbf{k} = (k_1, k_2, \ldots, k_m, k_{\mathrm{w}})$ be the corresponding order vector. Let Π' be a partition of $\mathcal{A}$ incompatible with P, and $\mathbf{k}'$ an arbitrary order vector of dimension $|\Pi'| + 1$. Then, for a sample z^n from P, and for any $\beta \geq 0$, we have

$$C_{\Pi',\mathbf{k}'}(z^n) > C_{\Pi,\mathbf{k}}(z^n) \quad \text{a.s. as } n \to \infty\,.$$

Proof: Let F^+ be a common refinement[8] of $F = \mathcal{F}_{\Pi,\mathbf{k}}$ and $F' = \mathcal{F}_{\Pi',\mathbf{k}'}$. Let $V = \mathcal{V}(F^+)$ denote the space of all valid parameter vectors for FSM sources based on F^+, and let $\overline{\mathcal{V}}(F^+)$ denote its topological closure. The constraints satisfied by IMP sources based on F and F' are extended to their representations in V (notice that a refinement increases the dimension of the parameter vector by "cloning" parameters, together with their constraints). Thus, as mentioned in the

[8] It is always possible to construct a common refinement of two FSMs, e.g., one whose state set is the Cartesian product of the state sets of the refined FSMs.

discussion immediately preceding the lemma, the set of all IMP-constrained FSM sources based on F' maps to a lower-dimensional hypersurface $V' = \mathcal{V}_{\mathcal{I}}(F^+) \subseteq V$, with closure $\overline{V}'$. We claim that the representation of P in V is outside the closed hypersurface $\overline{V}'$, and, thus, at positive Euclidean (or L_1) distance from it. To prove the claim, we first notice that since Π' is, by assumption, incompatible with P, no valid IMP-constrained assignment of parameters for F' can generate P, and, thus, $P \notin V'$. Furthermore, since points in $\overline{V}' \setminus V'$ correspond to "invalid" IMPs with unreachable states, we must have $P \notin \overline{V}'$, and, therefore, P is at positive distance from $\overline{V}'$, as claimed. The ergodicity of P also implies that, in its representation in V, all the states of F^+ have positive stationary probabilities. Applying Pinsker's inequality on a state by state basis in (**??**) for F^+, we conclude that for any process $P' \in V'$, we have

$$D(P||P') \geq \Delta\,, \tag{28}$$

for some constant $\Delta > 0$. Now, recall that $\hat{P}^*_{F^+}(z^n)$ denotes the ML probability of z^n with respect to F^+ for an *unconstrained* FSM source. It follows from the definition of $\hat{P}^*_{F^+}(z^n)$ and of the divergence $D(\cdot||\cdot)$ in (**??**) that for any process Q generated by F^+, we have

$$-\log Q(z^n) = -\log \hat{P}^*_{F^+}(z^n) + nD\big(\hat{P}^*_{F^+}\big|\big|Q\big)\,. \tag{29}$$

In particular, since F^+ can generate any process that either F or F' can generate, it can assign to z^n its IMP-constrained ML probabilities with respect to F and F' which are, respectively, $\hat{P}_{\Pi,\mathbf{k}}(z^n) = 2^{-\hat{H}_{\Pi,\mathbf{k}}(z^n)}$ and $\hat{P}_{\Pi',\mathbf{k}'}(z^n) = 2^{-\hat{H}_{\Pi',\mathbf{k}'}(z^n)}$. Applying (**??**) to $Q = \hat{P}_{\Pi,\mathbf{k}}$ and $Q = \hat{P}_{\Pi',\mathbf{k}'}$ separately, subtracting on each side of the resulting equations, and dividing by n, we obtain

$$\frac{1}{n}\left(\hat{H}_{\Pi',\mathbf{k}'}(z^n) - \hat{H}_{\Pi,\mathbf{k}}(z^n)\right) \\ = D\big(\hat{P}^*_{F^+}\big|\big|\hat{P}_{\Pi',\mathbf{k}'}\big) - D\big(\hat{P}^*_{F^+}\big|\big|\hat{P}_{\Pi,\mathbf{k}}\big)\,. \tag{30}$$

Now, since z^n is a sample from P, the empirical measures $\hat{P}^*_{F^+}$ and $\hat{P}_{\Pi,\mathbf{k}}$ tend to the true process P almost surely in the divergence sense, i.e., $D\big(\hat{P}^*_{F^+}\big|\big|P\big) \to 0$ and $D\big(\hat{P}_{\Pi,\mathbf{k}}\big|\big|P\big) \to 0$ a.s. as $n \to \infty$. Also, an empirical conditional probability value in either $\hat{P}^*_{F^+}$ or $\hat{P}_{\Pi,\mathbf{k}}$ is surely zero if the corresponding parameter in P is zero, and almost surely bounded away from zero otherwise. Hence, we also have $D\big(\hat{P}^*_{F^+}\big|\big|\hat{P}_{\Pi,\mathbf{k}}\big) \to 0$ a.s. as $n \to \infty$. On the other hand, since $\hat{P}_{\Pi',\mathbf{k}'} \in V'$, (**??**) applies with $P' = \hat{P}_{\Pi',\mathbf{k}'}$, so we have $D\big(P\big|\big|\hat{P}_{\Pi',\mathbf{k}'}\big) \geq \Delta > 0$, and, using a similar convergence argument, $D\big(\hat{P}^*_{F^+}\big|\big|\hat{P}_{\Pi',\mathbf{k}'}\big) \geq \Delta > 0$ a.s. as $n \to \infty$. Thus, it follows from (**??**) that

$$\frac{1}{n}\left(\hat{H}_{\Pi',\mathbf{k}'}(z^n) - \hat{H}_{\Pi,\mathbf{k}}(z^n)\right) \geq \Delta > 0 \quad \text{a.s. as } n \to \infty,$$

which implies, by (11),

$$\frac{1}{n}\Big(C_{\Pi',\mathbf{k}'}(z^n) - C_{\Pi,\mathbf{k}}(z^n)\Big) \geq \Delta > 0 \;\text{ a.s. as } n{\to}\infty\,, \tag{31}$$

since the contribution of the $O(\log n)$ penalty terms to the costs vanishes asymptotically in this case, for any choice of $\beta \geq 0$. ■

The following lemma, in turn, will be useful in establishing the second claim of Theorem 4.

Lemma 11: Let Π, Π', $\mathcal{I}_\Pi$ and $\mathcal{I}_{\Pi'}$ be as defined in Lemma 1, so that $\mathcal{I}_{\Pi'}$ is a memoryless refinement of $\mathcal{I}_\Pi$. Let $\mathbf{k} = (0, k_2, \ldots, k_m, k_{\mathrm{w}})$ be the order vector corresponding to $\mathcal{I}_\Pi$, and $\mathbf{k}' = (0, 0, k_2, \ldots, k_m, k_{\mathrm{w}})$ that of $\mathcal{I}_{\Pi'}$. For a sample z^n from P, and an appropriate choice of β, we have: if $k_{\mathrm{w}} > 0$, then

$$C_{\Pi',\mathbf{k}'}(z^n) > C_{\Pi,\mathbf{k}}(z^n) \quad \text{a.s.} \;\; \text{as } n \to \infty\,, \tag{32}$$

while if $k_{\mathrm{w}} = 0$, then

$$C_{\Pi',\mathbf{k}'}(z^n) = C_{\Pi,\mathbf{k}}(z^n)\,. \tag{33}$$

Proof: We first notice that, by Lemma 1, $\mathcal{P}_{\mathcal{I}}(\mathcal{F}_{\Pi,\mathbf{k}})$ can alternatively be characterized as the subset of $\mathcal{P}_{\mathcal{I}}(\mathcal{F}_{\Pi',\mathbf{k}'})$ formed by distributions such that the switch process P'_{w} satisfies the following two constraints, where Ψ denotes the mapping defined prior to Lemma 1:

a) If $S', S'' \in \mathcal{S}(P'_{\mathrm{w}})$ satisfy $\Psi(S') = \Psi(S'')$ then the corresponding conditional distributions coincide;
b) For every $S \in \mathcal{S}(P'_{\mathrm{w}})$, $P'_{\mathrm{w}}(B_2|S) = \gamma P'_{\mathrm{w}}(B_1|S)$ for some parameter γ, independent of S.

Clearly, the dimension of both parametrizations remains $\kappa(\Pi, \mathbf{k})$. It then follows from the definition of empirical entropy of an IMP and from (10) that

$$\hat{H}_{\Pi,\mathbf{k}}(z^n) = \hat{H}_0(z^n[B_1]) + \hat{H}_0(z^n[B_2]) \\ + \sum_{i=2}^{m} \hat{H}_{k_i}(\mathbf{z}_i) - \log \tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n)) \tag{34}$$

where $\tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n))$ denotes the ML probability, subject to the above two constraints, of the switch sequence $\mathbf{A}_{\Pi'}(z^n)$. Therefore,

$$\hat{H}_{\Pi,\mathbf{k}}(z^n) {-} \hat{H}_{\Pi',\mathbf{k}'}(z^n) = \\ -\log \tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n)) - \hat{H}_{k_{\mathrm{w}}}(\mathbf{A}_{\Pi'}(z^n)) \tag{35}$$

which depends on z^n only through $\mathbf{A}_{\Pi'}(z^n)$. The above difference is obviously nonnegative, since Π' is a refinement of Π; equivalently, looking at the right-hand side of (**??**), the maximization leading to $\tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n))$ involves more constraints than the one leading to $\hat{H}_{k_{\mathrm{w}}}(\mathbf{A}_{\Pi'}(z^n))$. Recalling the difference in model sizes computed in (13), we obtain, together with (**??**), that

$$C_{\Pi',\mathbf{k}'}(z^n) - C_{\Pi,\mathbf{k}}(z^n) = \\ \hat{H}_{k_{\mathrm{w}}}(\mathbf{A}_{\Pi'}(z^n)) \;+\; \beta m(m{+}1)^{k_{\mathrm{w}}} \log(n{+}1) \\ -\Big(-\log \tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n)){+}\beta\left((m{-}1)m^{k_{\mathrm{w}}}{+}1\right)\log(n{+}1)\Big)\,. \tag{36}$$

Thus, the left-hand side of (**??**) is equal to the difference between penalized ML probabilities for a switch sequence of length n on Π', for two candidate models. The first model is Markov of order k_{w}, whereas the second model differs from the plain Markov one in that states of $(\Pi')^{k_{\mathrm{w}}}$ have merged according to the mapping Ψ, so that the number of states is now $m^{k_{\mathrm{w}}}$ (constraint (**??**) above), and imposes the additional constraint (**??**) on the conditional probabilities of B_1 and B_2

(notice that the number of free parameters in this model is indeed $(m-1)m^{k_{\mathrm{w}}}+1$). Since, by our assumptions, the number of states of the underlying switch process is $m^{k_{\mathrm{w}}}$ and the process does satisfy the additional constraint (**??**), the left-hand side of (**??**) can be viewed as a penalized ML test of two models, the minimal, "true" one, and a refinement of it. When $k_{\mathrm{w}}=0$, the refinement is trivial and the penalty difference is 0, implying (**??**). When $k_{\mathrm{w}}>0$, our analysis, presented next, will rely on tools developed in [**?**] to study refinements of the type given by constraint (**??**), which will be extended here to deal also with the type of refinement given by constraint (**??**). As in [**?**], we will show the strong consistency of the penalized ML test for suitable β.

Specifically, given a sequence Z^n over $(\Pi')^n$, we start by defining the following "semi-ML" Markov probability distribution $\tilde{\tilde{P}}'_{\mathrm{w}}$ of order k_{w}: For every $S\in(\Pi')^{k_{\mathrm{w}}}$ and $i=2,\cdots,m$, we define $\tilde{\tilde{P}}'_{\mathrm{w}}(A_i|S)=P_{\mathrm{w}}(A_i|S)$ if $S\in\Pi^{k_{\mathrm{w}}}$ (i.e., S is a k_{w}-tuple over $(\Pi')^{k_{\mathrm{w}}}$ not containing either B_1 or B_2, and is therefore an unrefined state of $\Pi^{k_{\mathrm{w}}}$), and $\tilde{\tilde{P}}'_{\mathrm{w}}(A_i|S)=\hat{P}'_{\mathrm{w}}(A_i|\Psi(S))$ otherwise, where $\hat{P}'_{\mathrm{w}}(A_i|\bar{S})$ denotes the ratio between the number of occurrences of A_i following a state $\bar{S}$ in Z^n, and the number of occurrences of $\bar{S}$, where $\bar{S}$ can be either in $\Pi^{k_{\mathrm{w}}}$ (as is $\Psi(S)$ in this case) or, more generally, in $(\Pi')^{k_{\mathrm{w}}}$. The distribution is completely determined by further setting, for every $S\in(\Pi')^{k_{\mathrm{w}}}$, the relation $\tilde{\tilde{P}}'_{\mathrm{w}}(B_2|S)=\hat{\gamma}\tilde{\tilde{P}}'_{\mathrm{w}}(B_1|S)$, where

$$\hat{\gamma}\triangleq\frac{N_{B_2}(Z^n)}{N_{B_1}(Z^n)}$$

is the ML estimate of γ based on Z^n, given by the ratio between the number of occurrences of B_2 and B_1 in Z^n (independent of S), provided $N_{B_1}(Z^n)>0$. Otherwise, if $N_{B_1}(Z^n)=0$, we let $\tilde{\tilde{P}}'_{\mathrm{w}}(B_1|S)=0$. Notice that $\hat{P}'_{\mathrm{w}}(A_i|S)$ is the ML estimate of $P'_{\mathrm{w}}(A_i|S)$ regardless of the constraint relating $P'_{\mathrm{w}}(B_2|S)$ and $P'_{\mathrm{w}}(B_1|S)$. Since, in order to obtain the (constrained) ML probability $\tilde{P}'_{\mathrm{w}}(Z^n)$, one can first maximize over γ and then perform independent maximizations of the conditional probabilities for each state, it is easy to see that, for any $Z^n\in(\Pi')^n$, we have

$$P'_{\mathrm{w}}(Z^n)\le\tilde{\tilde{P}}'_{\mathrm{w}}(Z^n)\le\tilde{P}'_{\mathrm{w}}(Z^n) \tag{37}$$

justifying our reference to $\tilde{\tilde{P}}'_{\mathrm{w}}$ as a "semi-ML" Markov probability distribution.

Another (non-constrained) "semi-ML" Markov probability distribution $\hat{\hat{P}}'_{\mathrm{w}}$ of order k_{w} is defined as follows: For every $S\in(\Pi')^{k_{\mathrm{w}}}\cap\Pi^{k_{\mathrm{w}}}$ we define $\hat{\hat{P}}'_{\mathrm{w}}(A_i|S)=P_{\mathrm{w}}(A_i|S)$, $i=2,\cdots,m$, and $\hat{\hat{P}}'_{\mathrm{w}}(B_2|S)=\hat{\gamma}_S\hat{\hat{P}}'_{\mathrm{w}}(B_1|S)$, where $\hat{\gamma}_S$ denotes the ratio between the number of occurrences of B_2 and B_1 following state S in Z^n, provided the latter number is positive (otherwise, we let $\hat{\hat{P}}'_{\mathrm{w}}(B_1|S)=0$). For all other states $S\in(\Pi')^{k_{\mathrm{w}}}$ and every $Z\in\Pi'$, we define $\hat{\hat{P}}'_{\mathrm{w}}(Z|S)=\hat{P}'_{\mathrm{w}}(Z|S)$.

Notice that for states in $(\Pi')^{k_{\mathrm{w}}}\cap\Pi^{k_{\mathrm{w}}}$, $\hat{\hat{P}}'_{\mathrm{w}}$ differs from $\tilde{\tilde{P}}'_{\mathrm{w}}$ in that the ratio between the conditional probabilities of B_2 and B_1 depends on S (while the conditional probabilities of all A_i, $i=2,\cdots,m$, under the two measures, coincide, and are independent of Z^n). For the other states, both $\hat{\hat{P}}'_{\mathrm{w}}$ and $\tilde{\tilde{P}}'_{\mathrm{w}}$ use ML estimates (which are constrained for the latter distribution). The key observation is then that

$$\begin{aligned}-\log\tilde{P}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n))-\hat{H}_{k_{\mathrm{w}}}(\mathbf{A}_{\Pi'}(z^n))=\\ -\log\tilde{\tilde{P}}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n))+\log\hat{\hat{P}}'_{\mathrm{w}}(\mathbf{A}_{\Pi'}(z^n))\,.\end{aligned} \tag{38}$$

Now, the probability $P_{\mathrm{err}}(n)$ of the error event is given by

$$P_{\mathrm{err}}(n)\triangleq\sum_{z^n:C_{\Pi',\mathbf{k}'}(z^n)\le C_{\Pi,\mathbf{k}}(z^n)}P(z^n)=\sum_{Z^n\in\mathcal{E}}P'_{\mathrm{w}}(Z^n) \tag{39}$$

where $\mathcal{E}$ denotes the subset of switch sequences Z^n over $(\Pi')^n$ satisfying

$$\begin{aligned}\hat{H}_{k_{\mathrm{w}}}(Z^n)+\beta m(m+1)^{k_{\mathrm{w}}}\log(n+1)\le\\ -\log\tilde{P}'_{\mathrm{w}}(Z^n)+\beta\left((m-1)m^{k_{\mathrm{w}}}+1\right)\log(n+1)\end{aligned}$$

and the second equality in (**??**) follows from (**??**). By (**??**), and writing for succinctness $\mu\triangleq m(m+1)^{k_{\mathrm{w}}}-(m-1)m^{k_{\mathrm{w}}}-1$, we have $Z^n\in\mathcal{E}$ if and only if

$$-\log\tilde{\tilde{P}}'_{\mathrm{w}}(Z^n)\ge-\log\hat{\hat{P}}'_{\mathrm{w}}(Z^n)+\beta\mu\log(n+1)$$

or, equivalently,

$$\tilde{\tilde{P}}'_{\mathrm{w}}(Z^n)\le(n+1)^{-\beta\mu}\hat{\hat{P}}'_{\mathrm{w}}(Z^n)\,.$$

Therefore, by the first inequality in (**??**), the rightmost summation in (**??**) can be upper-bounded to obtain

$$P_{\mathrm{err}}(n)\le(n+1)^{-\beta\mu}\sum_{Z^n\in(\Pi')^n}\hat{\hat{P}}'_{\mathrm{w}}(Z^n)\,. \tag{40}$$

Notice that the probability distributions in the summation in the right-hand side of (**??**) depend on Z^n. Clearly, when restricted to sequences Z^n giving rise to the same distribution, the partial sum is upper-bounded by 1. Therefore, the overall sum is upper-bounded by the number N of distinct such distributions. Now, there are $(m+1)^{k_{\mathrm{w}}}-(m-1)^{k_{\mathrm{w}}}$ states given by k_{w}-tuples containing either B_1 or B_2 and, by the definition of $\hat{\hat{P}}'_{\mathrm{w}}$, for each of these states there are at most $(n+1)^{m+1}$ possible conditional distributions, given by the composition of the corresponding substring in Z^n. For each of the remaining $(m-1)^{k_{\mathrm{w}}}$ states, the definition of $\hat{\hat{P}}'_{\mathrm{w}}$ implies that there are at most $(n+1)^2$ possible conditional distributions. Therefore, writing $\mu'\triangleq\left((m+1)^{k_{\mathrm{w}}}-(m-1)^{k_{\mathrm{w}}}\right)(m+1)$, we obtain

$$N\le(n+1)^{2(m-1)^{k_{\mathrm{w}}}+\mu'}$$

implying

$$P_{\mathrm{err}}(n)\le(n+1)^{2(m-1)^{k_{\mathrm{w}}}+\mu'-\beta\mu}\,. \tag{41}$$

Since $m\ge2$ and $k_{\mathrm{w}}\ge1$, and recalling the definitions of μ and μ' above, it can be readily shown that, for any $\beta>3$, the exponent on the right-hand side of (**??**) is less than -1. Thus, $P_{\mathrm{err}}(n)$ is summable and the result follows from the Borel-Cantelli lemma. ■

With these tools in hand, we are now ready to prove Theorem 4.

Proof of Theorem 4: Define the set

$$\mathbf{\Pi}'=\{\,(\Pi',\mathbf{k}')\mid\Pi'\ \text{is } \textit{incompatible} \text{ with }\ P\,\}\ .$$

To establish the first claim of the theorem, we will prove that $\left(\hat{\Pi}(z^n), \hat{\mathbf{k}}(z^n)\right) \notin \mathbf{\Pi}'$ a.s. as $n \to \infty$. Consider a partition $\bar{\Pi}$ compatible with P, denote by $\bar{\mathbf{k}}$ the associated order vector, and let $\bar{\kappa} = \kappa(\bar{\Pi}, \bar{\mathbf{k}})$. Let $\kappa_0 > \bar{\kappa}$ denote a threshold for model sizes, which is independent of n, and will be specified in more detail later on. Write $\mathbf{\Pi}' = \mathbf{\Pi}_1 \cup \mathbf{\Pi}_2$, where

$$\mathbf{\Pi}_1 = \{\, (\Pi', \mathbf{k}') \in \mathbf{\Pi}' \mid \kappa(\Pi', \mathbf{k}') < \kappa_0 \,\} ,$$

and $\mathbf{\Pi}_2 = \mathbf{\Pi}' \setminus \mathbf{\Pi}_1$. Clearly, $\mathbf{\Pi}_1$ is finite and its size is independent of n. By Lemma **??**, for each pair $(\Pi', \mathbf{k}') \in \mathbf{\Pi}_1$, we have $C_{\Pi',\mathbf{k}'}(z^n) > C_{\bar{\Pi},\bar{\mathbf{k}}}(z^n)$ a.s. as $n \to \infty$, for any penalization coefficient $\beta \geq 0$. Thus, the search in (12), almost surely, will not return a pair from $\mathbf{\Pi}_1$. It remains to prove that it will not return a pair from $\mathbf{\Pi}_2$ either. As mentioned, the difficulty here is that the size of $\mathbf{\Pi}_2$ (and of the IMP models associated with pairs in $\mathbf{\Pi}_2$) is not bounded as $n \to \infty$, and we cannot establish the desired result with a finite number of applications of Lemma **??**. As before, we adapt some tools from [**?**] to IMP-constrained FSM sources.

For $(\Pi', \mathbf{k}') \in \mathbf{\Pi}_2$, let $\mathcal{P}_{\Pi',\mathbf{k}'}$ denote the probability that a solution with $(\Pi', \mathbf{k}')$ is preferred over $(\bar{\Pi}, \bar{\mathbf{k}})$ in the minimization. Define

$$B_{\Pi',\mathbf{k}'} = \left\{ z^n \,|\, C_{\Pi',\mathbf{k}'}(z^n) \leq C_{\bar{\Pi},\bar{\mathbf{k}}}(z^n) \,\right\} .$$

Clearly, we have

$$\mathcal{P}_{\Pi',\mathbf{k}'} \leq \sum_{z^n \in B_{\Pi',\mathbf{k}'}} P(z^n) \,. \tag{42}$$

By the definitions of $B_{\Pi',\mathbf{k}'}$ and of the cost function in (11), and denoting $\kappa' = \kappa(\Pi', \mathbf{k}')$, we have, for $z^n \in B_{\Pi',\mathbf{k}'}$,

$$\hat{H}_{\bar{\Pi},\bar{\mathbf{k}}}(z^n) \geq \hat{H}_{\Pi',\mathbf{k}'}(z^n) + \beta(\kappa' - \bar{\kappa}) \log(n+1) \,. \tag{43}$$

Recalling that $P(z^n) \leq \hat{P}_{\bar{\Pi},\bar{\mathbf{k}}}(z^n)$ by (10), and that $\hat{H}_{\Pi',\mathbf{k}'}(z^n) = -\log \hat{P}_{\Pi',\mathbf{k}'}(z^n)$, it follows from (**??**) that

$$P(z^n) \leq (n+1)^{\beta(\bar{\kappa}-\kappa')} \hat{P}_{\Pi',\mathbf{k}'}(z^n) \,, \quad z^n \in B_{\Pi',\mathbf{k}'} \,,$$

and, hence, together with (**??**), and applying an obvious bound, we obtain

$$\begin{aligned} \mathcal{P}_{\Pi',\mathbf{k}'} &\leq (n+1)^{\beta(\bar{\kappa}-\kappa')} \sum_{z^n \in B_{\Pi',\mathbf{k}'}} \hat{P}_{\Pi',\mathbf{k}'}(z^n) \\ &\leq (n+1)^{\beta(\bar{\kappa}-\kappa')} \sum_{z^n \in \mathcal{A}^n} \hat{P}_{\Pi',\mathbf{k}'}(z^n) \,. \end{aligned} \tag{44}$$

In analogy to the reasoning following (**??**) in the proof of Lemma **??**, the summation on the right-hand side of (**??**) can be upper-bounded by the number of different empirical distributions (or *types*) for IMPs based on $\mathcal{F}_{\Pi',\mathbf{k}'}$ and sequences of length n. It is well established (see, e.g., [**?**]) that $(\alpha_i - 1)\alpha_i^{k_i}$ counts suffice to determine the empirical distribution for the Markov component P_i (and similarly for the switch P_{w}). Hence, recalling (5), we conclude that $\kappa' = \kappa(\Pi', \mathbf{k}')$ counts suffice to determine an empirical distribution $\hat{P}_{\Pi',\mathbf{k}'}(z^n)$, and, therefore, the number of such distributions is upper-bounded (quite loosely) by $(n+1)^{\kappa'}$. Thus, it follows from (**??**) that

$$\mathcal{P}_{\Pi',\mathbf{k}'} \leq (n+1)^{\beta(\bar{\kappa}-\kappa')+\kappa'} \,. \tag{45}$$

We next bound the number of pairs $(\Pi', \mathbf{k}')$ satisfying $\kappa(\Pi', \mathbf{k}') = \kappa'$ for a given $\kappa' \geq \kappa_0$. The number of partitions Π' is upper-bounded by α^α, where $\alpha = |\mathcal{A}|$. For a given partition, with, say $|\Pi'| = m$, we need an assignment of process orders k'_i, $i \in \{1, 2, \ldots, m, \mathrm{w}\}$. If $|A'_i| = 1$, the only valid assignment is $k'_i = 0$, while if $|A'_i| \geq 2$, we must have $k'_i \leq \log \kappa'$. Thus, since $m \leq \alpha$, the number of pairs sought is upper-bounded by $\alpha^\alpha (\log \kappa')^{\alpha+1}$. We notice also that, for $z^n \in B_{\Pi',\mathbf{k}'}$ and sufficiently large n, we must have $\kappa' \leq n$ (actually, $\kappa' = o(n)$), for otherwise the penalty component of $C_{\Pi',\mathbf{k}'}(z^n)$ on its own would surpass $C_{\bar{\Pi},\bar{\mathbf{k}}}(z^n)$, which is $O(n)$. Hence, for sufficiently large n, denoting by $P_{\mathrm{err}}(n)$ the probability of a pair from $\mathbf{\Pi}_2$ prevailing over $(\bar{\Pi}, \bar{\mathbf{k}})$ in (12), and observing that for $n \geq 1$ we have $\alpha^\alpha(\log(n+1))^{\alpha+1} \leq (n+1)^\xi$ with $\xi = \alpha \log \alpha + \alpha + 1$, it follows from (**??**) that

$$\begin{aligned} P_{\mathrm{err}}(n) &\leq \sum_{(\Pi',\mathbf{k}'):\kappa' \geq \kappa_0} \mathcal{P}_{\Pi',\mathbf{k}'} \\ &\leq \sum_{\kappa'=\kappa_0}^{n} \alpha^\alpha \left(\log(n+1)\right)^{\alpha+1} (n+1)^{\beta(\bar{\kappa}-\kappa')+\kappa'} \\ &\leq \sum_{\kappa'=\kappa_0}^{n} (n+1)^{\kappa'(1-\beta)+\beta\bar{\kappa}+\xi} \\ &\leq (n+1)^{\kappa_0(1-\beta)+\beta\bar{\kappa}+\xi+1} , \end{aligned}$$

where the last inequality holds for $\beta > 1$. Choosing

$$\kappa_0 > \frac{\beta\bar{\kappa} + \xi + 2}{\beta - 1} \,,$$

we get

$$P_{\mathrm{err}}(n) \leq (n+1)^\delta \,,$$

for a constant $\delta < -1$. Therefore, $P_{\mathrm{err}}(n)$ is summable, and, applying again Borel-Cantelli's lemma, $(\hat{\Pi}, \hat{\mathbf{k}}) \notin \mathbf{\Pi}_2$ a.s. as $n \to \infty$. We conclude that $(\hat{\Pi}, \hat{\mathbf{k}})$ is compatible with P a.s. as $n \to \infty$, as claimed. The fact that $\hat{\mathbf{k}}$ is, almost surely, the correct order vector follows from the well known consistency of penalized ML estimators for Markov order [**?**] (recall, from the discussion following (12), that the order of each subprocess is estimated independently).

The second claim of the theorem is proved by applying Lemma **??**, which implies that in the domination-free case, the canonical partition beats other compatible partitions with more subalphabets. When $k_{\mathrm{w}}>0$, this follows from (**??**), while when $k_{\mathrm{w}}=0$, it follows from (**??**) and our tie-breaking convention. ■

Acknowledgment. Thanks to Erik Ordentlich and Farzad Parvaresh for stimulating discussions.

References

[1] T. Batu, S. Guha, and S. Kannan, “Inferring mixtures of Markov chains,” in *Computational Learning Theory—COLT*, 2004, pp. 186–199.

[2] N. Landwehr, “Modeling interleaved hidden processes,” in *ICML ’08: Proceedings of the 25th International Conference on Machine Learning*. New York, NY, USA: ACM, 2008, pp. 520–527.

[3] D. Gillblad, R. Steinert, and D. Ferreira, “Estimating the parameters of randomly interleaved Markov models,” in *IEEE International Conference on Data Mining Workshops, 2009. ICDMW ’09.*, Dec. 2009, pp. 308 –313.

[4] G. Schwartz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[5] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.

[6] I. Csiszár and P. C. Shields, “The consistency of the BIC Markov order estimator,” *Annals of Stat.*, vol. 28, pp. 1601–1619, 2000.

[7] Y. M. Shtarkov, “Switching discrete sources and its universal encoding,” *Probl. Inform. Transm.*, vol. 28, no. 3, pp. 95–111, 1992.

[8] R. L. Dobrushin, “Unified methods of information transmission—general case,” *Dokl. Akad. Nauk SSSR*, vol. 163, no. 1, pp. 16–19, 1963.

[9] R. B. Ash, *Information Theory*. John Wiley, 1967.

[10] W. Feller, *Probability theory and its applications*, 3rd ed. New York: John Wiley, 1968, vol. 1.

[11] M. J. Weinberger and M. Feder, “Predictive stochastic complexity and model estimation for finite-state processes,” *Journal of Statistical Planning and Inference*, vol. 39, pp. 353–372, 1994.

[12] D. Blackwell and L. Koopmans, “On the identifiability problem for functions of finite Markov chains,” *Ann. Math. Statist.*, vol. 28, pp. 1011–1015, 1957.

[13] M. Feder, N. Merhav, and M. Gutman, “Universal prediction of individual sequences,” *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1258–1270, Jul. 1992.

[14] R. P. Stanley, *Enumerative Combinatorics*. Cambridge: Cambridge University Press, 1997, vol. 1.

[15] G. Seroussi, W. Szpankowski, and M. J. Weinberger, “Deinterleaving Markov processes via penalized ML,” in *Proc. ISIT’2009*, Seoul, South Korea, June–July 2009, pp. 1739–1743.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: John Wiley & Sons, Inc., 2006.

[17] I. Csiszar, T. M. Cover, and B.-S. Choi, “Conditional limit theorems under Markov conditioning,” *IEEE Trans. Inform. Theory*, vol. 33, no. 6, pp. 788–801, Nov. 1987.

[18] A. Martín, G. Seroussi, and M. J. Weinberger, “Types classes of context trees,” *IEEE Trans. Inform. Theory*, vol. 58, no. 7, pp. 4077–4093, Jul. 2012.

BIOGRAPHIES

Gadiel Seroussi (M'87–SM'91–F'98) was born in Montevideo, Uruguay. He received the B.Sc. degree in electrical engineering, and the M.Sc. and D.Sc. degrees in computer science from Technion – Israel Institute of Technology, Haifa, Israel, in 1977, 1979, and 1981, respectively.

From 1981 to 1987 he was with the faculty of the Computer Science Department at Technion. During the 1982–1983 academic year, he was a Postdoctoral Fellow at the IBM T.J. Watson Research Center, Yorktown Heights, NY. From 1986 to 1988 he was a Senior Research Scientist at Cyclotomics Inc., Berkeley, CA. Since 1988 he has been with Hewlett–Packard Laboratories, Palo Alto, California, where he founded the Information Theory Research Group and was its director until 2005. During the 2005–2006 academic year, he was Associate Director of the Mathematical Sciences Research Institute in Berkeley, California. Since 2004, he has held a joint appointment in Computer Science and Electrical Engineering at Universidad de la República, Montevideo, Uruguay. He is a co-author of the book *Elliptic Curves in Cryptography* (1999), and a co-editor of *Advances in Elliptic Curve Cryptography* (2005), both published by Cambridge University Press. His research interests include the mathematical foundations and practical applications of information theory, error correcting codes, data compression, image processing, and cryptography.

Dr. Seroussi served as Associate Editor for Coding Theory of the IEEE TRANSACTIONS ON INFORMATION THEORY from 2006 to 2009. He is a co-recipient of the 2006 IEEE Joint Communications/Information Theory Paper Award.

Wojciech Szpankowski (F'04) is Saul Rosen Professor of Computer Science and (by courtesy) Electrical and Computer Engineering at Purdue University where he teaches and conducts research in analysis of algorithms, information theory, bioinformatics, analytic combinatorics, random structures, and stability problems of distributed systems. He received his M.S. and Ph.D. degrees in Electrical and Computer Engineering from Gdansk University of Technology.

He held several Visiting Professor/Scholar positions, including McGill University, INRIA, France, Stanford, Hewlett-Packard Labs, Université de Versailles, University of Canterbury, New Zealand, Ecole Polytechnique, France, and the Newton Institute, Cambridge, UK. He is a Fellow of IEEE, and the Erskine Fellow.

In 2010 he received the Humboldt Research Award. In 2001 he published the book "Average Case Analysis of Algorithms on Sequences", John Wiley & Sons, 2001. He has been a guest editor and an editor of technical journals, including THEORETICAL COMPUTER SCIENCE, the ACM TRANSACTION ON ALGORITHMS, the IEEE TRANSACTIONS ON INFORMATION THEORY, FOUNDATION AND TRENDS IN COMMUNICATIONS AND INFORMATION THEORY, COMBINATORICS, PROBABILITY, AND COMPUTING, and ALGORITHMICA. In 2008 he launched the interdisciplinary Institute for Science of Information, and in 2010 he became the Director of the newly established NSF Science and Technology Center for Science of Information.

Marcelo J. Weinberger (M'90–SM'98–F'07) received the Electrical Engineer degree from the Universidad de la República, Montevideo, Uruguay, in 1983, and the M.Sc. and D.Sc. degrees from Technion – Israel Institute of Technology, Haifa, Israel, in 1987 and 1991, respectively, both in electrical engineering.

From 1985 to 1992 he was with the Department of Electrical Engineering at Technion, joining the faculty for the 1991–1992 academic year. During 1992–1993 he was a Visiting Scientist at IBM Almaden Research Center, San Jose, California. Since 1993 he has been with Hewlett-Packard Laboratories, Palo Alto, California, where he is a Distinguished Scientist and leads the Information Theory Research group. His research interests include source coding, sequential decision problems, statistical modeling, and image compression. He is a coauthor of the algorithm at the core of the JPEG-LS lossless image compression standard, and was an editor of the standard specification. He also contributed to the coding algorithm of the JPEG2000 image compression standard.

Dr. Weinberger served as an Associate Editor for Source Coding of the IEEE TRANSACTIONS ON INFORMATION THEORY from 1999 to 2002. He is a co-recipient of the 2006 IEEE Communications/Information Theory Societies Joint Paper Award.