The Complete Realization Problem for Hidden Markov Models: A Survey and Some New Results

M. Vidyasagar
Tata Consultancy Services
No. 1 Software Units Layout, Madhapur
Hyderabad 500 081, INDIA
sagar@atc.tcs.com

March 23, 2009

Abstract

Suppose m is a positive integer, and let $\mathcal{M} := \{1, \ldots, m\}$. Suppose $\{\mathcal{Y}_t\}$ is a stationary stochastic process assuming values in $\mathcal{M}$. In this paper we study the question: When does there exist a hidden Markov model (HMM) that reproduces the statistics of this process? This question is more than forty years old, and as yet no complete solution is available. In this paper, we begin by surveying several known results, and then we present some new results that provide 'almost' necessary and sufficient conditions for the existence of a HMM for a mixing and ultra-mixing process (where the notion of ultra-mixing is introduced here).

In the survey part of the paper, consisting of Sections 2 through 8, we rederive the following known results: (i) Associate an infinite matrix H with the process, and call it a 'Hankel' matrix (because of some superficial similarity to a Hankel matrix). Then the process has a HMM realization *only if* H has finite rank. (ii) However, the finite Hankel rank condition is *not sufficient* in general. There exist processes with finite Hankel rank that do not admit a HMM realization. (iii) An abstract necessary and sufficient condition states that a frequency distribution has a realization as an HMM if and only if it belongs to a 'stable polyhedral' convex set within the set of all frequency distributions on $\mathcal{M}^*$, the set of all finite strings over $\mathcal{M}$. While this condition may be 'necessary and sufficient,' it virtually amounts to a restatement of the problem rather than a solution of it, as observed by Anderson [1]. (iv) Suppose a process has finite Hankel rank, say r. Then there always exists a 'regular quasi-realization' of the process. That is, there exist a row vector, a column vector, and a set of matrices, each of dimension r or $r \times r$ as appropriate, such that the frequency of arbitrary strings is given by a formula that is similar to the corresponding formula

for HMM's. Moreover, *all* quasi-regular realizations of the process can be obtained from one of them via a similarity transformation. Hence, given a finite Hankel-rank process, it is a simple matter to determine whether or not it has a regular HMM in the conventional sense, by testing the feasibility of a linear programming problem. (v) If in addition the process is α-mixing, every regular quasi-realization has additional features. Specifically, a matrix associated with the quasi-realization (which plays the role of the state transition matrix in a HMM) is 'quasi-row stochastic' (in that its rows add up to one, even though the matrix may not be nonnegative), and it also satisfies the 'quasi-strong Perron property' (its spectral radius is one, the spectral radius is a simple eigenvalue, and there are no other eigenvalues on the unit circle). A corollary is that if a finite Hankel rank α-mixing process has a regular HMM in the conventional sense, then the associated Markov chain is irreducible and aperiodic. While this last result is not surprising, it it does not seem to have been stated explicitly. While the above results are all 'known,' they are scattered over the literature; moreover, the presentation here is unified and occasionally consists of relatively simpler proofs than are found in the literature.

Next we move on to present some new results. The key is the introduction of a property called 'ultra-mixing.' The following results are established: (a) Suppose a process has finite Hankel rank, is both α-mixing as well as 'ultra-mixing,' and in addition satisfies a technical condition. Then it has an irreducible HMM realization (and not just a quasi-realization). Moreover, the Markov process underlying the HMM is either aperiodic (and is thus α-mixing), or else satisfies a 'consistency condition.' (b) In the other direction, suppose a HMM satisfies the consistency condition plus another technical condition. Then the associated output process has finite Hankel rank, is α-mixing and is also ultra-mixing. Moreover, it is shown that under a natural topology on the set of HMMs, both 'technical' conditions are indeed satisfied by an open dense set of HMMs. Taken together, these two results show that, modulo two technical conditions, the finite Hankel rank condition, α-mixing, and ultra-mixing are 'almost' necessary and sufficient for a process to have an irreducible and aperiodic HMM.

1 Introduction

1.1 General Remarks

Hidden Markov models (HMM's) were originally introduced in the statistics literature as far back as 1957; see [8, 22]. Subsequently, they were used with partial success in a variety of applications in the engineering world, starting in the late 1970's. Some of these applications include speech processing [35, 27] and source coding. In recent years, HMM's have also been used in some problems in computational biology, such identifying the genes of an organism from its DNA [30, 37, 13] and classifying proteins into a small number of families [29]. The bibliographies of [12, 32] contain many references in this area. In spite of there being so many applications of hidden Markov models, many of the underlying statistical questions remain unanswered. The aim of this paper is to address some of these issues.

Without going into details (which are given in Section 2), the problem under study can be stated as follows. Suppose m is a positive integer and let $\mathcal{M} := \{1, \ldots, m\}$. Suppose $\{\mathcal{Y}_t\}$ is a stationary stochastic process assuming values in $\mathcal{M}$. We are interested in the following kinds of questions:

1. Suppose the complete statistics of the process $\{\mathcal{Y}_t\}$ are known. Under what conditions is it possible to construct a hidden Markov model (HMM) for this process? This is the most general question and is referred to as the 'complete' realization problem.

2. Much of the complexity of the complete realization problem stems from the requirement that various vectors and matrices must have nonnegative entries. Is it possible to construct at least a 'quasi' HMM for the process by dropping the nonnegativity requirement? If so, what properties does such a quasi-realization have?

3. How can one construct a 'partial realization' for the process, that faithfully reproduces the statistics of the process only up to some finite order?

4. Suppose one has access not to the entire statistics of the process, but merely several sample paths, each of finite length. How can one compute approximations to the true statistics of the process on the basis of these observations, and what is the confidence one has in the accuracy of these estimates?

5. Suppose one has constructed a partial realization of the process on the basis of a finite length sample path. How are the accuracy and confidence in the estimates of the

statistics translated into accuracy and confidence estimates on the parameters in the model?

Ideally, we would like to be able to say something about *all* of these questions. In a 'practical' application, the last three questions are the ones to which we would most like to have an answer. However, these are also the most difficult questions to answer. In this paper, we provide nearly complete answers to the first two questions. In a companion paper, we provide nearly complete answers to the remaining three questions.

The subject of hidden Markov models (HMM's) is more than forty years old. Section 2 contains a detailed historical review, but for the purposes of an introduction, the situation can be summarized as follows:

1. Associate an infinite matrix H with the process. This matrix is usually called a 'Hankel' matrix (because of some superficial similarity to a Hankel matrix). Then the process has a HMM realization *only if* H has finite rank. Such processes can be referred to as 'finite Hankel rank' processes.

2. The converse is not true in general: There exist processes with finite Hankel rank that do not have a HMM realization.

3. If the process has finite Hankel rank (meaning that H has finite rank), and if in addition the process is α-mixing, then a *sampled version* of the process has a HMM. But in general, even with the α-mixing assumption, the full process need not have a HMM realization.

4. It is possible to give an abstract 'necessary and sufficient' condition for a given frequency distribution to have a HMM realization. However, as remarked by Anderson [1], this condition is more a restatement of the problem than a solution of it.

5. Suppose a process has finite Hankel rank. Then there always exists a 'quasi-realization' of the process. That is, there exist a row vector, a column vector, and a set of matrices, together with a formula for computing the frequencies of arbitrary strings that is similar to the corresponding formula for HMM's. Moreover, the quasi-realization can be chosen to be 'regular,' in the sense that the size of the 'state space' in the quasi-realization can always be chosen to equal the rank of the Hankel matrix. Hence every finite Hankel rank stochastic process has a 'regular quasi-realization,' whether or not it has a regular realization. Further, two different regular quasi-realizations of the same

process are related through a similarity transformation. Hence, given a finite Hankel-rank process, it is a simple matter to determine whether or not it has a regular HMM in the conventional sense, by testing the feasibility of a linear programming problem.

6. Suppose that in addition the process to be modelled is α-mixing.[1] In this case, every regular quasi-realization has additional features. Specifically, a matrix associated with the quasi-realization (which plays the role of the state transition matrix in a HMM) is 'quasi-row stochastic' (in that its rows add up to one, even though the matrix may not be nonnegative), and it also satisfies the 'quasi-strong Perron property' (its spectral radius is one, the spectral radius is also an eigenvalue, and there are no other eigenvalues on the unit circle). A corollary is that if a finite Hankel rank α-mixing process has a regular HMM in the conventional sense, then the associated Markov chain is irreducible and aperiodic. While this last result is not surprising, it does not seem to have been stated explicitly.

7. Assuming beforehand that the process under study is generated by an irreducible (but otherwise unknown) HMM that satisfies a few other technical conditions, it is possible to give a synthesis procedure that produces another irreducible HMM.

In the 'survey' part of the paper consisting of Sections 2 through 8, we rederive many of the above results. Though the rederived results are 'known,' the proofs given here are in some cases simpler than those in the original papers. Moreover, many of the relevant results are collected in one place for the convenience of the reader.

Then we move on to the new results. A property called 'ultra-mixing' is introduced, and it plays a crucial role in the study of HMMs. Ultra-mixing is also a kind of long-term asymptotic independence, which neither implies nor is implied by α-mixing. With the new notion in place, two results are established. First, suppose a process has finite Hankel rank, is both α-mixing as well as 'ultra-mixing,' and in addition satisfies a technical condition. Then it has an irreducible HMM realization (not just a quasi-realization). Moreover, the Markov process underlying the HMM is either aperiodic (and is thus α-mixing), or else satisfies a 'consistency condition.' In the other direction, suppose a HMM satisfies the consistency condition plus another technical condition. Then the associated output process has finite Hankel rank, is α-mixing and is also ultra-mixing.

[1] A precise definition of α-mixing is given in Section 8. In simple terms, α-mixing is a kind of long-term asymptotic independence. Thus a process $\{\mathcal{Y}_t\}$ is α-mixing if $\mathcal{Y}_t$ and $\mathcal{Y}_{t+k}$ are 'nearly' independent for k 'sufficiently large.'

Finally, we tackle the question of just how 'technical' the two technical conditions really are. Using a very natural topology on the set of HMMs, it is shown that both of the 'technical' conditions are satisfied by an *open dense set* of HMMs. Thus in some sense 'nearly all' HMMs satisfy these conditions. Taken together, these two results show that, modulo two technical conditions, the finite Hankel rank condition, α-mixing, and ultra-mixing are 'almost' necessary and sufficient for a process to have an irreducible and aperiodic HMM. Thus the results presented here are tantamount to nearly necessary and sufficient conditions for the existence of a HMM, for processes that satisfy appropriate mixing conditions.

1.2 Nature of Contributions of the Present Paper

Now an attempt is made to explain the contribution of the present paper. The basic ideas of HMM realization theory are more than forty years old. The fact that a stochastic process has to have finite Hankel rank in order to have a HMM was established by Gilbert [22], though his notation was slightly different. Dharmadhikari [14] gives an example of a process that has finite Hankel rank, but does not have a regular HMM realization. Fox and Rubin [21] extend the argument by presenting an example of a finite Hankel rank process that does not have a HMM realization at all, regular or otherwise. In [18], Dharmadhikari and Nadkarni simplify the example of Fox and Rubin and also, in the opinion of the present author, correct an error in the Fox-Rubin paper. A sufficient condition for the existence of a HMM, involving the existence of a suitable polyhedral cone, was established by Dharmadhikari [15]. An abstract necessary and sufficient condition is given in [23] and the argument is considerably streamlined in [34].

So what can possibly be new forty years later? In [1], Anderson says that "The use of a cone condition, described by some as providing a solution to the realization problem, constitutes (in this author's opinion) a restatement of the problem than a solution of it. This is because the cone condition is encapsulated by a set of equations involving unknowns; there is no standard algorithm for checking the existence of a solution or allowing construction of a solution;" In other words, the original 'solution' given in [23] is no solution at all in the opinion of Anderson (and also in the opinion of the present author). He then proceeds to give sufficient conditions for the existence of a suitable cone, as well as a procedure for constructing it. However, in order to do this he *begins* with the assumption that the process under study has a HMM; see Assumption 1 on p. 84 of [1]. As a consequence, some of the proofs in that paper make use of the properties of the unknown but presumed to exist HMM realization.

In contrast, in the present paper the objective is to *state all conditions only in terms of the process under study, and nothing else*. This objective is achieved. Given this background, it is hardly surprising that many of the theorems and proofs of the present paper bear a close resemblance to their counterparts in [1]. Indeed, it would be accurate to say that the present paper, to a very large extent, represents a reworking of the arguments in [1] while rephrasing any conditions that cannot be directly expressed in terms of the process under study. The one major departure from [1] is the delineation of a property referred to here as 'ultra-mixing.' This property is *proved* as a consequence of the various assumptions in [1]; see Theorem 6. Here the property is *a part of the assumptions on the process*. It turns out that 'ultra-mixing' has been introduced to the statistics literature by Kalikow [28] under the name of the 'uniform martingale' property. Kalikow also shows that ultra-mixing is equivalent to another property that he calls 'random Markov' property. While this connection is interesting, this fact by itself does not assist us in constructing a HMM for a stationary process. In the final theorem that establishes the existence of a HMM, we make a technical assumption about the behaviour of the cluster points of a countable set, that is suggested by a similar technical condition in the positive realization literature. See [6], Theorem 11.

Thus, in summary, the present paper pulls together several existing ideas in the literature, and gives conditions that are 'almost' necessary and sufficient for the existence of an aperiodic and irreducible HMM for a given process. Moreover, so far as the author has been able to determine, this is the first paper wherein all the requisite conditions for the existence of a HMM are stated solely in terms of the process under study.

The realization problem for processes that do not satisfy any kind of mixing properties is still open. However, in the opinion of the present author, this problem is virtually impossible to tackle – in the absence of any kind of mixing assumptions, there is simply far too much anomalous behaviour possible to permit the development of any kind of coherent realization theory.

2 Historical Review

The historical beginning of the study of HMM's can be said to be the papers by Blackwell and Koopmans [8] and Gilbert [22]. Blackwell and Koopmans study the case where a stationary process $\{\mathcal{Y}_t\}$ is a function of a finite Markov chain, and ask when the underlying Markov chain can be uniquely identified. Gilbert considered the case where $\{\mathcal{X}_t\}_{t\geq 0}$ is a Markov chain

assuming values in a state space X, and observed that if $f : X \to \mathbb{R}$ is some function, then the stochastic process $\{f(\mathcal{X}_t)\}$ need not be Markov. He then asked the question as to when a given stationary stochastic process $\{\mathcal{Y}_t\}_{t\geq 0}$ over a finite set $\mathcal{M} := \{1, \ldots, m\}$ can be realized as $\{f(\mathcal{X}_t)\}$ where $\{\mathcal{X}_t\}_{t\geq 0}$ is a Markov chain over a finite state space $\mathcal{N} := \{1, \ldots, n\}$. For each output state $u \in \mathcal{M}$, he defined an integer $n(u)$ which he called the rank of the variable u, and showed that *if* $\{\mathcal{Y}_t\}$ is a function of a finite-state Markov chain, then $n(u)$ is finite for each $u \in \mathcal{M}$, and moreover, $\sum_{u\in\mathcal{M})} n(u) \leq n$, where n is the size of the state space of the underlying Markov chain.[2] He conjectured that the condition $\sum_{u\in\mathcal{M})} n(u) =: s < \infty$ is also sufficient for the given stationary process $\{\mathcal{Y}_t\}$ to be a function of a finite-state Markov chain. Further, he defined the process $\{\mathcal{Y}_t\}$ to be a **regular** function of a finite-state Markov chain if the state space of the Markov chain has dimension s. He then went on to study the problem of identifying the underlying Markov chain, *assuming* that in fact the process under study was a function of a Markov chain.

Subsequently, in a series of definitive papers, Dharmadhikari shed considerable light on this question. He first showed [14] that if the process $\{\mathcal{Y}_t\}_{t\geq 0}$ has finite Hankel rank and is α-mixing, then there exists an integer r such that the 'sampled' process $\{\mathcal{Y}_{rt}\}_{t\geq 0}$ has a HMM realization. However, in general the integer r cannot be chosen as one, meaning that the original 'unsampled' process may or may not have a HMM realization. In another paper [17], he showed that if the process $\{\mathcal{Y}_t\}_{t\geq 0}$ has finite Hankel rank *and is exchangeable*, then it has a HMM realization. Since an exchangeable process is in some sense 'maximally non-mixing,' this result is quite counter-intuitive when posited against that of [14]. In yet another paper [15], he postulated a 'cone condition,' and showed that if a process satisfies the cone condition in addition to having finite Hankel rank, then it has a HMM realization. In [16], he showed that the class of processes that satisfy the cone condition is strictly larger than the class of processes having a regular HMM realization.

Fox and Rubin [21] showed that the conjecture of Gilbert is false in general, by giving an example of a process over a countable state space having the finite Hankel rank property, whereby the underlying Markov chain cannot be over a finite state space. This example was simplified by Dharmadhikari and Nadkarni [18], and in the opinion of the present author, an error in the Fox-Rubin example was corrected.

If an infinite matrix has finite rank, it is clear that the elements of that matrix must satisfy

[2] In the statistics literature, it is common to refer to the set $\mathcal{M}$ as the 'state space.' Here we stick to the convention in the engineering literature, and refer to the range of $\mathcal{Y}$ as the output space, and the range of $\mathcal{X}$ as the state space.

various recursive relationships. In [19], these recursive relationships are expressed in the form of what we call here a 'quasi-realization,' which is like a HMM realization except that we do not require all the vectors and matrices to be nonnegative. In [26], these relationships are studied further.

In [23], an abstract 'necessary and sufficient condition' is given to the effect that a frequency distribution has a HMM realization if and only if it belongs to a stable polyhedral convex set within the set of frequency distributions on $\mathcal{M}^*$, the set of all finite strings on $\mathcal{M}$. The original proof is very difficult to read, but a highly readable proof is given in [34].

Thus much of the fundamental work in this area was done nearly four decades ago. Subsequent work has mostly refined the notation and/or clarified the arguments, but there has not been much progress on improving on the cone condition as a sufficient condition for the existence of a HMM.

In a recent paper, Anderson [1] *starts* with the assumption that the process at hand has a HMM realization with an irreducible state transition matrix, and then gives a constructive procedure for constructing a HMM realization where the underlying matrix A is irreducible. Thus the paper of Anderson contains assumptions that cannot be directly stated in terms of the properties of the output process $\{\mathcal{Y}_t\}$. Moreover, there is no guarantee that the HMM constructed using his procedure has the same sized state space as the one that generated the process under study. But the ideas put forward in that paper are very useful in proving the results presented here; see Section 9.

Before concluding this historical review, we mention also the so-called 'positive realization problem' from control theory, which has a close relationship to the cone condition. Consider the linear recursion

$$x_{t+1} = Ax_t + Bu_t,\ y_t = Cx_t$$

defined over the real number system. Thus $x_t \in \mathbb{R}^n$, $y_t, u_t \in \mathbb{R}$, and the matrices A, B, C have dimensions $n \times n, n \times 1$ and $1 \times n$ respectively. If the input sequence $\{u_t\}$ equals the 'unit pulse' sequence $\{1, 0, 0, \ldots\}$, then the corresponding output sequence $\{y_t\}$, known as the 'unit pulse response,' equals

$$y_0 = 0, y_{t+1} = CA^tB\ \forall t \geq 0.$$

Define $h_t := CA^tB$. Suppose A, B, C are all nonnegative matrices/vectors. Then clearly $h_t \geq 0\ \forall t$. The positive realization problem is the converse: Suppose $\{h_t\}_{t\geq 0}$ is a nonnegative sequence, and that the z-transform of this sequence is a rational function. When do there exist *nonnegative* matrices A, B, C such that $h_t = CA^tB\ \forall t$? Some results on this problem

can be found in [25, 2, 24]. See [6] for a review of the current status of this problem. Some ideas from positive realization theory are also used in constructing HMM's; see Section 9.

3 Equivalence Between Several Stochastic Models

In this section, we consider three distinct-looking definitions of hidden Markov models that are prevalent in the literature, and show that they are all equivalent when it comes to expressive power. In other words, if a stationary stochastic process over a finite alphabet has any one of the three kinds of HMM, then it has all three kinds of HMM. However, the size of the state space is in general different in the three types of HMM's. In this respect, the 'joint Markov process' definition of a HMM found in [1] is the most economical in terms of the size of the state space, while the 'deterministic function of a Markov process' definition introduced in [22] of a HMM is the least economical.

Let us begin by introducing the three distinct types of stochastic models. The first model was originally introduced by Gilbert [22] in the paper that began the development of HMM theory, and is quite popular in the statistics community.

Definition 1 *Suppose $\{\mathcal{Y}_t\}$ is a stationary stochastic process assuming values in a finite set $\mathcal{M}$. We say that $\{\mathcal{Y}_t\}$ has a* **HMM of the deterministic function of a Markov chain type** *if there exist a Markov process $\{\mathcal{X}_t\}$ assuming values in a finite set $\mathcal{N} := \{1, \ldots, n\}$ and a function $f : \mathcal{N} \to \mathcal{M}$ such that $\mathcal{Y}_t = f(\mathcal{X}_t)$.*

Note that the existence of a HMM becomes an issue only if one insists on a *finite* state space. If one allows a state space of infinite cardinality, then one can always construct a HMM (of the deterministic function of a Markov chain type), One begins with the set of *all* strings (not necessarily of finite length) over $\mathcal{M}$ as the state space; this is an uncountable set. However, by taking equivalence classes it is possible to make the state space into a countable set; see [10] for details.

Second, we introduce a model that is very popular in the engineering community. It appears to have been first introduced in [4].

Definition 2 *Suppose n is a finite integer. Then we say that $\{\mathcal{Y}_t\}$ has a* **HMM of the random function of a Markov chain type** *if there exist an integer n, and a pair of matrices $A \in [0,1]^{n \times n}$ and $B \in [0,1]^{n \times m}$ such that the properties hold:*

1. *A is row stochastic; that is, the sum of every row of A equals one.*

2. *B is row stochastic; that is, the sum of every row of B equals one.*

3. *Choose $\pi \in [0,1]^n$ to be a row eigenvector of A corresponding to the eigenvalue one, such that its entries add up to one.* [3] *Suppose $\{\mathcal{X}_t\}$ is a Markov chain assuming values in $\mathcal{N} := \{1, \ldots, n\}$ with state transition matrix A and initial distribution π. Suppose $\mathcal{Z}_t$ is selected at random from $\mathcal{M}$ according to the law*

$$\Pr\{\mathcal{Z}_t = u | \mathcal{X}_t = j\} = b_{ju}.$$

Then this process $\{\mathcal{Z}_t\}$ has the same law as $\{\mathcal{Y}_t\}$.

In such a case we refer to A and B as the state transition matrix and output matrix, respectively, of the HMM.

Finally, we introduce a definition that is used in [1]. The antecedents of this definition are not clear. However, for the purposes of various proofs, it is the most convenient one. Moreover, as shown in Lemma 3.2 below, it is also the most economical in terms of the size of the state space.

Definition 3 *Suppose $\{\mathcal{Y}_t\}$ is a stationary stochastic process on the finite alphabet $\mathcal{M} := \{1, \ldots, m\}$. We say that the process $\{\mathcal{Y}_t\}$ has a* **HMM of the 'joint Markov process' type** *if there exists another stationary stochastic process $\{\mathcal{X}_t\}$ over a finite state space $\mathcal{N} := \{1, \ldots, n\}$ such that the following properties hold:*

1. *The joint process $\{(\mathcal{X}_t, \mathcal{Y}_t\}$ is Markov. Hence*

$$\Pr\{(\mathcal{X}_t, \mathcal{Y}_t) | \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}, \mathcal{X}_{t-2}, \mathcal{Y}_{t-2}, \ldots\} = \Pr\{(\mathcal{X}_t, \mathcal{Y}_t) | \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}\}. \tag{3.1}$$

2. *In addition, it is true that*

$$\Pr\{(\mathcal{X}_t, \mathcal{Y}_t) | \mathcal{X}_{t-1}, \mathcal{Y}_{t-1}\} = \Pr\{(\mathcal{X}_t, \mathcal{Y}_t) | \mathcal{X}_{t-1}\}. \tag{3.2}$$

From the definition, it is clear that

$$\Pr\{\mathcal{X}_t | \mathcal{X}_{t-1}, \mathcal{X}_{t-2}, \ldots\} = \Pr\{\mathcal{X}_t | \mathcal{X}_{t-1}\}.$$

In other words, $\{\mathcal{X}_t\}$ by itself is a Markov process. Let us define the $n \times n$ matrices $M^{(u)}, u \in \mathcal{M}$ as follows:

$$m_{ij}^{(u)} := \Pr\{\mathcal{X}_t = j \& \mathcal{Y}_t = u | \mathcal{X}_{t-1} = i\}. \tag{3.3}$$

[3] Note that, by [7], Theorem 1.1, p. 26, such an invariant probability vector always exists. However, unless additional conditions are imposed on A, π is not unique in general.

Next, let us define

$$a_{ij} := \sum_{u \in \mathcal{M}} m_{ij}^{(u)}, \ \forall i, j. \tag{3.4}$$

Then it is clear that the state transition matrix of the Markov process $\{\mathcal{X}_t\}$ is precisely A. Moreover, the condition(3.2) also implies that both $\mathcal{X}_t$ and $\mathcal{Y}_t$ are 'random functions' of the previous state $\mathcal{X}_{t-1}$. We say that the HMM is irreducible or primitive if the state transition matrix of the process $\{\mathcal{X}_t\}$ is irreducible or primitive.

Now it is shown that all these models are equivalent.

Lemma 3.1 *The following statements are equivalent:*

(i) The process $\{\mathcal{Y}_t\}$ has a HMM of the deterministic function of a Markov chain type.

(ii) The process $\{\mathcal{Y}_t\}$ has a HMM of the random function of a Markov chain type.

(iii) The process $\{\mathcal{Y}_t\}$ has a HMM of the joint Markov process type.

Proof: (i) $\Rightarrow$ (ii) Clearly every deterministic function of a Markov chain is also a 'random' function of the same Markov chain, with every element of B equal to zero or one. Precisely, since both $\mathcal{N}$ and $\mathcal{M}$ are finite sets, the function f simply induces a partition of the state space $\mathcal{N}$ into m subsets $\mathcal{N}_1, \ldots, \mathcal{N}_m$, where $\mathcal{N}_u := \{j \in \mathcal{N} : f(j) = u\}$. Thus two states in $\mathcal{N}_u$ are indistinguishable through the measurement process $\{\mathcal{Y}_t\}$. Now set $b_{ju} = 1$ if $j \in \mathcal{N}_u$ and zero otherwise.

(ii) $\Rightarrow$ (iii) If $\{\mathcal{Y}_t\}$ is modelled as a random function of a Markov process HMM with $\{\mathcal{X}_t\}$ as the underlying Markov chain, then the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov. Indeed, if we define $(\mathcal{X}_t, \mathcal{Y}_t) \in \mathcal{N} \times \mathcal{M}$, then it readily follows from the HMM conditions that

$$\Pr\{(\mathcal{X}_{t+1}, \mathcal{Y}_{t+1}) = (j, u) | (\mathcal{X}_t, \mathcal{Y}_t) = (i, v)\} = a_{ij} b_{ju},$$

Now define

$$M^{(u)} := [a_{ij} b_{ju}] \in [0, 1]^{n \times n}.$$

Then the process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is Markov, and its state transition matrix is given by

$$\begin{bmatrix} M^{(1)} & M^{(2)} & \ldots & M^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ M^{(1)} & M^{(2)} & \ldots & M^{(m)} \end{bmatrix}.$$

Finally, note that the probability that $(\mathcal{X}_{t+1}, \mathcal{Y}_{t+1}) = (j, u)$ depends only on $\mathcal{X}_t$ but not on $\mathcal{Y}_t$. Hence the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ satisfies all the conditions required of the joint Markov process HMM model.

(iii) $\Rightarrow$ **(i)** Suppose $\mathcal{X}_t$ is a Markov process such that the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ is also Markov. Then clearly $\mathcal{Y}_t = f[(\mathcal{X}_t, \mathcal{Y}_t)]$ for a suitable function f. Hence this is also a HMM of the deterministic function of a Markov chain type. ■

It is easy to verify that the above lemma remains valid if we add the requirement that the associated Markov process $\{\mathcal{X}_t\}$ is irreducible. In other words, a process $\{\mathcal{Y}_t\}$ is a function of an irreducible Markov chain, if and only if $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$ where $\{\mathcal{X}_t\}$ is an irreducible Markov chain, if and only if $\mathcal{Y}_t$ is a random function of $\mathcal{Z}_t$ where $\{\mathcal{Z}_t\}$ is an irreducible Markov chain.[4]

Up to now we have considered only the 'expressive power' of the various HMM types. However, this is only part of the problem of stochastic modelling. An equally, if not more, important issue is the 'economy' of the representation, that is, the number of states in the underlying Markov chain. Clearly, given a sample path of finite length, the fewer the number of parameters that need to be estimated, the more confidence we can have in the estimated parameter values. The next lemma summarizes the situation and shows that, so far as the economy of the representation is concerned, the joint Markov process model is the most economical.

Lemma 3.2 *(i) Suppose a process $\{\mathcal{Y}_t\}$ has a HMM of the random function of a Markov chain type, and let $\{\mathcal{X}_t\}$ denote the associated Markov chain. Let A and B denote respectively the state transition matrix and output matrix of the HMM. Then $\mathcal{Y}_t$ is a deterministic function of $\mathcal{X}_t$ if and only if every row of the matrix B contains one* 1 *and the remaining elements are zero.*

(ii) Suppose a process $\{\mathcal{Y}_t\}$ has a HMM of the joint Markov process type, and let $\{\mathcal{X}_t\}$ denote the associated Markov chain. Define the matrices $M^{(u)}$ as in (3.3), Then $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$ (and not just $\mathcal{X}_{t-1}$) if and only if the following consistency conditions hold: Define

$$a_{ij} := \sum_{u \in \mathcal{M}} m_{ij}^{(u)}, 1 \leq i, j \leq n.$$

If $a_{ij} \neq 0$, then the ratio

$$\frac{m_{ij}^{(u)}}{a_{ij}}$$

[4] All Markov chains must have finite state spaces.

is independent of i.

Proof: The first statement is obvious. Let us consider the second statement. Suppose the process $\{\mathcal{Y}_t\}$ has a joint Markov process type of HMM, and let $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ be the associated Markov process. Define the matrices $M^{(u)}$ as in (3.3). Then we already know that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_{t-1}$. The aim is to show that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$ (and not just $\mathcal{X}_{t-1}$) if and only if the stated condition holds.

'Only if': From (3.4), we know that the state transition matrix of the process $\{\mathcal{X}_t\}$ is given by $A = \sum_{u \in \mathcal{M}} M^{(u)}$. Now suppose that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$, and not just $\mathcal{X}_{t-1}$, and define

$$b_{ju} := \Pr\{\mathcal{Y}_t = u | \mathcal{X}_t = j\}, \ \forall u \in \mathcal{M}, j \in \mathcal{N}.$$

Then we must have $m_{ij}^{(u)} = a_{ij} b_{ju}$ for all i, j, u. If $a_{ij} = 0$ for some i, j, then perforce $m_{ij}^{(u)} = 0 \ \forall u \in \mathcal{M}$. Suppose $a_{ij} \neq 0$. Then it is clear that

$$b_{ju} = \frac{m_{ij}^{(u)}}{a_{ij}} \ \forall i$$

and is therefore independent of i.

'If': This consists of simply reversing the arguments. Suppose the ratio is indeed independent of i, and define b_{ju} as above. Then clearly $m_{ij}^{(u)} = a_{ij} b_{ju}$ and as a result $\mathcal{Y}_t$ is a random function of $\mathcal{X}_t$. ■

As a simple example, suppose $n = m = 2$,

$$M^{(1)} = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.4 \end{bmatrix}, M^{(2)} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}, A = \begin{bmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{bmatrix}.$$

Then

$$\frac{m_{11}^{(1)}}{a_{11}} = 5/7, \ \frac{m_{21}^{(1)}}{a_{21}} = 1/2 \neq 5/7.$$

Since the ratio $m_{ij}^{(u)}/a_{ij}$ fails to be independent of i for the choice $j = 1, u = 1$, it follows that $\mathcal{Y}_t$ is a random function of $\mathcal{X}_{t-1}$ but not a random function of $\mathcal{X}_t$.

4 Preliminaries

4.1 The Hankel Matrix

Some terminology is introduced to facilitate subsequent discussion.

Given an integer l, the set $\mathcal{M}^l$ consists of l-tuples. These can be arranged either in first-lexical order (flo) or last-lexical order (llo). First-lexical order refers to indexing the first element, then the second, and so on, while last-lexical order refers to indexing the last element, then the next to last, and so on. For example, suppose $m = 2$ so that $\mathcal{M} = \{1, 2\}$. Then

$$\mathcal{M}^3 \text{ in llo } = \{111, 112, 121, 122, 211, 212, 221, 222\},$$

$$\mathcal{M}^3 \text{ in flo } = \{111, 211, 121, 221, 112, 212, 122, 222\}.$$

Given any finite string $\mathbf{u} \in \mathcal{M}^*$, we can speak of its frequency $f_{\mathbf{u}}$. Thus, if $|\mathbf{u}| = l$ and $\mathbf{u} = u_1 \dots u_l$, we have

$$f_{\mathbf{u}} := \Pr\{(\mathcal{Y}_{t+1}, \mathcal{Y}_{t+2}, \dots, \mathcal{Y}_{t+1}) = (u_1, u_2, \dots, u_l)\}.$$

Since the process $\{\mathcal{Y}_t\}$ is assumed to be stationary, the above probability is independent of t.

Note the following fundamental properties of the frequency $f_{\mathbf{u}}$.

$$f_{\mathbf{u}} = \sum_{v \in \mathcal{M}} f_{\mathbf{u}v} = \sum_{w \in \mathcal{M}} f_{w\mathbf{u}}, \ \forall \mathbf{u} \in \mathcal{M}^*. \tag{4.1}$$

More generally,

$$f_{\mathbf{u}} = \sum_{\mathbf{v} \in \mathcal{M}^r} f_{\mathbf{uv}} = \sum_{\mathbf{w} \in \mathcal{M}^s} f_{\mathbf{wu}}, \ \forall \mathbf{u} \in \mathcal{M}^*, \tag{4.2}$$

where as usual $\mathcal{M}^*$ denotes the set of all strings of finite length over $\mathcal{M}$. These properties are known as 'right-consistency' and 'left-consistency' respectively.

Given integers $k, l \geq 1$, the matrix $F_{k,l}$ is defined as

$$F_{k,l} = [f_{\mathbf{uv}}, \mathbf{u} \in \mathcal{M}^k \text{ in flo}, \mathbf{v} \in \mathcal{M}^l \text{ in llo}] \in [0, 1]^{m^k \times m^l}.$$

Thus the rows of $F_{k,l}$ are indexed by an element of $\mathcal{M}^k$ in flo, while the columns are indexed by an element of $\mathcal{M}^l$ in llo. For example, suppose $m = 2$. Then

$$F_{1,2} = \begin{bmatrix} f_{111} & f_{112} & f_{121} & f_{122} \\ f_{211} & f_{212} & f_{221} & f_{222} \end{bmatrix},$$

whereas

$$F_{2,1} = \begin{bmatrix} f_{111} & f_{112} \\ f_{211} & f_{212} \\ f_{121} & f_{122} \\ f_{221} & f_{222} \end{bmatrix}.$$

In general, for a given integer s, the matrices $F_{0,s}, F_{1,s-1}, \ldots, F_{s-1,1}, F_{s,0}$ all contain frequencies of the m^s s-tuples. However, the dimensions of the matrices are different, and the elements are arranged in a different order. Note that by convention $F_{0,0}$ is taken as the 1×1 matrix 1 (which can be thought of as the frequency of occurence of the empty string).

Given integers $k, l \geq 1$, we define the matrix $H_{k,l}$ as

$$H_{k,l} := \begin{bmatrix} F_{0,0} & F_{0,1} & \ldots & F_{0,l} \\ F_{1,0} & F_{1,1} & \ldots & F_{1,l} \\ \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \ldots & F_{k,l} \end{bmatrix}.$$

Note that $H_{k,l}$ has $1 + m + \ldots + m^k$ rows, and $1 + m + \ldots + m^l$ columns. In general, $H_{k,l}$ is not a 'true' Hankel matrix, since it is not constant along backward diagonals. It is not even 'block Hankel.' However, it resembles a Hankel matrix in the sense that the matrix in the (i, j)-th block consists of frequencies of strings of length $i + j$. Finally, we define H (without any subscripts) to be the infinite matrix of the above form, that is,

$$H := \begin{bmatrix} F_{0,0} & F_{0,1} & \ldots & F_{0,l} & \ldots \\ F_{1,0} & F_{1,1} & \ldots & F_{1,l} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_{k,0} & F_{k,1} & \ldots & F_{k,l} & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Through a mild abuse of language we refer to H as the Hankel matrix associated with the process $\{\mathcal{Y}_t\}$.

4.2 A Necessary Condition for the Existence of Hidden Markov Models

In this section, it is shown that a process $\{\mathcal{Y}_t\}$ has a HMM *only if* the matrix H has finite rank. However, the finiteness of the rank of H is only necessary, but not sufficient in general.

Theorem 4.1 *Suppose $\{\mathcal{Y}_t\}$ has a 'joint Markov process' type of HMM with the associated $\{\mathcal{X}_t\}$ process having n states. Then Rank$(H) \leq n$.*

Proof: As shown in Section 3, the process $\{\mathcal{X}_t\}$ is Markov. Moreover, since it is a stationary process, it must have a stationary distribution π that satisfies $\pi = \pi A$, where the

$n \times n$ matrix A is defined in (3.4).[5] Note that $f_u = \Pr\{\mathcal{Y}_1 = u\}$, for every $u \in \mathcal{M}$. Since the state $\mathcal{X}_0$ is distributed according to π, it follows that

$$f_u = \Pr\{\mathcal{Y}_1 = u\} = \sum_{i=1}^{n} \sum_{j=1}^{n} \Pr\{\mathcal{X}_1 = j \& \mathcal{Y}_1 = u | \mathcal{X}_0 = i\} \cdot \pi_i.$$

From the definition of the matrices $M^{(u)}$, it follows that

$$f_u = \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i m_{ij}^{(u)}.$$

Here the first summation is over the initial state $\mathcal{X}_0$ and the second summation is over the subsequent state $\mathcal{X}_1$. This relationship can be expressed compactly as

$$f_u = \pi M^{(u)} \mathbf{e}_n.$$

More generally, let $\mathbf{u} \in \mathcal{M}^l$. Suppose to be specific that $\mathbf{u} = u_1 \dots u_l$. Then an easy generalization of the preceding argument shows that

$$f_{\mathbf{u}} = \sum_{i=1}^{n} \sum_{j_1=1}^{n} \dots \sum_{j_l=1}^{n} \pi_i m_{ij_1}^{(u_1)} \cdots m_{j_{l-1}j_l}^{(u_l)} = \pi M^{(u_1)} \cdots M^{(u_l)} \mathbf{e}_n. \tag{4.3}$$

Note that

$$\sum_{l \in \mathcal{M}} M^{(l)} = A, \ \pi \left[\sum_{l \in \mathcal{M}} M^{(l)} \right] = \pi, \text{ and } \left[\sum_{l \in \mathcal{M}} M^{(l)} \right] \mathbf{e}_n = \mathbf{e}_n. \tag{4.4}$$

Thus the sum of the matrices $M^{(u)}$ is the state transition matrix of the Markov chain, and π and $\mathbf{e}_n$ are respectively a row eigenvector and a column eigenvector of A corresponding to the eigenvalue 1.

Now let us return to the matrix H. Using (4.3), we see at once that H can be factored as

$$H = \begin{bmatrix} \pi \\ \pi M^{(1)} \\ \vdots \\ \pi M^{(m)} \\ \pi M^{(1)} M^{(1)} \\ \vdots \\ \pi M^{(m)} M^{(m)} \\ \vdots \end{bmatrix} [\mathbf{e}_n \mid M^{(1)} \mathbf{e}_n | \dots | M^{(m)} \mathbf{e}_n \mid M^{(1)} M^{(1)} \mathbf{e}_n | \ \dots | \ M^{(m)} M^{(m)} \mathbf{e}_n | \dots].$$

[5] Since we are *not* assuming that A is irreducible, A may have more than one stationary distribution. Hence the relation $\pi = \pi A$ need not determine π uniquely.

In other words, the rows consist of $\pi M^{(u_1)} \cdots M^{(u_l)}$ as $\mathbf{u} \in \mathcal{M}^l$ is in flo and l increases, whereas the columns consist of $M^{(u_1)} \cdots M^{(u_l)} \mathbf{e}_n$ as $\mathbf{u} \in \mathcal{M}^l$ is in llo and l increases. Now note that the first factor has n columns whereas the second factor has n rows. Hence $\text{Rank}(H) \leq n$. ■

We conclude this subsection by recalling a negative result of Sontag [39], in which he shows that the problem of deciding whether or not a given 'Hankel' matrix has finite rank is undecidable.

5 Non-Sufficiency of the Finite Hankel Rank Condition

Let us refer to the process $\{\mathcal{Y}_t\}$ as 'having finite Hankel rank' if $\text{Rank}(H) < \infty$. Thus Theorem 4.1 shows that $\text{Rank}(H)$ being finite is a *necessary* condition for the given process to have a HMM. However, the converse is *not true in general* – it is possible for a process to have finite Hankel rank and yet not have a realization as a HMM. The original example in this direction was given by Fox and Rubin [21]. However, their proof contains an error, in the opinion of this author. In a subsequent paper, Dharmadhikari and Nadkarni [18] quietly and without comment simplified the example of Fox and Rubin and also gave a correct proof (without explicitly pointing out that the Fox-Rubin proof is erroneous). In this section, we review the example of [18] and slightly simplify their proof. It is worth noting that the example crucially depends on rotating a vector by an angle α that is not commensurate with π, that is, α/π is not a rational number. A similar approach is used by Benvenuti and Farina [6], Example 4 to construct a nonnegative impulse response with finite Hankel rank which does not have a finite rank *nonnegative* realization.

Let us begin by choosing numbers $\lambda \in (0, 0.5], \alpha \in (0, 2\pi)$ such that α and π are non-commensurate. In particular, this rules out the possibility that $\alpha = \pi$. Now define

$$h_l := \lambda^l \sin^2(l\alpha/2), \ \forall l \geq 1.$$

Note that we can also write

$$h_l = \lambda^l \frac{(e^{\mathbf{i}l\alpha/2} - e^{-\mathbf{i}l\alpha/2})^2}{4},$$

where (just in this equation) $\mathbf{i}$ denotes $\sqrt{-1}$. Simplifying the expression for h_l shows that

$$h_l = \frac{\lambda^l}{4}(\zeta^l + \zeta^{-l} - 2), \tag{5.5}$$

where $\zeta := e^{\mathbf{i}\alpha}$. Because h_l decays at a geometric rate with respect to l, the following properties are self-evident.

1. $h_i > 0 \; \forall i$. Note that $l\alpha$ can never equal a multiple of π because α and π are noncommensurate.

2. We have that
$$\sum_{i=1}^{\infty} h_i =: \delta < 1. \tag{5.6}$$

3. We have that
$$\sum_{i=1}^{\infty} i h_i < \infty.$$

4. The infinite Hankel matrix
$$\bar{H} := \begin{bmatrix} h_1 & h_2 & h_3 & \dots \\ h_2 & h_3 & h_4 & \dots \\ h_3 & h_4 & h_5 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$
has finite rank of 3.

Given a sequence $\{h_i\}_{i \geq 1}$, let us define its z-transform $\tilde{h}(\cdot)$ by[6]
$$\tilde{h}(z) := \sum_{i=1}^{\infty} h_i z^{i-1}.$$

Thanks to an old theorem of Kronecker [31], it is known that the Hankel matrix $\bar{H}$ has finite rank if and only if $\tilde{h}$ is a *rational* function of z, in which case the rank of the Hankel matrix is the same as the degree of the rational function $\tilde{h}(z)$. Now it is a ready consequence of (5.5) that
$$\tilde{h}(z) = \frac{1}{4}\left[\frac{\lambda\zeta}{1-\lambda\zeta z} + \frac{\lambda\zeta^{-1}}{1-\lambda\zeta^{-1}z} - 2\frac{\lambda}{1-\lambda z}\right].$$
Hence the infinite matrix $\bar{H}$ has rank 3.

The counterexample is constructed by defining a Markov process $\{\mathcal{X}_t\}$ with a countable state space and another process $\{\mathcal{Y}_t\}$ with just two output values such that $\mathcal{Y}_t$ is a function of $\mathcal{X}_t$. The process $\{\mathcal{Y}_t\}$ satisfies the finite Hankel rank condition; in fact $\text{Rank}(H) \leq 5$.

[6]Normally in z-transformation theory, the sequence $\{h_i\}$ is indexed starting from $i = 0$, whereas here we have chosen to begin with $i = 1$. This causes the somewhat unconventional-looking definition.

And yet no Markov process with a finite state space can be found such that $\mathcal{Y}_t$ is a function of that Markov process. Since we already know from Section 3 that the existence of all the three kinds of HMMs is equivalent, this is enough to show that the process $\{\mathcal{Y}_t\}$ does not have a joint Markov process type of HMM.

The process $\{\mathcal{X}_t\}$ is Markovian with a countable state space $\{0, 1, 2, \ldots\}$. The transition probabilities of the Markov chain are defined as follows:

$$\Pr\{\mathcal{X}_{t+1} = 0 | \mathcal{X}_t = 0\} = 1 - \delta = 1 - \sum_{i=1}^{\infty} h_i,$$

$$\Pr\{\mathcal{X}_{t+1} = i | \mathcal{X}_t = 0\} = h_i \text{ for } i = 1, 2, \ldots,$$

$$\Pr\{\mathcal{X}_{t+1} = i | \mathcal{X}_t = i + 1\} = 1 \text{ for } i = 1, 2, \ldots,$$

and all other probabilities are zero. Thus the dynamics of the Markov chain are as follows: If the chain starts in the initial state 0, then it makes a transition to state i with probability h_i, or remains in 0 with the probability $1 - \sum_i h_i = 1 - \delta$. Once the chain moves to the state i, it then successively goes through the states $i - 1, i - 2, \ldots, 1, 0$. Then the process begins again. Thus the dynamics of the Markov chain consist of a series of cycles beginning and ending at state 0, but where the lengths of the cycles are random, depending on the transition out of the state 0.

Clearly $\{\mathcal{X}_t\}$ is a Markov process. Now we define $\{\mathcal{Y}_t\}$ to be a function of this Markov process. Let $\mathcal{Y}_t = a$ if $\mathcal{X}_t = 0$, and let $\mathcal{Y}_t = b$ otherwise, i.e., if $\mathcal{X}_t = i$ for some $i \geq 1$. Thus the output process $\{\mathcal{Y}_t\}$ assumes just two values a and b. Note that in the interests of clarity we have chosen to denote the two output states as a and b instead of 1 and 2. For this process $\{\mathcal{Y}_t\}$ we shall show that (i)

$$\text{Rank}(H) \leq \text{Rank}(\bar{H}) + 2 = 5,$$

where H is the Hankel matrix associated with the process $\{\mathcal{Y}_t\}$, and (ii) there is no Markov process $\{\mathcal{Z}_t\}$ with a finite state space such that $\mathcal{Y}_t$ is a (deterministic) function of $\mathcal{Z}_t$.

The stationary distribution of the Markov chain is as follows:

$$\pi_0 = g := \left[1 + \sum_{i=1}^{\infty} i h_i\right]^{-1},$$

$$\pi_i = g \sum_{j=i}^{\infty} h_j, i \geq 1.$$

To verify this, note the structure of the state transition matrix A of the Markov chain: State 0 can be reached only from states 0 and 1. Thus column 0 of A has $1-\delta$ in row 0, 1 in row 1, and zeros in all other rows. For $i \geq 1$, state i can be reached only from states 0 and $i+1$. Hence column i has h_i in row 0, 1 in row $i+1$, and zeros elsewhere. As a result

$$(\pi A)_0 = g\left(1 - \delta + \sum_{j=1}^{\infty} h_j\right) = g = \pi_0,$$

while for $i \geq 1$,

$$(\pi A)_i = h_i \pi_0 + \pi_{i+1} = g\left[h_i + \sum_{j=i+1}^{\infty} h_j\right] = g \sum_{j=i}^{\infty} h_j = \pi_i.$$

To verify that this is indeed a probability vector, note that

$$\begin{aligned}
\sum_{i=0}^{\infty} \pi_i &= g\left[1 + \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} h_j\right] \\
&= g\left[1 + \sum_{j=1}^{\infty} \sum_{i=1}^{j} h_j\right] \\
&= g\left[1 + \sum_{j=1}^{\infty} j h_j\right] = 1
\end{aligned}$$

in view of the definition of g.

Next, let us compute the frequencies of various output strings. Note that if $\mathcal{Y}_t = a$, then certainly $\mathcal{X}_t = 0$. Hence, if $\mathcal{Y}_t = a$, then the conditional probability of $\mathcal{Y}_{t+1}$ does not depend on the values of $\mathcal{Y}_i, i < t$. Therefore, for arbitrary strings $\mathbf{u}, \mathbf{v} \in \{a, b\}^*$, we have

$$f_{\mathbf{u}a\mathbf{v}} = f_{\mathbf{u}a} \cdot f_{\mathbf{v}|\mathbf{u}a} = f_{\mathbf{u}a} \cdot f_{\mathbf{v}|a}.$$

Hence the infinite matrix $H^{(a)}$ defined by

$$H^{(a)} := [f_{\mathbf{u}a\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*]$$

has rank one. In such a case, it is customary to refer to a as a 'Markovian state.'

Next, let us compute the frequencies of strings of the form $ab^l a, ab^l, b^l a$, and b^l. A string of the form $ab^l a$ can occur only of $\mathcal{X}_t = 0, \mathcal{X}_{t+1} = l, \ldots, \mathcal{X}_{t+l} = 1, \mathcal{X}_{t+l+1} = 0$. All transitions except the first one have probability one, while the first transition has probability h_l. Finally, the probability that $\mathcal{X}_t = 0$ is π_0. Hence

$$f_{ab^l a} = \pi_0 h_l, \ \forall l.$$

Next, note that

$$f_{ab^l} = f_{ab^{l+1}} + f_{ab^l a}.$$

Hence, if we define

$$\pi_0 \gamma_l := f_{ab^l},$$

then γ_l satisfies the recursion

$$\pi_0 \gamma_l = \pi_0 \gamma_{l+1} + \pi_0 h_l.$$

To start the recursion, note that

$$\begin{aligned} \pi_0 \gamma_1 &= f_{ab} = f_a - f_{aa} = \pi_0 - \pi_0(1-\delta) \\ &= \pi_0 \delta = \pi_0 \sum_{i=1}^{\infty} h_i. \end{aligned}$$

Therefore

$$\pi_0 \gamma_l = \pi_0 \sum_{i=l}^{\infty} h_i, \text{ or } \gamma_l = \sum_{i=l}^{\infty} h_i.$$

Now we compute the frequencies f_{b^l} for all l. Note that

$$f_{b^l} = f_{b^{l+1}} + f_{ab^l} = f_{b^{l+1}} + \pi_0 \gamma_l.$$

Hence if we define $\pi_0 \eta_l := f_{b^l}$, then η_l satisfies the recursion

$$\eta_l = \eta_{l+1} + \gamma_l.$$

To start the recursion, note that

$$f_b = 1 - f_a = 1 - \pi_0.$$

Now observe that

$$\pi_0 = \left[1 + \sum_{i=1}^{\infty} i h_i \right]^{-1}$$

and as a result

$$\begin{aligned} 1 - \pi_0 &= \pi_0 \sum_{i=1}^{\infty} i h_i = \pi_0 \sum_{i=1}^{\infty} \sum_{j=1}^{i} h_i \\ &= \pi_0 \sum_{j=1}^{\infty} \sum_{i=j}^{\infty} h_i = \pi_0 \sum_{j=1}^{\infty} \gamma_j. \end{aligned}$$

Hence

$$f_{b^l} = \pi_0 \eta_l, \text{ where } \eta_l = \sum_{i=l}^{\infty} \gamma_i.$$

Finally, to compute $f_{b^l a}$, note that

$$f_{b^l a} + f_{b^{l+1}} = f_{b^l}.$$

Hence

$$f_{b^l a} = f_{b^l} - f_{b^{l+1}} = \pi_0(\eta_l - \eta_{l+1}) = \pi_0 \gamma_l.$$

Now let us look at the Hankel matrix H corresponding to the process $\{\mathcal{Y}_t\}$. We can think of H as the interleaving of two infinite matrices $H^{(a)}$ and $H^{(b)}$, where

$$H^{(a)} = [f_{\mathbf{u}a\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*],$$

$$H^{(b)} = [f_{\mathbf{u}b\mathbf{v}}, \mathbf{u}, \mathbf{v} \in \{a, b\}^*].$$

We have already seen that $H^{(a)}$ has rank one, since a is a Markovian state. Hence it follows that

$$\text{Rank}(H) \leq \text{Rank}(H^{(a)}) + \text{Rank}(H^{(b)}) = \text{Rank}(H^{(b)}) + 1.$$

To bound $\text{Rank}(H^{(b)})$, fix integers l, n, and define

$$H^{(b)}_{l,n} := [f_{\mathbf{u}b\mathbf{v}}, \mathbf{u} \in \{a, b\}^l, \mathbf{v} \in \{a, b\}^n].$$

Note that $H^{(b)}_{l,n} \in [0, 1]^{2^l \times 2^n}$. It is now shown that

$$\text{Rank}(H^{(b)}_{l,n}) \leq \text{Rank}(\bar{H}) + 1 = 4. \tag{5.7}$$

Since the right side is independent of l, n, it follows that

$$\text{Rank}(H^{(b)}) \leq 4,$$

whence

$$\text{Rank}(H) \leq 5.$$

To prove (5.7), suppose $\mathbf{u} \in \{a, b\}^{l-1}$ is arbitrary. Then

$$f_{\mathbf{u}ab\mathbf{v}} = f_{\mathbf{u}a} \cdot f_{b\mathbf{v}|\mathbf{u}a} = f_{\mathbf{u}a} \cdot f_{b\mathbf{v}|a},$$

because a is a Markovian state. Hence each of the 2^{l-1} rows $[f_{\mathbf{u}ab\mathbf{v}}, \mathbf{u} \in \{a, b\}^{l-1}]$ is a multiple of the row $[f_{b\mathbf{v}|a}]$, or equivalently, of the row $[f_{a^l b\mathbf{v}}]$. Hence $\text{Rank}(H^{(b)})$ is unaffected if we

keep only this one row and jettison the remaining $2^{l-1}-1$ rows. Similarly, as $\mathbf{u}$ varies over $\{a,b\}^{l-2}$, each of the rows $[f_{\mathbf{u}abb\mathbf{v}}]$ is proportional to $[f_{a^{l-2}abb\mathbf{v}}] = [f_{a^{l-1}b^2\mathbf{v}}]$. So we can again retain just the row $[f_{a^{l-1}b^2\mathbf{v}}]$ and discard the rest. Repeating this argument l times shows that $H^{(b)}$ has the same rank as the $(l+1)\times 2^n$ matrix

$$\begin{bmatrix} f_{a^l b\mathbf{v}} \\ f_{a^{l-1}b^2\mathbf{v}} \\ \vdots \\ f_{ab^l\mathbf{v}} \\ f_{b^{l+1}\mathbf{v}} \end{bmatrix}, \mathbf{v}\in\{a,b\}^n.$$

A similar exercise can now be repeated with $\mathbf{v}$. If $\mathbf{v}$ has the form $\mathbf{v}=a\mathbf{w}$, then

$$f_{a^i b^{l+1-i}a\mathbf{w}} = f_{a^i b^{l+1-i}a}\cdot f_{\mathbf{w}|a}.$$

So all 2^{n-1} columns $[f_{a^i b^{l+1-i}a\mathbf{w}}, \mathbf{w}\{a,b\}^{n-1}]$ are proportional to the single column $[f_{a^i b^{l+1-i}a^n}]$. So we can keep just this one column and throw away the rest. Repeating this argument shows that $H^{(b)}$ has the same rank as the $(l+1)\times(n+1)$ matrix

$$\begin{array}{c} \\ a^l \\ a^{l-1}b \\ \vdots \\ ab^{l-1} \\ b^l \end{array}\!\!\begin{array}{c} \begin{array}{ccccc} ba^n & b^2a^{n-1} & \dots & b^n a & b^{n+1} \end{array} \\ \begin{bmatrix} f_{a^l ba^n} & f_{a^l b^2 a^{n-1}} & \dots & f_{a^l b^n a} & f_{a^l b^{n+1}} \\ f_{a^{l-1}b^2a^n} & f_{a^{l-1}b^3a^{n-1}} & \dots & f_{a^{l-1}b^{n+1}a} & f_{a^{l-1}b^{n+2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{ab^l a^n} & f_{ab^{l+1}a^{n-1}} & \dots & f_{ab^{l+n-1}a} & f_{ab^{l+n}} \\ f_{b^{l+1}a^n} & f_{b^{l+2}a^{n-1}} & \dots & f_{b^{l+n}a} & f_{b^{l+n+1}} \end{bmatrix} \end{array}.$$

The structure of this matrix becomes clear if we note that

$$\begin{aligned} f_{a^i b^j a^t} &= f_{a^i}\cdot f_{b^j a^t|a} \\ &= f_{a^i}\cdot f_{b^j a|a}\cdot f_{a^{t-1}|a} \\ &= \pi_0(1-\delta)^{i-1}\cdot h_j\cdot(1-\delta)^{t-1}. \end{aligned} \tag{5.8}$$

The strings in the last row and column either do not begin with a, or end with a, or both. So let us divide the first row by $\pi_0(1-\delta)^{l-1}$, the second row by $\pi_0(1-\delta)^{l-2}$, etc., the l-th row by π_0, and do nothing to the last row. Similarly, let us divide the first column by $(1-\delta)^{n-1}$, the second column by $(1-\delta)^{n-2}$, etc., the n-th column by $(1-\delta)^0=1$, and leave the last column as is. The resulting matrix has the same rank as $H^{(b)}_{l,n}$, and the matrix is

$$\begin{bmatrix} h_1 & h_2 & \dots & h_n & \times \\ h_2 & h_3 & \dots & h_{n+1} & \times \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_l & h_{l+1} & \dots & h_{l+n} & \times \\ \times & \times & \dots & \times & \times \end{bmatrix},$$

where $\times$ denotes a number whose value does not matter. Now the upper left $l \times n$ submatrix is a submatrix of $\bar{H}$; as a result its rank is bounded by 3. This proves(5.7).[7]

To carry on our analysis of this example, we make use of z-transforms. This is not done in [18], but it simplifies the arguments to follow. As shown earlier, the z-transform of the sequence $\{h_i\}$ is given by

$$\tilde{h}(z) = \frac{1}{4}\left[\frac{\lambda\zeta}{1-\lambda\zeta z} + \frac{\lambda\zeta^{-1}}{1-\lambda\zeta^{-1}z} - 2\frac{\lambda}{1-\lambda z}\right] = \frac{\psi_h(z)}{\phi(z)},$$

where

$$\phi(z) := (1-\lambda\zeta)(1-\lambda\zeta^{-1})(1-\lambda), \tag{5.9}$$

and $\psi_h(z)$ is some polynomial of degree no larger than two; its exact form does not matter. Next, recall that

$$\gamma_i = \sum_{j=i}^{\infty} h_j.$$

Now it is an easy exercise to show that

$$\tilde{\gamma}(z) = \frac{\delta - \tilde{h}(z)}{1-z},$$

where, as defined earlier, $\delta = \sum_{i=1}^{\infty} h_i$. Even though we are dividing by $1-z$ in the above expression, in reality $\tilde{\gamma}$ does not have a pole at $z=1$, because $\tilde{h}(1) = \delta$. Hence we can write

$$\tilde{\gamma}(z) = \frac{\psi_\gamma(z)}{\phi(z)},$$

where again ψ_γ is some polynomial of degree no larger than two, and $\phi(z)$ is defined in (5.9). By entirely similar reasoning, it follows from the expression

$$\eta_i = \sum_{j=i}^{\infty} \gamma_j$$

that

$$\tilde{\eta}(z) = \frac{s - \tilde{\gamma}(z)}{1-z},$$

where

$$s := \sum_{i=1}^{\infty} \gamma_i = \sum_{i=1}^{\infty}\sum_{j=i}^{\infty} h_j = \sum_{j=1}^{\infty}\sum_{i=1}^{j} h_j = \sum_{j=1}^{\infty} j h_j.$$

[7]Through better book-keeping, Dharmadhikari and Nadkarni [18] show that the rank is bounded by 3, not 4. This slight improvement is not worthwhile since all that matters is that the rank is finite.

Here again, $\tilde{\eta}(\cdot)$ does not have a pole at $z = 1$, and in fact

$$\tilde{\gamma}(z) = \frac{\psi_\eta(z)}{\phi(z)},$$

where ψ_η is also a polynomial of degree no larger than two. The point of all these calculations is to show that each of the quantities γ_l, η_l has the form

$$\gamma_l = c_{0,\gamma}\lambda^l + c_{1,\gamma}\lambda^l\zeta^l + c_{2,\gamma}\lambda^l\zeta^{-l}, \tag{5.10}$$

$$\eta_l = c_{0,\eta}\lambda^l + c_{1,\eta}\lambda^l\zeta^l + c_{2,\eta}\lambda^l\zeta^{-l}, \tag{5.11}$$

for appropriate constants. Note that, even though ζ is a complex number, the constants occur in conjugate pairs so that γ_l, η_l are always real. And as we have already seen from (5.5), we have

$$h_l = -\frac{1}{2}\lambda^l + \frac{1}{4}\lambda^l\zeta^l + \frac{1}{4}\lambda^l\zeta^{-l}.$$

Now the expression (5.11) leads at once to two very important observations.

Observation 1: Fix some positive number ρ, and compute the weighted average

$$\frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\eta_l =: \theta(\rho, T).$$

Then it follows that

1. If $\rho < \lambda$, then $\theta(\rho, T) \to \infty$ as $T \to \infty$.
2. If $\rho > \lambda$, then $\theta(\rho, T) \to 0$ as $T \to \infty$.
3. If $\rho = \lambda$, then $\theta(\rho, T) \to c_{0,\eta}$ as $T \to \infty$, where $c_{0,\eta}$ is the constant in (5.11).

If $\rho \neq \lambda$, then the behavior of $\theta(\rho, T)$ is determined by that of $(\lambda/\rho)^l$. If $\rho = \lambda$, then the averages of the oscillatory terms $(\lambda\zeta/\rho)^l$ and $(\lambda/\rho\zeta)^l$ will both approach zero, and only the first term in (5.11) contributes to a nonzero average.

Observation 2: Let T be any fixed integer, and consider the moving average

$$\frac{1}{T}\sum_{j=l+1}^{l+T}\lambda^{-j}\eta_j =: \theta_l^T.$$

This quantity does not have a limit as $l \to \infty$ if α is not commensurate with π. To see this, take the z-transform of $\{\theta_l^T\}$. This leads to

$$\tilde{\theta}^T(z) = \frac{\beta^T(z)}{\phi(z)},$$

where $\beta^T(z)$ is some high degree polynomial. After dividing through by $\phi(z)$, we get

$$\tilde{\theta}^T(z) = \beta_q^T(z) + \frac{\beta_r^T(z)}{\phi(z)},$$

where β_q^T is the quotient and β_r^T is the remainder (and thus has degree no more than two). By taking the inverse z-transform, we see that the sequence $\{\theta_l^T\}$ is the sum of two parts: The first part is a sequence having finite support (which we can think of as the 'transient'), and the second is a sequence of the form

$$c_{0,\theta}\lambda^l + c_{1,\theta}\lambda^l\zeta^l + c_{2,\theta}\lambda^l\zeta^{-l}.$$

From this expression it is clear that if α is noncommensurate with π, then θ_l^T does not have a limit as $l \to \infty$.

These two observations are the key to the concluding part of this very long line of reasoning. Suppose by way of contradiction that the output process $\{\mathcal{Y}_t\}$ can be expressed as a function of a Markov process $\{\mathcal{Z}_t\}$ with a finite state space. Let $\mathcal{N} = \{1, \ldots, n\}$ denote the state space, and let π, A denote the stationary distribution and state transition matrix of the Markov chain $\{\mathcal{Z}_t\}$. Earlier we had used these symbols for the Markov chain $\{\mathcal{X}_t\}$, but no confusion should result from this recycling of notation. From the discussion in Chapter 1 of [38], it follows that by a symmetric permutation of rows and columns (which corresponds to permuting the labels of the states), A can be arranged in the form

$$A = \begin{bmatrix} P & 0 \\ R & Q \end{bmatrix},$$

where the rows of P correspond to the recurring states and those of R to transient states. Similarly, it follows from the discussion in Chapter 4 of [38] that the components of π corresponding to transient states are all zero. Hence the corresponding states can be dropped from the set $\mathcal{N}$ without affecting anything. So let us assume that all states are recurrent.

Next, we can partition the state space $\mathcal{N}$ into those states that map into a, and those states that map into b. With the obvious notation, we can partition π as $[\pi_a \;\; \pi_b]$ and the state transition matrix as

$$A = \begin{bmatrix} A_{aa} & A_{ab} \\ A_{ba} & A_{bb} \end{bmatrix}.$$

Moreover, again following the discussion in Chapter 1 of [38], we can arrange A_{bb} in the form

$$A_{bb} = \begin{bmatrix} A_{11} & 0 & \ldots & 0 \\ A_{21} & A_{22} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ A_{s1} & A_{s1} & \ldots & A_{ss} \end{bmatrix},$$

where s is the number of communicating classes within those states that map into the output b, and each of the diagonal matrices A_{ii} is irreducible. Of course, the fact that each of the diagonal blocks is irreducible does still not suffice to determine π uniquely, but as before we can assume that no component of π is zero, because if some component of π is zero, then we can simply drop that component from the state space.

Now it is claimed that $\rho(A_{bb}) = \lambda$, where $\rho(\cdot)$ denotes the spectral radius. To show this, recall that if B is an irreducible matrix with spectral radius $\rho(B)$, the (unique strictly positive) row eigenvector θ and the column eigenvector ϕ corresponding to the eigenvalue $\rho(B)$, then the 'ergodic average'

$$\frac{1}{T}\sum_{l=1}^{T}[\rho(B)]^{-l}B^l$$

converges to the rank one matrix $\phi\theta$ as $T \to \infty$. Now from the triangular structure of A_{bb}, it is easy to see that $\rho(A_{bb})$ is the maximum amongst the numbers $\rho(A_{ii}), i = 1, \ldots, s$. If we let θ_i, ϕ_i denote the unique row and column eigenvectors of A_{ii} corresponding to $\rho(A_{ii})$, it is obvious that

$$\frac{1}{T}\sum_{l=1}^{T}[\rho(A_{bb}]^{-l}A_{bb}^l \to \text{ Block Diag } \{\phi_i\theta_i I_{\{\rho(A_{ii})=\rho(A_{bb})\}}\}. \tag{5.12}$$

In other words, if $\rho(A_{ii}) = \rho(A_{bb})$, then the corresponding term $\phi_i\theta_i$ is present in the block diagonal matrix; if $\rho(A_{ii}) < \rho(A_{bb})$, then the corresponding entry in the block diagonal matrix is the zero matrix. Let D denote the block diagonal in (5.12), and note that at least one of the $\rho(A_{ii})$ equals $\rho(A_{bb})$. Hence at least one of the products $\phi_i\theta_i$ is present in the block diagonal matrix D.

From the manner in which the HMM has been set up, it follows that

$$\eta_l = f_{b^l} = \pi_b A_{bb}^l \mathbf{e}.$$

In other words, the only way in which we can observe a sequence of l symbols b in succession is for all states to belong to the subset of $\mathcal{N}$ that map into the output b. Next, let us examine the behavior of the quantity

$$\frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\eta_l = \frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\pi_b A_{bb}^l \mathbf{e},$$

where $\rho = \rho(A_{bb})$. Now appealing to (5.12) shows that the above quantity has a definite limit as $T \to \infty$. Moreover, since π_b and $\mathbf{e}$ are strictly positive, and the block diagonal matrix D

has at least one positive block $\phi_i\theta_i$, it follows that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{l=1}^{T}\rho^{-l}\eta_l = \pi D\mathbf{e} \in (0,\infty).$$

By Observation 1, this implies that $\rho(A_{bb}) = \lambda$.

Finally (and at long last), let us examine those blocks A_{ii} which have the property that $\rho(A_{ii}) = \rho(A_{bb}) = \rho$. Since each of these is an irreducible matrix, it follows from the discussion in Chapter 4 of [38] that each such matrix has a unique 'period' n_i, which is an integer. Moreover, A_{ii} has eigenvalues at $\rho\exp(\mathbf{i}2\pi j/n_i)$, $j = 1,\ldots,n_i-1$, and all other eigenvalues of A_{ii} have magnitude strictly less than ρ. This statement applies *only* to those indices i such that $\rho(A_{ii}) = \rho(A_{bb}) = \rho$. Now let N denote the least common multiple of all these integers n_i. Then it is clear that the matrix A_{bb} has a whole lot of eigenvalues of the form $\rho\exp(\mathbf{i}2\pi j/N)$ for some (though not necessarily all) values of j ranging from 0 to $N-1$; all other eigenvalues of A have magnitude strictly less than ρ. As a result, the quantity

$$\frac{1}{N}\sum_{l=t+1}^{t+N}A_{bb}^{l}$$

has a definite limit at $t\to\infty$. In turn this implies that the quantity

$$\frac{1}{N}\sum_{l=t+1}^{t+N}\pi_b A_{bb}^{l}\mathbf{e} = \frac{1}{N}\sum_{l=t+1}^{t+N}\eta_l$$

has a definite limit at $t\to\infty$. However, this contradicts Observation 2, since α is non-commensurate with π. This contradiction shows that the stochastic process $\{\mathcal{Y}_t\}$ cannot be realized as a function of a finite state Markov chain.

6 An Abstract Necessary and Sufficient Condition

In [23], Heller stated and proved an abstract necessary and sufficient condition for a given probability law to have an HMM realization. Heller's paper is very difficult to follow since it adopts a 'coordinate-free' approach. Picci [34] gave a very readable proof of Heller's theorem, which is reproduced here with minor variations.

Recall that $\mathcal{M}^*$, the set of all finite strings over $\mathcal{M} = \{1,\ldots,m\}$, is a countable set. We let $\mu(\mathcal{M}^*)$ denote the set of all maps $p : \mathcal{M}^* \to [0,1]$ satisfying the following two conditions:

$$\sum_{u\in\mathcal{M}} p_u = 1, \tag{6.1}$$

$$\sum_{v \in \mathcal{M}} p_{\mathbf{u}v} = p_{\mathbf{u}}, \ \forall \mathbf{u} \in \mathcal{M}^*. \tag{6.2}$$

Note that by repeated application of (6.2), we can show that

$$\sum_{\mathbf{v} \in \mathcal{M}^l} p_{\mathbf{u}\mathbf{v}} = p_{\mathbf{u}}, \ \forall \mathbf{u} \in \mathcal{M}^*. \tag{6.3}$$

By taking $\mathbf{u}$ to be the empty string, so that $p_{\mathbf{u}} = 1$, we get from the above that

$$\sum_{\mathbf{v} \in \mathcal{M}^l} p_{\mathbf{v}} = 1, \ \forall l. \tag{6.4}$$

We can think of $\mu(\mathcal{M}^*)$ as the set of all frequency assignments to strings in $\mathcal{M}^*$ that are **right-consistent** by virtue of satisfying (6.2).

Definition 4 *Given a frequency assignment $p \in \mu(\mathcal{M}^*)$, we say that $\{\pi, M^{(1)}, \ldots, M^{(m)}\}$ is a* **HMM realization** *of p if*

$$\pi \in \mathbb{R}^n_+, \sum_{i=1}^{n} \pi_i = 1, \tag{6.5}$$

$$M^{(u)} \in [0,1]^{n \times n} \ \forall u \in \mathcal{M}, \tag{6.6}$$

$$\left[\sum_{u \in \mathcal{M}} M^{(u)} \right] \mathbf{e}_n = \mathbf{e}_n, \tag{6.7}$$

and finally

$$p_{\mathbf{u}} = \pi M^{(u_1)} \ldots M^{(u_l)} \mathbf{e}_n \ \forall \mathbf{u} \in \mathcal{M}^l. \tag{6.8}$$

Given a frequency distribution $p \in \mu(\mathcal{M}^*)$, for each $u \in \mathcal{M}$ we define the conditional distribution

$$p(\cdot|u) := \mathbf{v} \in \mathcal{M}^* \mapsto \frac{p_{u\mathbf{v}}}{p_u}. \tag{6.9}$$

If by chance $p_u = 0$, we define $p(\cdot|u)$ to equal p. Note that $p(\cdot|u) \in \mu(\mathcal{M}^*)$; that is, $p(\cdot|u)$ is also a frequency assignment map. By applying (6.9) repeatedly, for each $\mathbf{u} \in \mathcal{M}^*$ we can define the conditional distribution

$$p(\cdot|\mathbf{u}) := \mathbf{v} \in \mathcal{M}^* \mapsto \frac{p_{\mathbf{u}\mathbf{v}}}{p_{\mathbf{u}}}. \tag{6.10}$$

Again, for each $\mathbf{u} \in \mathcal{M}^*$, the conditional distribution $p(\cdot|\mathbf{u})$ is also a frequency assignment. Clearly conditioning can be applied recursively and the results are consistent. Thus

$$p((\cdot|\mathbf{u})|\mathbf{v}) = p(\cdot|\mathbf{u}\mathbf{v}), \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^*. \tag{6.11}$$

It is easy to verify that if p satisfies the right-consistency condition (6.2), then so do all the conditional distributions $p(\cdot|\mathbf{u})$ for all $\mathbf{u} \in \mathcal{M}^*$. Thus, if $p \in \mu(\mathcal{M}^*)$, then $p(\cdot|\mathbf{u}) \in \mu(\mathcal{M}^*)$ for all $\mathbf{u} \in \mathcal{M}^*$.

A set $\mathcal{C} \subseteq \mu(\mathcal{M}^*)$ is said to be **polyhedral** if there exist an integer n and distributions $q^{(1)}, \ldots, q^{(n)} \in \mu(\mathcal{M}^*)$ such that $\mathcal{C}$ is the convex hull of these $q^{(i)}$, that is, every $q \in \mathcal{C}$ is a convex combination of these $q^{(i)}$. A set $\mathcal{C} \subseteq \mu(\mathcal{M}^*)$ is said to be **stable** if

$$q \in \mathcal{C} \;\Rightarrow\; q(\cdot|\mathbf{u}) \in \mathcal{C}\ \forall \mathbf{u} \in \mathcal{M}^*. \tag{6.12}$$

In view of (6.11), (6.12) can be replaced by weaker-looking condition

$$q \in \mathcal{C} \;\Rightarrow\; q(\cdot|u) \in \mathcal{C}\ \forall u \in \mathcal{M}. \tag{6.13}$$

Now we are ready to state the main result of this section, first proved in [23]. However, the proof below follows [34] with some slight changes in notation.

Theorem 6.1 *A frequency distribution $p \in \mu(\mathcal{M}^*)$ has a HMM realization if and only if there exists a stable polyhedral set $\mathcal{C} \subseteq \mu(\mathcal{M}^*)$ containing p.*

Proof: "If" Suppose $q^{(1)}, \ldots, q^{(n)} \in \mu(\mathcal{M}^*)$ are the generators of the polyhedral set $\mathcal{C}$. Thus every $q \in \mathcal{C}$ is of the form

$$q = \sum_{i=1}^{n} a_i q^{(i)}, a_i \geq 0, \sum_{i=1}^{n} a_i = 1.$$

In general neither the integer n nor the individual distributions $q^{(i)}$ are unique, but this does not matter. Now, since $\mathcal{C}$ is stable, $q(\cdot|u) \in \mathcal{C}$ for all $a \in \mathcal{C}, u \in \mathcal{M}$. In particular, for each i, u, there exist constants $\alpha_{ij}^{(u)}$ such that

$$q^{(i)}(\cdot|u) = \sum_{j=1}^{n} \alpha_{ij}^{(u)} q^{(j)}(\cdot), \alpha_{ij}^{(u)} \geq 0, \sum_{j=1}^{n} \alpha_{ij}^{(u)} = 1.$$

Thus from (6.9) it follows that

$$\begin{aligned} q_{u\mathbf{v}}^{(i)} &= \sum_{j=1}^{n} q_u^{(i)} \alpha_{ij}^{(u)} q_{\mathbf{v}}^{(j)} \\ &= \sum_{j=1}^{n} m_{ij}^{(u)} q_{\mathbf{v}}^{(j)}, \end{aligned} \tag{6.14}$$

where

$$m_{ij}^{(u)} := q_u^{(i)} \alpha_{ij}^{(u)}, \ \forall i, j, u. \tag{6.15}$$

We can express (6.14) more compactly by using matrix notation. For $\mathbf{u} \in \mathcal{M}^*$, define

$$\mathbf{q_u} := [q_{\mathbf{u}}^{(1)} \dots q_{\mathbf{u}}^{(n)}]^t \in [0,1]^{n \times 1}.$$

Then (6.14) states that

$$\mathbf{q}_{u\mathbf{v}} = M^{(u)} \mathbf{q_v} \ \forall u \in M, \mathbf{v} \in \mathcal{M}^*,$$

where $M^{(u)} = [m_{ij}^{(u)}] \in [0,1]^{n \times n}$. Moreover, it follows from (6.11) that

$$\mathbf{q_{uv}} = M^{(u_1)} \dots M^{(u_l)} \mathbf{q_v} \ \forall \mathbf{u} \mathcal{M}^l, \mathbf{v} \in \mathcal{M}^*.$$

If we define

$$M^{(\mathbf{u})} := M^{(u_1)} \dots M^{(u_l)} \ \forall \mathbf{u} \in \mathcal{M}^l,$$

then the above equation can be written compactly as

$$\mathbf{q_{uv}} = M^{(\mathbf{u})} M^{(\mathbf{v})} \mathbf{q_v} \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^*. \tag{6.16}$$

By assumption, $p \in \mathcal{C}$. Hence there exist numbers $\pi_1, \dots, \pi_n$, not necessarily unique, such that

$$p(\cdot) = \sum_{i=1}^{n} \pi_i q^{(i)}(\cdot), \pi_i \geq 0 \ \forall i, \sum_{i=1}^{n} \pi_i = 1. \tag{6.17}$$

We can express (6.17) as

$$p(\cdot) = \pi \mathbf{q}(\cdot).$$

Hence, for all $\mathbf{u}, \mathbf{v} \in \mathcal{M}^*$, it follows from (6.16) that

$$p_{\mathbf{uv}} = \pi \mathbf{q_{uv}} = \pi M^{(\mathbf{u})} \mathbf{q_v}, \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^*. \tag{6.18}$$

In particular, if we let $\mathbf{v}$ equal the empty string, then $\mathbf{q_v} = \mathbf{e}_n$, and $p_{\mathbf{uv}} = p_{\mathbf{u}}$. Thus (6.18) becomes

$$p_{\mathbf{u}} = \pi M^{(\mathbf{u})} \mathbf{e}_n,$$

which is the same as (6.8).

Next, we verify (6.7) by writing it out in component form. We have

$$\begin{aligned}\sum_{j=1}^{n}\sum_{u\in\mathcal{M}} m_{ij}^{(u)} &= \sum_{u\in\mathcal{M}}\sum_{j=1}^{n} m_{ij}^{(u)} \\ &= \sum_{u\in\mathcal{M}} q_u^{(i)}\left[\sum_{j=1}^{n}\alpha_{ij}^{(u)}\right] \\ &= \sum_{u\in\mathcal{M}} q_u^{(i)} \text{ because } \sum_{j=1}^{n}\alpha_{ij}^{(u)} = 1 \\ &= 1,\ \forall i \text{ because } q^{(i)} \in \mu(\mathcal{M}^*) \text{ and (6.1).}\end{aligned}$$

Before leaving the "If" part of the proof, we observe that if the probability distribution $p \in \mu(\mathcal{M}^*)$ is also *left-consistent* by satisfying

$$\sum_{u\in\mathcal{M}} p_{u\mathbf{v}} = p_{\mathbf{v}}\ \forall u \in \mathcal{M}, \mathbf{v} \in \mathcal{M}^*,$$

then *it is possible* to choose the vector π such that

$$\pi\left[\sum_{u\in\mathcal{M}} M^{(u)}\right] = \pi. \tag{6.19}$$

To see this, we substitute into (6.8) which has already been established. This gives

$$\pi M^{(\mathbf{v})}\mathbf{e}_n = p_{\mathbf{v}} = \sum_{u\in\mathcal{M}} p_{u\mathbf{v}} = \pi\left[\sum_{u\in\mathcal{M}} M^{(u)}\right] M^{(\mathbf{v})}\mathbf{e}_n,\ \forall \mathbf{v} \in \mathcal{M}^*.$$

Now it is not possible to "cancel" $M^{(\mathbf{v})}\mathbf{e}_n$ from both sides of the above equation. However, it is always possible to choose the coefficient vector π so as to satisfy (6.19).

"Only if" Suppose p has a HMM realization $\{\pi, M^{(1)}, \ldots, M^{(m)}\}$. Let n denote the dimension of the matrices $M(u)$ and the vector π. Define the distributions $q^{(1)}, \ldots, q^{(n)}$ by

$$\mathbf{q_u} = [q_{\mathbf{u}}^{(1)} \ldots q_{\mathbf{u}}^{(n)}]^t := M^{(\mathbf{u})}\mathbf{e}_n,\ \forall \mathbf{u} \in \mathcal{M}^*. \tag{6.20}$$

Thus $q_{\mathbf{u}}^{(i)}$ is the i-th component of the column vector $M^{(\mathbf{u})}\mathbf{e}_n$. First it is shown that each $q^{(i)}$ is indeed a frequency distribution. From (6.20), it follows that

$$\sum_{u\in\mathcal{M}} \mathbf{q}_u = \left[\sum_{u\in\mathcal{M}} M^{(u)}\right]\mathbf{e}_n = \mathbf{e}_n,$$

where we make use of (6.7). Thus each $q^{(i)}$ satisfies (6.1) (with p replaced by $q^{(i)}$). Next, to show that each $q^{(i)}$ is right-consistent, observe that for each $\mathbf{u} \in \mathcal{M}^*, v \in \mathcal{M}$ we have

$$\sum_{v \in \mathcal{M}} \mathbf{q}_{\mathbf{u}v} = M^{(\mathbf{u})} \left[\sum_{v \in \mathcal{M}} M^{(v)} \right] \mathbf{e}_n = M^{(\mathbf{u})} \mathbf{e}_n = \mathbf{q}_{\mathbf{u}}.$$

Thus each $q^{(i)}$ is right-consistent. Finally, to show that the polyhedral set consisting of all convex combinations of $q^{(1)}, \ldots, q^{(n)}$ is stable, observe that

$$q^{(i)}(\mathbf{v}|u) = \frac{q^{(i)}(u\mathbf{v})}{q^{(i)}(u)}.$$

Substituting from (6.19) gives

$$\begin{aligned} q^{(i)}(\mathbf{v}|u) &= \frac{1}{q^{(i)}(u)} M^{(u)} M^{(\mathbf{v})} \mathbf{e}_n \\ &= \mathbf{a}_u^{(i)} M^{(\mathbf{v})} \mathbf{e}_n, \\ &= \mathbf{a}_u^{(i)} \mathbf{q}_{\mathbf{v}}, \end{aligned} \tag{6.21}$$

where

$$\mathbf{a}_u^{(i)} := \left[\frac{m_{ij}^{(u)}}{q^{(i)}(u)}, j = 1, \ldots, n \right] \in [0,1]^{1 \times n}.$$

Thus each conditional distribution $q^{(i)}(\cdot|u)$ is a linear combination of $q^{(1)}(\cdot), \ldots, q^{(n)}(\cdot)$. It remains only to show that $q^{(i)}(\cdot|u)$ is a *convex* combination, that is, that each $\mathbf{a}_u^{(i)} \in \mathbb{R}_+^n$ and that $\mathbf{a}_u^{(i)} \mathbf{e}_n = 1$. The first is obvious from the definition of the vector $\mathbf{a}^{(i)}$. To establish the second, substitute $\mathbf{v}$ equal to the empty string in (6.21). Then $q^{(i)}(\mathbf{v}|u) = 1$ for all i, u, and $\mathbf{q}_{\mathbf{v}} = \mathbf{e}_n$. Substituting these into (6.20) shows that

$$1 = \mathbf{a}_u^{(i)} \mathbf{e}_n,$$

as desired. Thus the polyhedral set $\mathcal{C}$ consisting of all convex combinations of the $q^{(i)}$ is stable. Finally, it is obvious from (6.8) that p is a convex combination of the $q^{(i)}$ and thus belongs to $\mathcal{C}$. ■

7 Existence of Regular Quasi-Realizations for Finite Hankel Rank Processes

In this section, we study processes whose Hankel rank is finite, and show that it is *always* possible to construct a 'quasi-realization' of such a process. Moreover, any two *regular* quasi-realizations of a finite Hankel rank process are related through a similarity transformation.

Definition 5 *Suppose a process* $\{\mathcal{Y}_t\}$ *has finite Hankel rank* r*. Suppose* $n \geq r$*,* $\mathbf{x}$ *is a row vector in* $\mathbb{R}^n$*,* $\mathbf{y}$ *is a column vector in* $\mathbb{R}^n$*, and* $C^{(u)} \in \mathbb{R}^{n \times n}$ $\forall u \in \mathcal{M}$*. Then we say that* $\{n, \mathbf{x}, \mathbf{y}, C^{(u)}, \mathbf{u} \in \mathcal{M}\}$ *is a* **quasi-realization** *of the process if three conditions hold. First,*

$$f_{\mathbf{u}} = \mathbf{x} C^{(u_1)} \ldots C^{(u_l)} \mathbf{y} \ \forall \mathbf{u} \in \mathcal{M}^*, \tag{7.1}$$

where $l = |\mathbf{u}|$*. Second,*

$$\mathbf{x} \left[\sum_{u \in \mathcal{M}} C^{(u)} \right] = \mathbf{x}. \tag{7.2}$$

Third,

$$\left[\sum_{u \in \mathcal{M}} C^{(u)} \right] \mathbf{y} = \mathbf{y}. \tag{7.3}$$

We say that $\{n, \mathbf{x}, \mathbf{y}, C^{(u)}, \mathbf{u} \in \mathcal{M}\}$ *is a* **regular quasi-realization** *of the process if* $n = r$*, the rank of the Hankel matrix.*

The formula (7.1) is completely analogous to (4.3). Similarly, (7.2) and (7.3) are analogous to (4.4). The only difference is that the various quantities are not required to be nonnegative. This is why we speak of a 'quasi-realization' instead of a true realization. With this notion, it is possible to prove the following powerful statements:

1. Suppose the process $\{\mathcal{Y}_t\}$ has finite Hankel rank, say r. Then the process always has a regular quasi-realization.

2. Suppose a process $\{\mathcal{Y}_t\}$ has finite Hankel rank r, and suppose $\{\theta_1, \phi_1, D_1^{(u)}, u \in \mathcal{M}\}$ and $\{\theta_2, \phi_2, D_2^{(u)}, u \in \mathcal{M}\}$ are two regular quasi-realizations of this process. Then there exists a nonsingular matrix T such that

$$\theta_2 = \theta_1 T^{-1}, D_2^{(u)} = T D_1^{(u)} T^{-1} \ \forall u \in \mathcal{M}, \phi_2 = T \phi_1.$$

These two statements are formally stated and proven as Theorem 7.1 and Theorem 7.2 respectively.

The results of this section are not altogether surprising. Given that the infinite matrix H has finite rank, it is clear that there *must exist* recursive relationships between its various elements. Earlier work, most notably [14, 11], contains some such recursive relationships. However, the present formulae are the cleanest, and also the closest to the conventional formula (4.3). Note that Theorem 7.1 is more or less contained in the work of Erickson [19]. In

[26], the authors generalize the work of Erickson by studying the relationship between two quasi-realizations, *without* assuming that the underlying state spaces have the same dimension. In this case, in place of the similarity transformation above, they obtain 'intertwining' conditions of the form $D_2^{(u)}T = TD_1^{(u)}$, where the matrix T may now be rectangular. In the interests of simplicity, in the present case we do not study this more general case. Moreover, the above formulae are the basis for the construction of a 'true' (as opposed to quasi) HMM realization in subsequent sections.

Some notation is introduced to facilitate the subsequent proofs. Suppose k, l are integers, and $I \subseteq \mathcal{M}^k$, $J \subseteq \mathcal{M}^l$; thus every element of I is a string of length k, while every element of J is a string of length l. Specifically, suppose $I = \{\mathbf{i}_1, \ldots, \mathbf{i}_{|I|}\}$, and $J = \{\mathbf{j}_1, \ldots, \mathbf{j}_{|J|}\}$. Then we define

$$F_{I,J} := \begin{bmatrix} f_{\mathbf{i}_1\mathbf{j}_1} & f_{\mathbf{i}_1\mathbf{j}_2} & \cdots & f_{\mathbf{i}_1\mathbf{j}_{|J|}} \\ f_{\mathbf{i}_2\mathbf{j}_1} & f_{\mathbf{i}_2\mathbf{j}_2} & \cdots & f_{\mathbf{i}_2\mathbf{j}_{|J|}} \\ \vdots & \vdots & \vdots & \vdots \\ f_{\mathbf{i}_{|I|}\mathbf{j}_1} & f_{\mathbf{i}_{|I|}\mathbf{j}_2} & \cdots & f_{\mathbf{i}_{|I|}\mathbf{j}_{|J|}} \end{bmatrix}. \tag{7.4}$$

Thus $F_{I,J}$ is a submatrix of $F_{k,l}$ and has dimension $|I| \times |J|$. This notation is easily reconciled with the earlier notation. Suppose k, l are integers. Then we can think of $F_{k,l}$ as shorthand for $F_{\mathcal{M}^k,\mathcal{M}^l}$. In the same spirit, if I is a subset of $\mathcal{M}^k$ and l is an integer, we use the 'mixed' notation $F_{I,l}$ to denote $F_{I,\mathcal{M}^l}$. This notation can be extended in an obvious way to the case where either k or l equals zero. If $l = 0$, we have that $\mathcal{M}^0 := \{\emptyset\}$. In this case

$$F_{I,0} := [f_{\mathbf{i}} : \mathbf{i} \in I] \in \mathbb{R}^{|I| \times 1}.$$

Similarly if $J \subseteq \mathcal{M}^l$ for some integer l, then

$$F_{0,J} := [f_{\mathbf{j}} : \mathbf{j} \in J] \in \mathbb{R}^{1 \times |J|}.$$

Finally, given any string $\mathbf{u} \in \mathcal{M}^*$, we define

$$F_{k,l}^{(\mathbf{u})} := [f_{\mathbf{iuj}}, \mathbf{i} \in \mathcal{M}^k \text{ in flo}, \mathbf{j} \in \mathcal{M}^l \text{ in llo}], \tag{7.5}$$

$$F_{I,J}^{(\mathbf{u})} := [f_{\mathbf{iuj}}, \mathbf{i} \in I, \mathbf{j} \in J]. \tag{7.6}$$

Lemma 7.1 *Suppose H has finite rank. Then there exists a smallest integer k such that*

$$Rank(F_{k,k}) = Rank(H).$$

Moreover, for this k, we have

$$Rank(F_{k,k}) = Rank(H_{k+l,k+s}),\ \forall l, s \geq 0. \tag{7.7}$$

Proof: We begin by observing that, for every pair of integers k, l, we have

$$\text{Rank}(H_{k,l}) = \text{Rank}(F_{k,l}). \tag{7.8}$$

To see this, observe that the row indexed by $\mathbf{u} \in \mathcal{M}^{k-1}$ in $F_{k-1,s}$ is the sum of the rows indexed by $v\mathbf{u}$ in $F_{k,s}$, for each s. This follows from (4.1). Similarly each row in $F_{k-2,s}$ is the sum of m rows in $F_{k-1,s}$ and thus of m^2 rows of $F_{k,s}$, and so on. Thus it follows that every row of $F_{t,s}$ for $t < k$ is a sum of m^{k-t} rows of $F_{k,s}$. Therefore

$$\text{Rank}(H_{k,l}) = \text{Rank}([F_{k,0} \;\; F_{k,1} \;\; \ldots F_{k,l}]).$$

Now repeat the same argument for the columns of this matrix. Every column of $F_{k,t}$ is the sum of m^{k-t} columns of $F_{k,l}$. This leads to the desired conclusion (7.8).

To complete the proof, observe that, since $H_{l,l}$ is a submatrix of $H_{l+1,l+1}$, we have that

$$\text{Rank}(H_{1,1}) \leq \text{Rank}(H_{2,2}) \leq \ldots \leq \text{Rank}(H).$$

Now at each step, there are only two possibilities: Either $\text{Rank}(H_{l,l}) < \text{Rank}(H_{l+1,l+1})$, or else $\text{Rank}(H_{l,l}) = \text{Rank}(H_{l+1,l+1})$. Since $\text{Rank}(H)$ is finite, the first possibility can only occur finitely many times. Hence there exists a smallest integer k such that

$$\text{Rank}(H_{k,k}) = \text{Rank}(H).$$

We have already shown that $\text{Rank}(H_{k,k}) = \text{Rank}(F_{k,k})$. Finally, since $H_{k+l,k+s}$ is a submatrix of H and contains $H_{k.k}$ as a submatrix, the desired conclusion (7.7) follows. ■

Note: Hereafter, the symbol k is used *exclusively* for this integer and nothing else. Similarly, hereafter the symbol r is used exclusively for the (finite) rank of the Hankel matrix H and nothing else.

Now consider the matrix $F_{k,k}$, which is chosen so as to have rank r. Thus there exist sets $I, J \subseteq \mathcal{M}^k$, such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. (Recall the definition of the matrix $F_{I,J}$ from (7.4).) In other words, the index sets I, J are chosen such that $F_{I,J}$ is any full rank nonsingular submatrix of $F_{k,k}$. Of course the choice of I and J is not unique. However, once I, J are chosen, there exist *unique* matrices $U \in \mathbb{R}^{m^k \times r}, V \in \mathbb{R}^{r \times m^k}$ such that $F_{k,k} = UF_{I,J}V$. Hereafter, the symbols U, V are used only for these matrices and nothing else.

The next lemma shows that, once the index sets I, J are chosen (thus fixing the matrices U and V), the relationship $F_{k,k} = UF_{I,J}V$ can be extended to strings of *arbitrary* lengths.

Lemma 7.2 *With the various symbols defined as above, we have*

$$F_{k,k}^{(\mathbf{u})} = UF_{I,J}^{(\mathbf{u})}V, \ \forall \mathbf{u} \in \mathcal{M}^*. \tag{7.9}$$

This result can be compared to [1], Lemma 1, p. 99.

Proof: For notational convenience only, let us suppose I, J consist of the first r elements of $\mathcal{M}^r$. The more general case can be handled through more messy notation. The matrix U can be partitioned as follows:

$$U = \begin{bmatrix} I_r \\ \bar{U} \end{bmatrix}.$$

This is because $F_{I,k}$ is a submatrix of $F_{k,k}$. (In general we would have to permute the indices so as to bring the elements of I to the first r positions.) Now, by the rank condition and the assumption that $F_{k,k} = UF_{I,J}V(= UF_{I,k})$, it follows that

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k - r} \end{bmatrix} H_{k,\cdot} = \begin{bmatrix} F_{I,k} & F_{I,\cdot} \\ \mathbf{0} & F_{\mathcal{M}^k \setminus I,\cdot} - \bar{U}F_{I,\cdot} \end{bmatrix},$$

where

$$F_{I,\cdot} = [F_{I,k+1} \ \ F_{I,k+2} \ldots], \text{ and } F_{\mathcal{M}^k \setminus I,\cdot} = [F_{\mathcal{M}^k \setminus I,k+1} \ \ F_{\mathcal{M}^k \setminus I,k+2} \ldots].$$

This expression allows us to conclude that

$$F_{\mathcal{M}^k \setminus I,\cdot} = \bar{U}F_{I,\cdot}. \tag{7.10}$$

Otherwise the $(2,2)$-block of the above matrix would contain some nonzero element, which would in turn imply that $\text{Rank}(H_{k,\cdot}) > r$, a contradiction. Now the above relationship implies that

$$F_{k,k}^{(\mathbf{u})} = UF_{I,K}^{(\mathbf{u})}, \ \forall \mathbf{u} \in \mathcal{M}^*.$$

Next, as with U, partition V as $V = [I_r \ \ \bar{V}]$. (In general, we would have to permute the columns to bring the elements of J to the first positions.) Suppose $N > k$ is some integer. Observe that $F_{N,k}$ is just $[F_{k,k}^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^{N-k}$ in flo$]$. Hence

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k - r} \end{bmatrix} H_{\cdot,k} = \begin{bmatrix} F_{I,k}^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^* \text{ in flo} \\ \mathbf{0} \end{bmatrix} = \left[\begin{array}{c} F_{I,k} \\ \mathbf{0} \\ \hline F_{I,k}^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^* \setminus \emptyset \text{ in flo} \\ \mathbf{0} \end{array}\right].$$

Now post-multiply this matrix as shown below:

$$\begin{bmatrix} I_r & \mathbf{0} \\ -\bar{U} & I_{m^k-r} \end{bmatrix} H_{.,k} \begin{bmatrix} I_r & -\bar{V} \\ \mathbf{0} & I_{m^k-r} \end{bmatrix} = \begin{bmatrix} F_{I,J} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \\ F_{I,J}^{(\mathbf{u})} & F_{I,M^k\backslash J}^{(\mathbf{u})} - F_{I,J}^{(\mathbf{u})}\bar{V}, \mathbf{u} \in \mathcal{M}^* \text{in flo} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

So if $F_{I,M^k\backslash J}^{(\mathbf{u})} \neq F_{I,J}^{(\mathbf{u})}\bar{V}$ for some $\mathbf{u} \in \mathcal{M}^*$, then Rank$(H_{.,k})$ would exceed Rank$(F_{I,J})$, which is a contradiction. Thus it follows that

$$F_{I,M^k\backslash J}^{(\mathbf{u})} = F_{I,J}^{(\mathbf{u})}\bar{V}, \ \forall \mathbf{u} \in \mathcal{M}^*. \tag{7.11}$$

The two relationships (7.10) and (7.11) can together be compactly expressed as (7.9), which is the desired conclusion. ■

Lemma 7.3 *Choose unique matrices $\bar{D}^{(u)}, u \in \mathcal{M}$, such that*

$$F_{I,J}^{(u)} = F_{I,J}\bar{D}^{(u)}, \ \forall u \in \mathcal{M}. \tag{7.12}$$

Then for all $\mathbf{u} \in \mathcal{M}^*$, we have

$$F_{I,J}^{(\mathbf{u})} = F_{I,J}\bar{D}^{(u_1)} \dots \bar{D}^{(u_l)}, \ \textit{where } l = |\mathbf{u}|. \tag{7.13}$$

Choose unique matrices $D^{(u)}, u \in \mathcal{M}$, such that

$$F_{I,J}^{(u)} = D^{(u)}F_{I,J}, \ \forall u \in \mathcal{M}. \tag{7.14}$$

Then for all $\mathbf{u} \in \mathcal{M}^*$, we have

$$F_{I,J}^{(\mathbf{u})} = D^{(u_1)} \dots D^{(u_l)}F_{I,J}, \ \textit{where } l = |\mathbf{u}|. \tag{7.15}$$

This result can be compared to [1], Theorem 1, p. 90.

Proof: We prove only (7.13), since the proof of (7.15) is entirely similar. By the manner in which the index sets I, J are chosen, we have

$$\text{Rank}[F_{I,J} \ \ F_{I,J}^{(u)}] = \text{Rank}[F_{I,J}], \ \forall u \in \mathcal{M}.$$

Hence there exist unique matrices $\bar{D}^{(u)}, u \in \mathcal{M}$ such that (7.12) holds. Now suppose $\mathbf{v}$ is any *nonempty* string in $\mathcal{M}^*$. Then, since $F_{I,J}$ is a maximal rank submatrix of H, it follows that

$$\text{Rank} \begin{bmatrix} F_{I,J} & F_{I,J}^{(u)} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} \end{bmatrix} = \text{Rank} \begin{bmatrix} F_{I,J} \\ F_{I,J}^{(\mathbf{v})} \end{bmatrix}, \ \forall u \in \mathcal{M}.$$

Now post-multiply the matrix on the left side as shown below:

$$\begin{bmatrix} F_{I,J} & F_{I,J}^{(u)} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} \end{bmatrix} \begin{bmatrix} I & -\bar{D}^{(u)} \\ \mathbf{0} & I \end{bmatrix} = \begin{bmatrix} F_{I,J} & \mathbf{0} \\ F_{I,J}^{(\mathbf{v})} & F_{I,J}^{(\mathbf{v}u)} - F_{I,J}^{(\mathbf{v})}\bar{D}^{(u)} \end{bmatrix}.$$

This shows that

$$F_{I,J}^{(\mathbf{v}u)} = F_{I,J}^{(\mathbf{v})}\bar{D}^{(u)},\ \forall \mathbf{v} \in \mathcal{M}^*,\ \forall u \in \mathcal{M}. \tag{7.16}$$

Otherwise, the (2,2)-block of the matrix on the right side would contain a nonzero element and would therefore have rank larger than that of $F_{I,J}$, which would be a contradiction. Note that if $\mathbf{v}$ is the empty string in (7.16), then we are back to the definition of the matrix $\bar{D}^{(u)}$. Now suppose $\mathbf{u} \in \mathcal{M}^*$ has length l and apply (7.16) recursively. This leads to the desired formula (7.13). The proof of (7.15) is entirely similar. ■

Suppose $\mathbf{u} \in \mathcal{M}^*$ has length l. Then it is natural to define

$$\bar{D}^{(\mathbf{u})} := \bar{D}^{(u_1)} \dots \bar{D}^{(u_l)},\ D^{(\mathbf{u})} := D^{(u_1)} \dots D^{(u_l)}.$$

With this notation let us observe that the matrices $D^{(\mathbf{u})}$ and $\bar{D}^{(\mathbf{u})}$ 'intertwine' with the matrix $F_{I,J}$. That is,

$$F_{I,J}\bar{D}^{(\mathbf{u})} = D^{(\mathbf{u})}F_{I,J},\ \text{and}\ F_{I,J}^{-1}D^{(\mathbf{u})} = \bar{D}^{(\mathbf{u})}F_{I,J}^{-1}. \tag{7.17}$$

This follows readily from the original relationship

$$F_{I,J}\bar{D}^{(u)} = D^{(u)}F_{I,J}(= F_{I,J}^{(u)})\ \forall u \in \mathcal{M}$$

applied recursively.

Finally we come to the main theorem about quasi-realizations. We begin by formalizing the notion.

Note that a regular quasi-realization in some sense completes the analogy with the formulas (4.3) and (4.4).

Theorem 7.1 *Suppose the process $\{\mathcal{Y}_t\}$ has finite Hankel rank, say r. Then the process always has a regular quasi-realization. In particular, choose the integer k as in Lemma 7.1, and choose index sets $I, J \subseteq \mathcal{M}^k$ such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. Define the matrices $U, V, D^{(u)}, \bar{D}^{(u)}$ as before. The following two choices are regular quasi-realizations. First, let*

$$\mathbf{x} = \theta := F_{0,J}F_{I,J}^{-1},\ \mathbf{y} = \phi := F_{I,0},\ C^{(u)} = D^{(u)}\ \forall u \in \mathcal{M}. \tag{7.18}$$

Second, let

$$\mathbf{x} = \bar{\theta} := F_{0,J},\ \mathbf{y} = \bar{\phi} := F_{I,J}^{-1}F_{I,0},\ C^{(u)} = \bar{D}^{(u)}\ \forall u \in \mathcal{M}. \tag{7.19}$$

This result can be compared to [1], Theorem 1, p. 90 and Theorem 2, p. 92.

Proof: With all the spade work done already, the proof is very simple. For any string $\mathbf{u} \in \mathcal{M}^*$, it follows from (7.14) that

$$F_{I,J}^{(\mathbf{u})} = D^{(u_1)} \dots D^{(u_l)} F_{I,J}, \text{ where } l = |\mathbf{u}|.$$

Next, we have from (7.9) that

$$F_{k,k}^{(\mathbf{u})} = U F_{I,J}^{(\mathbf{u})} V, \ \forall \mathbf{u} \in \mathcal{M}^*.$$

Now observe that, by definition, we have

$$f_{\mathbf{u}} = \sum_{\mathbf{i} \in \mathcal{M}^k} \sum_{\mathbf{j} \in \mathcal{M}^k} f_{\mathbf{iuj}} = \mathbf{e}_{m^k}^t F_{k,k}^{(\mathbf{u})} \mathbf{e}_{m^k} = \mathbf{e}_{m^k}^t U D^{(u_1)} \dots D^{(u_l)} F_{I,J} V \mathbf{e}_{m^k},$$

where $\mathbf{e}_{m^k}$ is the column vector with m^k one's. Hence (7.1) is satisfied with the choice

$$n = r, \theta := \mathbf{e}_{m^k}^t U, \phi := F_{I,J} V \mathbf{e}_{m^k}, C^{(u)} = D^{(u)} \ \forall u \in \mathcal{M},$$

and the matrices $D^{(u)}$ as defined in (7.14). Since $D^{(\mathbf{u})} F_{I,J} = F_{I,J} \bar{D}^{(\mathbf{u})}$, we can also write

$$f_{\mathbf{u}} = \mathbf{e}_{m^k}^t U F_{I,J} \bar{D}^{(u_1)} \dots \bar{D}^{(u_l)} V \mathbf{e}_{m^k}.$$

Hence (7.1) is also satisfied with the choice

$$n = r, \bar{\theta} := \mathbf{e}_{m^k}^t U F_{I,J}, \bar{\phi} := V \mathbf{e}_{m^k}, C^{(u)} = \bar{D}^{(u)} \ \forall u \in \mathcal{M},$$

and the matrices $\bar{D}^{(u)}$ as defined in (7.13).

Next, we show that the vectors $\theta, \phi, \bar{\theta}, \bar{\phi}$ can also be written as in (7.18) and (7.19). For this purpose, we proceed as follows:

$$\theta = \mathbf{e}_{m^k}^t U = \mathbf{e}_{m^k}^t U F_{I,J} F_{I,J}^{-1} = \mathbf{e}_{m^k}^t F_{k,J} F_{I,J}^{-1} = F_{0,J} F_{I,J}^{-1}.$$

Therefore

$$\bar{\theta} = \theta F_{I,J} = F_{0,J}.$$

Similarly

$$\phi = F_{I,J} V \mathbf{e}_{m^k} = F_{I,k} \mathbf{e}_{m^k} = F_{I,0},$$

and

$$\bar{\phi} = F_{I,J}^{-1} F_{I,0}.$$

It remains only to prove the eigenvector properties. For this purpose, note that, for each $u \in \mathcal{M}$, we have

$$F_{0,J}\bar{D}^{(u)} = \mathbf{e}_{m^k}^t U F_{I,J}\bar{D}^{(u)} = \mathbf{e}_{m^k}^t U F_{I,J}^{(u)} = F_{0,J}^{(u)}.$$

Now

$$\theta D^{(u)} = F_{0,J}F_{I,J}^{-1}D^{(u)} = F_{0,J}\bar{D}^{(u)}F_{I,J}^{-1} = F_{0,J}^{(u)}F_{I,J}^{-1}.$$

Hence

$$\theta\left[\sum_{u\in\mathcal{M}} D^{(u)}\right] = \sum_{u\in\mathcal{M}} \theta D^{(u)} = \sum_{u\in\mathcal{M}} F_{0,J}^{(u)}F_{I,J}^{-1} = F_{0,J}F_{I,J}^{-1} = \theta,$$

since

$$\sum_{u\in\mathcal{M}} F_{0,J}^{(u)} = F_{0,J}.$$

As for ϕ, we have

$$D^{(u)}\phi = D^{(u)}F_{I,J}V\mathbf{e}_{m^k} = F_{I,J}^{(u)}V\mathbf{e}_{m^k} = F_{I,k}^{(u)}\mathbf{e}_{m^k} = F_{I,0}^{(u)}.$$

Hence

$$\left[\sum_{u\in\mathcal{M}} D^{(u)}\right]\phi = \sum_{u\in\mathcal{M}} D^{(u)}\phi = \sum_{u\in\mathcal{M}} F_{I,0}^{(u)} = F_{I,0} = \phi.$$

This shows that $\{r, \theta, \phi, D^{(u)}\}$ is a quasi-realization. The proof in the case of the barred quantities is entirely similar. We have

$$\bar{\theta}\bar{D}^{(u)} = F_{0,J}\bar{D}^{(u)} = F_{0,J}^{(u)},$$

so

$$\bar{\theta}\left[\sum_{u\in\mathcal{M}} \bar{D}^{(u)}\right] = \sum_{u\in\mathcal{M}} F_{0,J}^{(u)} = F_{0,J} = \bar{\theta}.$$

It can be shown similarly that

$$\left[\sum_{u\in\mathcal{M}} \bar{D}^{(u)}\right]\bar{\phi} = \bar{\phi}.$$

This completes the proof. ■

Next, it is shown that any two 'regular' quasi-realizations of the process are related through a similarity transformation.

Theorem 7.2 *Suppose a process $\{\mathcal{Y}_t\}$ has finite Hankel rank r, and suppose $\{\theta_1, \phi_1, D_1^{(u)}, u \in \mathcal{M}\}$ and $\{\theta_2, \phi_2, D_2^{(u)}, u \in \mathcal{M}\}$ are two regular quasi-realizations of this process. Then there exists a nonsingular matrix T such that*

$$\theta_2 = \theta_1 T^{-1}, D_2^{(u)} = T D_1^{(u)} T^{-1}\ \forall u \in \mathcal{M}, \phi_2 = T\phi_1.$$

Proof: Suppose the process has finite Hankel rank, and let r denote the rank of H. Choose the integer k as before, namely, the smallest integer k such that $\text{Rank}(F_{k,k}) = \text{Rank}(H)$. Choose subsets $I, J \subseteq \mathcal{M}^k$ such that $|I| = |J| = r$ and $\text{Rank}(F_{I,J}) = r$. Up to this point, all entities depend only on the process and its Hankel matrix (which depends on the law of the process), and not on the specific quasi-realization. Moreover, the fact that I, J are not unique is not important.

Now look at the matrix $F_{I,J}$, and express it in terms of the two quasi-realizations. By definition,

$$F_{I,J} = \begin{bmatrix} f_{\mathbf{i}_1\mathbf{j}_1} & \cdots & f_{\mathbf{i}_1\mathbf{j}_r} \\ \vdots & \vdots & \vdots \\ f_{\mathbf{i}_r\mathbf{j}_1} & \cdots & f_{\mathbf{i}_r\mathbf{j}_r} \end{bmatrix}.$$

Now, since we are given two quasi-realizations, the relationship (7.1) holds for each quasi-realization. Hence

$$F_{I,J} = \begin{bmatrix} \theta_s D_s^{(\mathbf{i}_1)} \\ \vdots \\ \theta_s D_s^{(\mathbf{i}_r)} \end{bmatrix} [D_s^{(\mathbf{j}_1)}\phi_s \ldots D_s^{(\mathbf{j}_r)}\phi_s], \text{ for } s = 1, 2.$$

Define

$$P_s := \begin{bmatrix} \theta_s D_s^{(\mathbf{i}_1)} \\ \vdots \\ \theta_s D_s^{(\mathbf{i}_r)} \end{bmatrix}, \; Q_s := [D_s^{(\mathbf{j}_1)}\phi_s \ldots D_s^{(\mathbf{j}_r)}\phi_s], \text{ for } s = 1, 2.$$

Then $F_{I,J} = P_1 Q_1 = P_2 Q_2$. Since $F_{I,J}$ is nonsingular, so are P_1, Q_1, P_2, Q_2. Moreover,

$$P_2^{-1} P_1 = Q_2 Q_1^{-1} =: T, \text{ say.}$$

Next, fix $u \in \mathcal{M}$ and consider the $r \times r$ matrix $F_{I,J}^{(u)}$. We have from (7.1) that

$$F_{I,J}^{(u)} = P_1 D_1^{(u)} Q_1 = P_2 D_2^{(u)} Q_2.$$

Hence

$$D_2^{(u)} = P_2^{-1} P_1 D_1^{(u)} Q_1 Q_2^{-1} = T D_1^{(u)} T^{-1}, \; \forall u \in \mathcal{M}.$$

Finally, we can factor the entire matrix H as

$$H = [\theta_s D_s^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^* \text{ in flo}][D_s^{(\mathbf{v})}\phi_s, \mathbf{v} \in \mathcal{M}^* \text{ in llo}], \; s = 1, 2,$$

where

$$D^{(\mathbf{u})} := D^{(u_1)} \ldots D^{(u_l)}, \; l = |\mathbf{u}|,$$

and $D^{(\mathbf{v})}$ is defined similarly. Note that the first matrix in the factorization of H has r columns and infinitely many rows, while the second matrix has r rows and infinitely many columns. Thus there exists a nonsingular matrix, say S, such that

$$[\theta_2 D_2^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^* \text{ in flo}] = [\theta_1 D_1^{(\mathbf{u})}, \mathbf{u} \in \mathcal{M}^* \text{ in flo}]S^{-1},$$

and

$$[D_2^{(\mathbf{v})}\phi_2, \mathbf{v} \in \mathcal{M}^* \text{ in llo}] = S[D_1^{(\mathbf{v})}\phi_1, \mathbf{v} \in \mathcal{M}^* \text{ in llo}].$$

Choosing $\mathbf{u} = \mathbf{i}_1, \ldots, \mathbf{i}_r$ and $\mathbf{v} = \mathbf{j}_1, \ldots, \mathbf{j}_r$ shows that in fact $S = T$. Finally, choosing $\mathbf{u} = \mathbf{v} = \emptyset$ shows that

$$\theta_2 = \theta_1 T^{-1}, \ \phi_2 = T\phi_1.$$

This completes the proof. ■

We conclude this section with an example from [14] of a regular quasi-realization that does not correspond to a regular realization.

Let $n = 4$, and define the 4×4 'state transition matrix'

$$A = \begin{bmatrix} \lambda_1 & 0 & 0 & 1-\lambda_1 \\ 0 & -\lambda_2 & 0 & 1+\lambda_2 \\ 0 & 0 & -\lambda_3 & 1+\lambda_3 \\ 1-\lambda_1 & c(1+\lambda_2) & -c(1+\lambda_3) & \lambda_1 + c(\lambda_3 - \lambda_2) \end{bmatrix},$$

as well as the 'output matrix'

$$B = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

It is easy to see that $A\mathbf{e}_4 = \mathbf{e}_4$, that is, the matrix A is 'stochastic.' Similarly $B\mathbf{e}_2 = \mathbf{e}_2$ and so B is stochastic (without quotes). Let $\mathbf{b}_i$ denote the i-th column of B, and let $\text{Diag}(\mathbf{b}_i)$ denote the diagonal 4×4 matrix with the elements of $\mathbf{b}_i$ on the diagonal. Let us define

$$C^{(1)} = A\text{Diag}(\mathbf{b}_1) = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & -\lambda_2 & 0 & 0 \\ 0 & 0 & -\lambda_3 & 0 \\ 1-\lambda_1 & c(1+\lambda_2) & -c(1+\lambda_3) & 0 \end{bmatrix},$$

$$C^{(2)} = A\text{Diag}(\mathbf{b}_2) = \begin{bmatrix} 0 & 0 & 0 & 1-\lambda_1 \\ 0 & 0 & 0 & 1+\lambda_2 \\ 0 & 0 & 0 & 1+\lambda_3 \\ 0 & 0 & 0 & \lambda_1 + c(\lambda_3 - \lambda_2) \end{bmatrix}.$$

Then $C^{(1)} + C^{(2)} = A$. Note that

$$\mathbf{x} = [0.5 \;\; 0.5c \;\; -0.5c \;\; 0.5]$$

is a 'stationary distribution' of A; that is, $\mathbf{x}A = \mathbf{x}$. With these preliminaries, we can define the 'quasi-frequencies'

$$f_{\mathbf{u}} = \mathbf{x}C^{(u_1)} \dots C^{(u_l)}\mathbf{e}_4,$$

where $\mathbf{u} = u_1 \dots u_l$. Because $\mathbf{x}$ and $\mathbf{e}_4$ are respectively row and column eigenvectors of A corresponding to the eigenvalue one, these quasi-frequencies satisfy the consistency conditions (4.1) and (4.2). Thus, in order to qualify as a quasi-realization, the only thing missing is the property that $f_{\mathbf{u}} \geq 0$ for all strings $\mathbf{u}$.

This nonegativity property is established in [14] using a Markov chain analogy, and is not reproduced here. All the frequencies will all be nonnegative provided the following inequalities are satisfied:

$$0 < \lambda_i < 1, i = 1, 2, 3; \lambda_1 > \lambda_i, i = 2, 3; 0 < c < 1,$$

$$\lambda_1 + c(\lambda_3 - \lambda_2) > 0; (1 - \lambda_1)^k > c(1 + \lambda_i)^k, i = 2, 3, k = 1, 2.$$

One possible choice (given in [14]) is

$$\lambda_1 = 0.5, \lambda_2 = 0.4, \lambda_3 = 0.3, c = 0.06.$$

Thus the above is a quasi-realization.

To test whether this quasi-realization can be made into a realization (with nonnegative elements), we can make use of Theorem 7.2. *All possible* quasi-realizations of this process can be obtained by performing a similarity transformation on the above quasi-realization. Thus there exists a regular realization (not quasi-realization) of this process if and only if there exists a nonsingular matrix T such that $\mathbf{x}T^{-1}, TC^{(i)}T^{-1}, T\mathbf{e}_4$ are all nonnegative. This can in turn be written as the feasibility of a linear program, namely:

$$\pi T = \mathbf{x}; TC^{(i)} = M^{(i)}T, i = 1, 2; T\mathbf{e}_4 = \mathbf{e}_4; M^{(i)} \geq \mathbf{0}, i = 1, 2; \pi \geq \mathbf{0}.$$

It can be readily verified that the above linear program is *not* feasible, so that there is no regular realization for this process, only regular quasi-realizations.

As pointed out above, it is possible to check in polynomial time whether a given regular quasi-realization can be converted into a regular realization of a stationary process. There is

a related problem that one can examine, namely: Suppose one is given a triplet $\{\mathbf{x}, C^{(u)}, u \in \mathcal{M}, \mathbf{y}\}$ with compatible dimensions. The problem is to determine whether the triple product

$$f_{\mathbf{u}} := \mathbf{x}C^{(\mathbf{u})}\mathbf{y} = \mathbf{x}C^{(u_1)} \cdots C^{(u_l)}\mathbf{y} \geq 0 \ \forall \mathbf{u} \in \mathcal{M}^l, \ \forall l.$$

This problem can be viewed as one of deciding whether a given rational power series always has nonnegative coefficients. This problem is known to be undecidable; see [36], Theorem 3.13. Even if $m = 2$, the above problem is undecidable if $n \geq 50$, where n is the size of the vector $\mathbf{x}$. The arguments of [9] can be adapted to prove this claim.[8] Most likely the problem remains undecidable even if we add the additional requirements that

$$\mathbf{x}\left[\sum_{u\in\mathcal{M}} C^{(u)}\right] = \mathbf{x},$$

$$\left[\sum_{u\in\mathcal{M}} C^{(u)}\right]\mathbf{y} = \mathbf{y},$$

because the above two conditions play no role in determining the nonnegativity or otherwise of the 'quasi-frequencies' $f_{\mathbf{u}}$, but serve only to assure that these quasi-frequencies are consistent.

8 Spectral Properties of Alpha-Mixing Processes

In this section, we add the assumption that the finite Hankel rank process under study is also α-mixing, and show that the regular quasi-realizations have an additional property, namely: The matrix that plays the role of the state transition matrix in the HMM has a spectral radius of one, this eigenvalue is simple, and all other eigenvalues have magnitude strictly less than one. This property is referred to as the 'quasi strong Perron property.' As a corollary, it follows that if an α-mixing process has a regular realization (and not just a quasi-realization), then the underlying Markov chain is irreducible and aperiodic.

We begin by reminding the reader about the notion of α-mixing. Suppose the process $\{\mathcal{Y}_t\}$ is defined on the probability space (S, Ω), where Ω is a σ-algebra on the set S. For each pair of indices s, t with $s < t$, define Σ_s^t to be the σ-algebra (a subalgebra of Ω) generated by the random variables $\mathcal{Y}_s, \ldots, \mathcal{Y}_t$. Then the **$\alpha$-mixing coefficient** $\alpha(l)$ of the process $\{\mathcal{Y}_t\}$ is defined as

$$\alpha(l) := \sup_{A\in\Sigma_0^t, B\in\Sigma_{t+l}^{\infty}} |P(A \cap B) - P(A)P(B)|.$$

[8] Thanks to Vincent Blondel for these references.

The process $\{\mathcal{Y}_t\}$ is said to be α-**mixing** if $\alpha(l) \to 0$ as $l \to \infty$. Note that in the definition above, A is an event that depends strictly on the 'past' random variables before time t, whereas B is an event that depends strictly on the 'future' random variables after time $t+l$. If the future were to be completely independent of the past, we would have $P(A \cap B) = P(A)P(B)$. Thus the α-mixing coefficient measures the extent to which the future is independent of the past.

Remark: As will be evident from the proofs below, actually we do *not* make use of the α-mixing property of the process $\{\mathcal{Y}_t\}$. Rather, what is needed is that

$$\sum_{\mathbf{w} \in \mathcal{M}^l} f_{\mathbf{uwv}} \to f_{\mathbf{u}} f_{\mathbf{v}} \text{ as } l \to \infty, \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^k, \tag{8.1}$$

where k is the *fixed* integer arising from the finite Hankel rank condition. Since the process assumes values in a finite alphabet, (8.1) is equivalent to the condition

$$\max_{A \in \Sigma_1^k, B \in \Sigma_{l+k+1}^{2k}} |P(A \cap B) - P(A)P(B)| \to 0 \text{ as } l \to \infty. \tag{8.2}$$

To see this, suppose that (8.2) holds, and choose A to be the event $(y_1, \ldots, y_k) = \mathbf{u}$, and similarly, choose B to be the event $(y_{l+k+1}, \ldots, y_{l+2k}) = \mathbf{v}$, for some $\mathbf{u}, \mathbf{v} \in \mathcal{M}^k$. Then it is clear that $A \cap B$ is the event that a string of length $l+2k$ begins with $\mathbf{u}$ and ends with $\mathbf{v}$. Thus

$$P(A) = f_{\mathbf{u}}, \ P(B) = f_{\mathbf{v}}, \ P(A \cap B) = \sum_{\mathbf{w} \in \mathcal{M}^l} f_{\mathbf{uwv}}.$$

Hence (8.2) implies (8.1). To show the converse, suppose (8.1) holds. Then (8.2) also holds for elementary events A and B. Since k is a *fixed* number and the alphabet of the process is finite, both of the σ-algebras Σ_1^k, Σ_{l+k+1}^{2k} are *finite* unions of elementary events. Hence (8.1) is enough to imply (8.2). It is not known whether (8.2) is *strictly weaker* than α-mixing for processes assuming values over a finite alphabet.

Now we state the main result of this section.

Theorem 8.1 *Suppose the process $\{\mathcal{Y}_t\}$ is α-mixing and has finite Hankel rank r. Let $\{r, \mathbf{x}, \mathbf{y}, C^{(\mathbf{u})}, u \in \mathcal{M}\}$ be any regular quasi-realization of the process, and define*

$$S := \sum_{u \in \mathcal{M}} C^{(\mathbf{u})}.$$

Then $S^l \to \mathbf{yx}$ as $l \to \infty$, $\rho(S) = 1$, $\rho(S)$ is a simple eigenvalue of S, and all other eigenvalues of S have magnitude strictly less than one.

This theorem can be compared with [1], Theorem 4, p. 94.

Proof: It is enough to prove the theorem for the *particular* quasi-realization $\{r, \theta, \phi, D^{(u)}, u \in \mathcal{M}\}$ defined in (7.2). This is because there exists a nonsingular matrix T such that $C^{(u)} = T^{-1}D^{(u)}T$ for all u, and as a result the matrices $\sum_{u\in\mathcal{M}} C^{(u)}$ and $\sum_{u\in\mathcal{M}} D^{(u)}$ have the same spectrum. The α-mixing property implies that, for each $\mathbf{i} \in I, \mathbf{j} \in J$, we have

$$\sum_{\mathbf{w}\in\mathcal{M}^l} f_{\mathbf{iwj}} \to f_{\mathbf{i}} f_{\mathbf{j}} \text{ as } l \to \infty. \tag{8.3}$$

This is a consequence of (8.1) since both I and J are subsets of $\mathcal{M}^k$. Now note that, for each fixed $\mathbf{w} \in \mathcal{M}^l$, we have from (7.1) that

$$[f_{\mathbf{iwj}}, \mathbf{i} \in I, \mathbf{j} \in J] = F_{I,J}^{(\mathbf{w})} = D^{(\mathbf{w})} F_{I,J}, \tag{8.4}$$

where, as per earlier convention, we write

$$D^{(\mathbf{w})} := D^{(w_1)} \dots D^{(w_l)}.$$

It is clear that

$$\sum_{\mathbf{w}\in\mathcal{M}^l} D^{(\mathbf{w})} = \left[\sum_{u\in\mathcal{M}} D^{(u)}\right]^l = S^l. \tag{8.5}$$

Now (8.3) implies that

$$\sum_{\mathbf{w}\in\mathcal{M}^l} [f_{\mathbf{iwj}}, \mathbf{i} \in I, \mathbf{j} \in J] \to [f_{\mathbf{i}}, i \in I][f_{\mathbf{j}}, j \in J] =: F_{I,0}F_{0,J},$$

where $F_{I,0}$ is an r-dimensional column vector and $F_{0,J}$ is an r-dimensional row vector. Moreover, combining (8.4) and (8.5) shows that

$$S^l F_{I,J} \to F_{I,0}F_{0,J},$$

and since $F_{I,J}$ is nonsingular, that

$$S^l \to F_{I,0}F_{0,J}F_{I,J}^{-1} = \phi\theta \text{ as } l \to \infty.$$

So the conclusion is that S^l approaches $\phi\theta$, which is a rank one matrix, as $l \to \infty$. Moreover, this rank one matrix has one eigenvalue at one and the rest at zero. To establish this, we show that

$$F_{0,J}F_{I,J}^{-1}F_{I,0} = 1.$$

This is fairly straight-forward. Note that $F_{0,J}F_{I,J}^{-1} = \theta$ and $F_{I,0} = \phi$ as defined in (7.2). Then taking $\mathbf{u}$ to be the empty string in (7.1) (and of course, substituting $\mathbf{x} = \theta, \mathbf{y} = \phi$) shows that $\theta\phi = 1$, which is the desired conclusion. Let A denote the rank one matrix

$$A := F_{I,0}F_{0,J}F_{I,J}^{-1}.$$

Then $S^l \to A$ as $l \to \infty$. Suppose the spectrum of the matrix S is $\{\lambda_1, \ldots, \lambda_n\}$, where $n = m^k$, and $|\lambda_1| = \rho(S)$. Then, since the spectrum of S^l is precisely $\{\lambda_1^l, \ldots, \lambda_n^l\}$, it follows that

$$\{\lambda_1^l, \ldots, \lambda_n^l\} \to \{1, 0, \ldots, 0\} \text{ as } l \to \infty.$$

Here we make use of the facts that A is a rank one matrix, and that its spectrum consists of $n-1$ zeros plus one. This shows that S has exactly one eigenvalue on the unit circle, namely at $\lambda = 1$, and the remaining eigenvalues are all inside the unit circle.

Corollary 8.1 *Suppose a stationary process $\{\mathcal{Y}_t\}$ is α-mixing and has a regular realization. Then the underlying Markov chain is aperiodic and irreducible.*

Proof: Suppose that the process under study has a regular realization (and not just a regular quasi-realization). Let A denote the state transition matrix of the corresponding Markov process $\{\mathcal{X}_t\}$. From Theorem 7.2, it follows that A is similar to the matrix S defined in Theorem 8.1. Moreover, if the process $\{\mathcal{Y}_t\}$ is α-mixing, then the matrix A (which is similar to S) satisfies the strong Perron property. In other words, it has only one eigenvalue on the unit circle, namely a simple eigenvalue at one. Hence the Markov chain $\{\mathcal{X}_t\}$ is irreducible and aperiodic. ■

9 Ultra-Mixing Processes and the Existence of HMM's

In the previous two sections, we studied the existence of quasi-realizations. In this section, we study the existence of 'true' (as opposed to quasi) realizations. We introduce a new property known as 'ultra-mixing' and show that if a process has finite Hankel rank, and is both α-mixing as well as ultra-mixing, then modulo a technical condition it has a HMM where the underlying Markov chain is itself α-mixing (and hence aperiodic and irreducible) or else satisfies a 'consistency condition.' The converse is also true, modulo another technical condition.

The material in this section is strongly influenced by [1]. In that paper, the author *begins* with the assumption that the stochastic process under study is generated by an irreducible

HMM (together with a few other assumptions), and then gives a constructive procedure for constructing an irreducible HMM for the process. Thus the paper does not give a set of conditions for the existence of a HMM *in terms of the properties of the process under study*. Moreover, even with the assumptions in [1], the order of the HMM constructed using the given procedure can in general be much larger than the order of the HMM that generates the process in the first place. In contrast, in the present paper we give conditions explicitly in terms of the process under study, that are sufficient to guarantee the existence of an irreducible HMM. However, the proof techniques used here borrow heavily from [1].

9.1 Constructing a Hidden Markov Model

We begin with a rather 'obvious' result that sets the foundation for the material to follow.

Lemma 9.1 *Suppose $\{\mathcal{Y}_t\}$ is a stationary process over a finite alphabet $\mathcal{M}$. Then the process $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM if and only if there exist an integer n, a stochastic row vector $\mathbf{h}$, and $n \times n$ nonnegative matrices $G^{(1)}, \ldots, G^{(m)}$ such that the following statements are true.*

1. *The matrix $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is stochastic, in that each of its rows adds up to one. Equivalently, $\mathbf{e}_n$ is a column eigenvector of Q corresponding to the eigenvalue one.*
2. *$\mathbf{h}$ is a row eigenvector $\mathbf{h}$ of Q corresponding to the eigenvalue one, i.e., $\mathbf{h}Q = \mathbf{h}$.*
3. *For every $\mathbf{u} \in \mathcal{M}^*$, we have*

$$f_{\mathbf{u}} = \mathbf{h} G^{(u_1)} \cdots G^{(u_l)} \mathbf{e}_n,$$

where $l = |\mathbf{u}|$.

In this case there exists a Markov process $\{\mathcal{X}_t\}$ evolving over $\mathcal{N} := \{1, \ldots, n\}$ such that the joint process $\{(\mathcal{X}_t, \mathcal{Y}_t)\}$ satisfies the conditions (3.1) and (3.2).

Proof: One half of this lemma has already been proven in the course of proving Theorem 4.1. Suppose $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM model. Let $\{\mathcal{X}_t\}$ denote the associated Markov process. Define the matrices $M^{(1)}, \ldots, M^{(m)}$ as in (3.3). and let π denote the stationary distribution of the process $\{\mathcal{X}_t\}$. Then it is clear that the conditions of the lemma are satisfied with $\mathbf{h} = \pi$ and $G^{(u)} = M^{(u)}$ for each $u \in \mathcal{M}$.

To prove the converse, suppose $\mathbf{h}, G^{(1)}, \ldots, G^{(m)}$ exist that satisfy the stated conditions. Let $\{\mathcal{Z}_t\}$ be a stationary Markov process with the state transition matrix

$$A_{\mathcal{Z}} := \begin{bmatrix} G^{(1)} & G^{(2)} & \ldots & G^{(m)} \\ \vdots & \vdots & \vdots & \vdots \\ G^{(1)} & G^{(2)} & \ldots & G^{(m)} \end{bmatrix}.$$

and the stationary distribution

$$\pi_{\mathcal{Z}} = [\mathbf{h}G^{(1)} | \ldots | \mathbf{h}G^{(m)}].$$

To show that $\pi_{\mathcal{Z}}$ is indeed a stationary distribution of $A_{\mathcal{Z}}$, partition $\pi_{\mathcal{Z}}$ in the obvious fashion as $[\pi_1 \ldots \pi_m]$, and observe that $\pi_v = \mathbf{h}G^{(v)}$. Then, because of the special structure of the matrix $A_{\mathcal{Z}}$, in order to be a stationary distribution of the Markov chain, the vector $\pi_{\mathcal{Z}}$ needs to satisfy the relationship

$$\left[\sum_{v \in \mathcal{M}} \pi_v\right] \cdot G^{(u)} = \pi_u. \tag{9.1}$$

Now observe that

$$\left[\sum_{v \in \mathcal{M}} \pi_v\right] = \mathbf{h} \sum_{v \in \mathcal{M}} G^{(v)} = \mathbf{h}Q = \mathbf{h}.$$

Hence the desired relationship (9.1) follows readily. Now the stationary distribution of the $\mathcal{X}_t$ process is clearly $\sum_{v \in \mathcal{M}} \mathbf{h}G^{(v)} = \mathbf{h}$. Hence, by the formula (4.3), it follows that the frequencies of the $\mathcal{Y}_t$ process are given by

$$f_{\mathbf{u}} = \mathbf{h}G^{(u_1)} \cdots G^{(u_l)}\mathbf{e}_n.$$

This is the desired conclusion. ■

9.2 The Consistency Condition

Before presenting the sufficient condition for the existence of a HMM, we recall a very important result from [1]. Consider a 'joint Markov process' HMM where the associated matrix A (the transition matrix of the $\{\mathcal{X}_t\}$ process) is irreducible. In this case, it is well known and anyway rather easy to show that the state process $\{\mathcal{X}_t\}$ is α-mixing if and only if the matrix A is aperiodic in addition to being irreducible. If A is aperiodic (so that the state process is α-mixing), then the output process $\{\mathcal{Y}_t\}$ is also α-mixing. However, the converse is not always true. It is possible for the output process to be α-mixing even if the state

process is not. Theorem 5 of [1] gives necessary and sufficient conditions for this to happen. We reproduce this important result below.

Suppose a 'joint Markov process' HMM has n states and that the state transition matrix A is irreducible. Let π denote the unique positive stationary probability distribution of the $\mathcal{X}_t$ process. As in (3.3), define the matrices $M^{(u)}, u \in \mathcal{M}$ by

$$m_{ij}^{(u)} = \Pr\{\mathcal{X}_1 = j \& \mathcal{Y}_1 = u | \mathcal{X}_0 = i\}, 1 \leq i, j \leq n, u \in \mathcal{M}.$$

Let p denote the number of eigenvalues of A on the unit circle (i.e., the period of the Markov chain). By renumbering the states if necessary, rearrange A so that it has the following cyclic form:

$$A = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & A_1 \\ A_p & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & A_{p-1} & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & A_2 & \mathbf{0} \end{bmatrix}, \tag{9.2}$$

where all blocks have the same size $(n/p) \times (n/p)$ (which clearly implies that p is a divisor of n). The matrices $M^{(u)}$ inherit the same zero block structure as A; so the notation $M_i^{(u)}$ is unambiguous. For a string $\mathbf{u} \in \mathcal{M}^l$, define

$$M_i^{(\mathbf{u})} := M_i^{(u_1)} M_{i+1}^{(u_2)} \dots M_{i+l-1}^{(u_l)},$$

where the subscripts on M are taken modulo p. Partition π into p equal blocks, and label them as π_1 through π_p.

Theorem 9.1 *The output process $\{\mathcal{Y}_t\}$ is α-mixing if and only if, for every string $\mathbf{u} \in \mathcal{M}^*$, the following 'consistency conditions' hold:*

$$\pi_1 M_1^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \pi_2 M_p^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \pi_3 M_{p-1}^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \dots = \pi_p M_2^{(\mathbf{u})} \mathbf{e}_{(n/p)} = \frac{1}{p} \pi M^{(\mathbf{u})} \mathbf{e}_n. \tag{9.3}$$

For a proof, see [1], Theorem 5.

9.3 The Ultra-Mixing Property

In earlier sections, we studied the spectrum of various matrices under the assumption that the process under study is α-mixing. For present purposes, we introduce a different kind of mixing property.

Definition 6 *Given the process $\{\mathcal{Y}_t\}$, suppose it has finite Hankel rank, and let k denote the unique integer defined in Lemma 7.1. Then the process $\{\mathcal{Y}_t\}$ is said to be* **ultra-mixing** *if there exists a sequence $\{\delta_l\} \downarrow 0$ such that*

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuv}}}{f_{\mathbf{uv}}} \right| \leq \delta_l, \ \forall \mathbf{i} \in \mathcal{M}^k, \mathbf{u} \in \mathcal{M}^l, \mathbf{v} \in \mathcal{M}^*. \tag{9.4}$$

Note that, the way we have defined it here, the notion of ultra-mixing is defined only for processes with finite Hankel rank.

In [28], Kalikow defines a notion that he calls a 'uniform martingale,' which is the same as an ultra-mixing stochastic process. He shows that a stationary stochastic process over a finite alphabet is a uniform martingale if and only if it is also a 'random Markov process,' which is defined as follows: A process $\{(\mathcal{Y}_t, N_t)\}$ where $\mathcal{Y}_t \in \mathcal{M}$ and N_t is a positive integer (natural number) for each t is said to be a 'random Markov process' if (i) The process $\{N_t\}$ is independent of the $\{\mathcal{Y}_t\}$ process, and (ii) for each t, we have

$$\Pr\{\mathcal{Y}_t | \mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \ldots\} = \Pr\{\mathcal{Y}_t | \mathcal{Y}_{t-1}, \mathcal{Y}_{t-2}, \ldots, \mathcal{Y}_{t-N_t}\}.$$

Observe that if N_t equals a fixed integer N for all t, then the above condition says that $\{\mathcal{Y}_t\}$ is an N-step Markov process. Hence a 'random Markov process' is an N_t-step Markov process where the length of the 'memory' N_t is itself random and independent of $\mathcal{Y}_t$. One of the main results of [28] is that the ultra-mixing property is equivalent to the process being random Markov. However, the random Markov property seems to be quite different in spirit from a process having a HMM.

The ultra-mixing property can be interpreted as a kind of long-term indepedence. It says that the conditional probability that a string begins with $\mathbf{i}$, given the next l entries, is just about the same whether we are given just the next l entries, or the next l entries as well as the still later entries. This property is also used in [1]. It does not appear straight-forward to relate ultra-mixing to other notions of mixing such as α-mixing. This can be seen from the treatment of [1], Section 11, where the author assumes (in effect) that the process under study is *both* ultra-mixing as well as α-mixing.

9.4 The Main Result

Starting with the original work of Dharmadhikari [15], 'cones' have played a central role in the construction of HMM's. The present paper continues that tradition. Moreover, cones also play an important role in the so-called positive realization problem. Hence it is not

surprising that the conditions given here also borrow a little bit from positive realization theory. See [6] for a survey of the current status of this problem.

Recall that a set $\mathcal{S} \subseteq \mathbb{R}^r$ is said to be a 'cone' if $\mathbf{x}, \mathbf{y} \in \mathcal{S} \ \Rightarrow \ \alpha\mathbf{x} + \beta\mathbf{y} \in \mathcal{S} \ \forall \alpha, \beta \geq 0$. The term 'convex cone' is also used to describe such an object. Given a (possibly infinite) set $\mathcal{V} \subseteq \mathbb{R}^r$, the symbol $\text{Cone}(\mathcal{V})$ denotes the smallest cone containing $\mathcal{V}$, or equivalently, the intersection of all cones containing $\mathcal{V}$. If $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ is a finite set, then it is clear that

$$\text{Cone}(\mathcal{V}) = \{\sum_{i=1}^{n} \alpha_i \mathbf{v}_i : \ \alpha_i \geq 0 \ \forall i\}.$$

In such a case, $\text{Cone}(\mathcal{V})$ is said to be 'polyhedral' and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are said to be 'generators' of the cone. Note that, in the way we have defined the concept here, the generators of a polyhedral cone are not uniquely defined. It is possible to refine the definition; however, the above definition is sufficient for the present purposes. Finally, given a cone $\mathcal{C}$ (polyhedral or otherwise), the 'polar cone' $\mathcal{C}^p$ is defined by

$$\mathcal{C}^p := \{\mathbf{y} \in \mathbb{R}^r : \mathbf{y}^t\mathbf{x} \geq 0 \ \forall \mathbf{x} \in \mathcal{C}\}.$$

It is easy to see that $\mathcal{C}^p$ is also a cone, and that $\mathcal{C} \subseteq (\mathcal{C}^p)^p$.

Next, we introduce two cones that play a special role in the proof. Suppose as always that the process under study has finite Hankel rank, and define the integer k as in Lemma 7.1. Throughout, we use the quasi-realization $\{r, \theta, \phi, D^{(u)}\}$ defined in (7.2). Now define

$$\mathcal{C}_c := \text{Cone}\{D^{(\mathbf{u})}\phi : \mathbf{u} \in \mathcal{M}^*\},$$

$$\mathcal{C}_o := \{\mathbf{y} \in \mathbb{R}^r : \theta D^{(\mathbf{v})}\mathbf{y} \geq 0, \ \forall \mathbf{v} \in \mathcal{M}^*\}.$$

The subscripts o and c have their legacy from positive realization theory, where $\mathcal{C}_c$ is called the 'controllability cone' and $\mathcal{C}_o$ is called the 'observability cone.' See for example [6]. However, in the present context, we could have used any other symbols. Note that from (7.1) and (7.2) we have

$$\theta D^{(\mathbf{v})} D^{(\mathbf{u})} \phi = f_{\mathbf{uv}} \geq 0, \ \forall \mathbf{u}, \mathbf{v} \in \mathcal{M}^*.$$

Hence $D^{(\mathbf{u})}\phi \in \mathcal{C}_o \ \forall \mathbf{u} \in \mathcal{M}^*$, and as a result $\mathcal{C}_c \subseteq \mathcal{C}_o$. Moreover, both $\mathcal{C}_c$ and $\mathcal{C}_o$ are invariant under $D^{(w)}$ for each $w \in \mathcal{M}$. To see this, let $w \in \mathcal{M}$ be arbitrary. Then $D^{(w)}D^{(\mathbf{u})}\phi = D^{(w\mathbf{u})}\phi$, for all $\mathbf{u} \in \mathcal{M}^*$. Hence

$$D^{(w)}\mathcal{C}_c = \text{Cone}\{D^{(w\mathbf{u})}\phi : \mathbf{u} \in \mathcal{M}^*\} \subseteq \mathcal{C}_c.$$

Similarly, suppose $\mathbf{y} \in \mathcal{C}_o$. Then the definition of $\mathcal{C}_o$ implies that $\theta D^{(\mathbf{v})}\mathbf{y} \geq 0$ for all $\mathbf{v} \in \mathcal{M}^*$. Therefore

$$\theta D^{(\mathbf{v})} D^{(w)}\mathbf{y} = \theta D^{(\mathbf{v}w)}\mathbf{y} \geq 0 \ \forall \mathbf{v} \in \mathcal{M}^*.$$

Hence $D^{(w)}\mathbf{y} \in \mathcal{C}_c$. The key difference between $\mathcal{C}_c$ and $\mathcal{C}_o$ is that the former cone need not be closed, whereas the latter cone is always closed (this is easy to show).

In order to state the sufficient condition for the existence of a HMM, a few other bits of notation are introduced. Suppose the process under study has finite Hankel rank, and let k be the unique integer defined in Lemma 7.1. Let r denote the rank of the Hankel matrix, and choose subsets $I, J \subseteq \mathcal{M}^k$ such that $|I| = |J| = r$ and $F_{I,J}$ has rank r. For each finite string $\mathbf{u} \in \mathcal{M}^*$, define the vectors

$$\mathbf{p_u} := \frac{1}{f_{\mathbf{u}}} F^{(\mathbf{u})}_{I,0} = [f_{\mathbf{iu}}/f_{\mathbf{u}}, i \in I] \in [0,1]^{r \times 1}, \ \mathbf{q_u} := \frac{1}{f_{\mathbf{u}}} F^{(\mathbf{u})}_{0,J} = [f_{\mathbf{uj}}/f_{\mathbf{u}}, j \in J] \in [0,1]^{1 \times r}.$$

The interpretation of $\mathbf{p_u}$ is that the $\mathbf{i}$-th component of this vector is the conditional probability, given that the last part of a sample path consists of the string $\mathbf{u}$, that the immediately preceding k symbols are $\mathbf{i}$. The vector $\mathbf{q_u}$ is interpreted similarly. The $\mathbf{j}$-th component of this vector is the conditional probability, given that the first part of a sample path consists of the string $\mathbf{u}$, that the next k symbols are $\mathbf{j}$.

Lemma 9.2 *Let $\| \cdot \|$ denote the ℓ_1-norm on $\mathbb{R}^r$. Then there exists a constant $\gamma > 0$ such that*

$$\gamma \leq \| \mathbf{p_u} \| \leq 1, \gamma \leq \| \mathbf{q_u} \| \leq 1, \ \forall \mathbf{u} \in \mathcal{M}^*.$$

Proof: Note that the vector $[f_{\mathbf{iu}}/f_{\mathbf{u}}, \mathbf{i} \in \mathcal{M}^k]$ is a probability vector, in the sense that its components are nonnegative and add up to one. Hence this vector has ℓ_1-norm of one. Since $\mathbf{p_u}$ is a subvector of it, it follows that $\| \mathbf{p_u} \| \leq 1$. On the other hand, we have

$$[f_{\mathbf{iu}}/f_{\mathbf{u}}, \mathbf{i} \in \mathcal{M}^k] = U\mathbf{p_u}, \ \forall \mathbf{u} \in \mathcal{M}^*,$$

and U has full column rank. Hence $\| \mathbf{p_u} \|$ is bounded away from zero independently of $\mathbf{u}$. Similar arguments apply to $\mathbf{q_u}$. ■

The vectors $\mathbf{p_u}$ and $\mathbf{q_u}$ satisfy some simple recurrence relationships.

Lemma 9.3 *Suppose $\mathbf{u}, \mathbf{v} \in \mathcal{M}^*$. Then*

$$D^{(\mathbf{u})}\mathbf{p_v} = \frac{f_{\mathbf{uv}}}{f_{\mathbf{v}}}\mathbf{p_{uv}}, \ \mathbf{q_u}C^{(\mathbf{v})} = \frac{f_{\mathbf{uv}}}{f_{\mathbf{u}}}\mathbf{q_{uv}}.$$

Proof: From Lemmas 7.2 and 7.3, it follows that

$$F_{I,0}^{(\mathbf{v})} = F_{I,k}^{(\mathbf{v})} \mathbf{e}_{m^k} = D^{(\mathbf{v})} F_{I,J} V \mathbf{e}_{m^k} = D^{(\mathbf{v})} \phi, \ \forall \mathbf{v} \in \mathcal{M}^*.$$

This shows that

$$\mathbf{p}_{\mathbf{v}} = \frac{1}{f_{\mathbf{v}}} F_{I,0}^{(\mathbf{v})} = \frac{1}{f_{\mathbf{v}}} D^{(\mathbf{v})} \phi.$$

Hence, for arbitrary $\mathbf{u}, \mathbf{v} \in \mathcal{M}^*$, we have

$$D^{(\mathbf{u})} \mathbf{p}_{\mathbf{v}} = \frac{1}{f_{\mathbf{v}}} D^{(\mathbf{u})} D^{(\mathbf{v})} \phi = \frac{1}{f_{\mathbf{v}}} D^{(\mathbf{uv})} \phi = \frac{f_{\mathbf{uv}}}{f_{\mathbf{v}}} \mathbf{p}_{\mathbf{uv}}.$$

The proof in the case of $\mathbf{q}_{\mathbf{v}}$ is entirely similar. ■

Now let us consider the countable collection of probability vectors $\mathcal{A} := \{\mathbf{p}_{\mathbf{u}} : \mathbf{u} \in \mathcal{M}^*\}$. Since $\mathbf{p}_{\mathbf{u}}$ equals $D^{(\mathbf{u})}\phi$ within a scale factor, it follows that $\mathcal{C}_c = \text{Cone}(\mathcal{A})$. Moreover, since $\mathcal{A} \subseteq \mathcal{C}_c \subseteq \mathcal{C}_o$ and $\mathcal{C}_o$ is a closed set, it follows that the set of cluster points of $\mathcal{A}$ is also a subset of $\mathcal{C}_o$.[9] Finally, it follows from Lemma 9.2 that every cluster point of $\mathcal{A}$ has norm no smaller than γ.

Now we state the main result of this section.

Theorem 9.2 *Suppose the process $\{\mathcal{Y}_t\}$ satisfies the following conditions:*

1. *It has finite Hankel rank.*
2. *It is ultra-mixing.*
3. *It is α-mixing.*
4. *The cluster points of the set $\mathcal{A}$ of probability vectors are finite in number and lie in the interior of the cone $\mathcal{C}_o$.*

Under these conditions, the process has an irreducible 'joint Markov process' hidden Markov model. Moreover the HMM satisfies the consistency conditions (9.3).

Remark: Among the hypotheses of Theorem 9.2, Conditions 1 through 3 are 'real' conditions, whereas Condition 4 is a 'technical' condition.

The proof proceeds via two lemmas. The first lemma gives insight into the behaviour of the matrix $D^{(\mathbf{u})}$ as $|\mathbf{u}| \to \infty$. To put these lemmas in context, define the matrix $S =$

[9] Recall that a vector $\mathbf{y}$ is said to be a 'cluster point' of $\mathcal{A}$ if there exists a sequence in $\mathcal{A}$, no entry of which equals $\mathbf{y}$, converging to $\mathbf{y}$. Equivalently, $\mathbf{y}$ is a cluster point if $\mathcal{A}$ if every neighbourhood of $\mathbf{y}$ contains a point of $\mathcal{A}$ not equal to $\mathbf{y}$.

$\sum_{u\in\mathcal{M}} D^{(u)}$. Then by Theorem 8.1, we know that if the process $\{\mathcal{Y}_t\}$ is α-mixing, then S^l approaches a rank one matrix as $l \to \infty$. In the present case it is shown that, if the process is ultra-mixing, then *each individual* matrix $D^{(\mathbf{u})}$ approaches a rank one matrix as $|\mathbf{u}| \to \infty$. This result has no counterpart in earlier literature and may be of independent interest.

Lemma 9.4 *Let $\|\cdot\|$ denote both the ℓ_1-norm of a vector in $\mathbb{R}^{m^k}$ as well as the corresponding induced norm on the set of $m^k \times m^k$ matrices. Suppose the process $\{\mathcal{Y}_t\}$ is ultra-mixing. Define*

$$\mathbf{b}_U := \mathbf{e}_{m^k}^t U \in \mathbb{R}^{1\times r}.$$

Then

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} - \frac{1}{f_{\mathbf{u}}} \mathbf{p}_{\mathbf{u}} \mathbf{b}_U D^{(\mathbf{u})} \| \leq r\delta_{|\mathbf{u}|} \| F_{I,J}^{-1} \|, \tag{9.5}$$

where $\{\delta_l\}$ is the sequence in the definition of the ultra-mixing property, and $|\mathbf{u}|$ denotes the length of the string u.

Proof: If we substitute $\mathbf{j}$ for $\mathbf{v}$ in (9.4), we get

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{uj}}} \right| \leq \delta_{|\mathbf{u}|}.$$

For each $\mathbf{j} \in J$, we have that $f_{\mathbf{uj}}/f_{\mathbf{u}} \leq 1$. Hence we can multiply both sides of the above equation by $f_{\mathbf{uj}}/f_{\mathbf{u}} \leq 1$, which gives

$$\left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{uj}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} \right| = \left| \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{u}}} \right| \leq \delta_{|\mathbf{u}|} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} \leq \delta_{|\mathbf{u}|}.$$

Now define the $r \times r$ matrix $R^{(\mathbf{u})}$ by

$$(R^{(\mathbf{u})})_{\mathbf{ij}} := \frac{f_{\mathbf{iu}}}{f_{\mathbf{u}}} \cdot \frac{f_{\mathbf{uj}}}{f_{\mathbf{u}}} - \frac{f_{\mathbf{iuj}}}{f_{\mathbf{u}}}.$$

Then (see for example [41])

$$\| R^{(\mathbf{u})} \| = \max_{\mathbf{j}\in\mathcal{M}^k} \sum_{\mathbf{i}\in\mathcal{M}^k} |(R^{(\mathbf{u})})_{\mathbf{ij}}| \leq r\delta_{|\mathbf{u}|}.$$

Next, note that

$$R^{(\mathbf{u})} = \mathbf{p}_{\mathbf{u}}\mathbf{q}_{\mathbf{u}} - \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} F_{I,J}.$$

Hence we have established that

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} F_{I,J} - \mathbf{p}_{\mathbf{u}}\mathbf{q}_{\mathbf{u}} \| \leq r\delta_{|\mathbf{u}|}. \tag{9.6}$$

Therefore

$$\| \frac{1}{f_{\mathbf{u}}} D^{(\mathbf{u})} - \mathbf{p}_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} F_{I,J}^{-1} \| \leq r \delta_{|\mathbf{u}|} \| F_{I,J}^{-1} \| .$$

Thus the proof is complete once it is shown that

$$\mathbf{q}_{\mathbf{u}} F_{I,J}^{-1} = \frac{1}{f_{\mathbf{u}}} \mathbf{b}_U D^{(\mathbf{u})}.$$

But this last step is immediate, because

$$f_{\mathbf{u}} \mathbf{q}_{\mathbf{u}} F_{I,J}^{-1} = F_{0,J}^{(\mathbf{u})} F_{I,J}^{-1} = \mathbf{e}_{m^k}^t U F_{I,J}^{(\mathbf{u})} F_{I,J}^{-1} == \mathbf{e}_{m^k}^t U D^{(\mathbf{u})} F_{I,J} F_{I,J}^{-1} = \mathbf{b}_U D^{(\mathbf{u})}.$$

This completes the proof. ■

The reader may wonder about the presence of the factor $1/f_{\mathbf{u}}$ in (9.5). Obviously, in any reasonable stochastic process, the probability $f_{\mathbf{u}}$ approaches zero as $|\mathbf{u}| \to \infty$. Hence, unless we divide by this quantity, we would get an inequality that is trivially true because both quantities individually approach zero. In contrast, (9.6) shows that the matrix $(1/f_{\mathbf{u}})D^{(\mathbf{u})}$ is both bounded and bounded away from zero for all $\mathbf{u} \in \mathcal{M}^*$.

Thus Lemma 9.4 serves to establish the behaviour of the matrix $D^{(\mathbf{u})}$ as $|\mathbf{u}| \to \infty$. Whatever be the vector $\mathbf{x} \in \mathbb{R}^r$, the vector $(1/f_{\mathbf{u}})D^{(\mathbf{u})}\mathbf{x}$ approaches $(1/f_{\mathbf{u}})\mathbf{p}_{\mathbf{u}}\mathbf{b}_U D^{(\mathbf{u})}\mathbf{x}$ and thus eventually gets 'aligned' with the vector $\mathbf{p}_{\mathbf{u}}$ as $|\mathbf{u}| \to \infty$.

Lemma 9.5 *Suppose the process under study is ultra-mixing, and that the cluster points of the probability vector set $\mathcal{A}$ are finite in number and belong to the interior of the cone $\mathcal{C}_c$. Then there exists a polyhedral cone $\mathcal{P}$ such that*

1. *$\mathcal{P}$ is invariant under each $D^{(u)}, \mathbf{u} \in \mathcal{M}$.*
2. *$\mathcal{C}_c \subseteq \mathcal{P} \subseteq \mathcal{C}_o$.*
3. *$\phi \in \mathcal{P}$.*
4. *$\theta^t \in \mathcal{P}^p$.*

Remark: In some sense this is the key lemma in the proof of the main theorem. It is noteworthy that the hypotheses do *not* include the assumption that the process under study is α-mixing.

Proof: First, note that, given any $\epsilon > 0$, there exists an $L = L(\epsilon)$ such that the following is true: For each $\mathbf{w} \in \mathcal{M}^*$ with $|\mathbf{w}| > L$, write $\mathbf{w} = \mathbf{uv}$ with $|\mathbf{u}| = L$. Then

$\| \mathbf{p_w} - \mathbf{p_u} \| \leq \epsilon$. To see this, given $\epsilon > 0$, choose L such that $\delta_L \leq \epsilon/m^k$. Then (9.4) implies that $\| \mathbf{p_u} - \mathbf{p_w} \| \leq \epsilon$. ■

By assumption, the set of probability vectors $\mathcal{A} := \{\mathbf{p}_u : \mathbf{u} \in \mathcal{M}^k\}$ has only finitely many cluster points. Let us denote them as $\mathbf{x}_1, \ldots, \mathbf{x}_n$. By assumption again, each of these vectors lies in the interior of $\mathcal{C}_o$. Hence there exists an $\epsilon > 0$ such that the sphere (in the ℓ_1-norm) centered at each $\mathbf{x}_i$ of radius 2ϵ is also contained in $\mathcal{C}_o$.

Next, note that there exists an integer L such that *every* vector $\mathbf{p_u}$ with $|\mathbf{u}| \geq L$ lies within a distance of ϵ (in the ℓ_1-norm) from at least one of the $\mathbf{x}_i$. In other words, there exists an integer L such that

$$\min_{1 \leq i \leq n} \| \mathbf{p_u} - \mathbf{x}_i \| \leq \epsilon, \ \forall \mathbf{u} \in \mathcal{M}^l \text{ with } l > L.$$

To see why this must be so, assume the contrary. Thus there exists a sequence $\mathbf{p}_{\mathbf{u}_j}$ such that $\| \mathbf{p}_{\mathbf{u}_j} - \mathbf{x}_i \| > \epsilon$ for all i, j. Now the sequence $\{\mathbf{p}_{\mathbf{u}_j}\}$ is bounded and therefore has a convergent subsequence. The limit of this convergent subsequence cannot be any of the $\mathbf{x}_i$ by the assumption that $\| \mathbf{p}_{\mathbf{u}_j} - \mathbf{x}_i \| > \epsilon$ for all i, j. This violates the earlier assumption that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are *all* the cluster points of the set $\mathcal{A}$.

Now choose a set $\mathbf{z}_1, \ldots, \mathbf{z}_r$ of basis vectors for $\mathbb{R}^r$ such that each $\mathbf{z}_j$ has unit norm. For instance, we can take $\mathbf{z}_j$ to be the unit vector with a 1 in position j and zeros elsewhere. With ϵ already defined above, define the unit vectors

$$\mathbf{y}^+_{i,j} := \frac{\mathbf{x}_i + 2\epsilon \mathbf{z}_j}{\| \mathbf{x}_i + 2\epsilon \mathbf{z}_j \|}, \ \mathbf{y}^-_{i,j} := \frac{\mathbf{x}_i - 2\epsilon \mathbf{z}_j}{\| \mathbf{x}_i - 2\epsilon \mathbf{z}_j \|}, \ 1 \leq i \leq n, \ 1 \leq j \leq s.$$

With this definition, it is clear that *every* vector in the ball of radius 2ϵ centered at each $\mathbf{x}_i$ can be written as a nonnegative combination of the set of vectors $\{\mathbf{y}^+_{i,j}, \mathbf{y}^-_{i,j}\}$.

Now define the cone

$$\mathcal{B} := \text{Cone}\{\mathbf{y}^+_{i,j}, \mathbf{y}^-_{i,j}\}.$$

We begin by observing that $\mathbf{p_u} \in \mathcal{B}$ whenever $|\mathbf{u}| \geq L$. This is because each such $\mathbf{p_u}$ lies within a distance of ϵ from one of the $\mathbf{x}_i$ whenever $|\mathbf{u}| \geq L$. In particular, $\mathbf{p_u} \in \mathcal{B}$ whenever $|\mathbf{u}| = L$. Moreover, by (4.1) and (4.2), every $\mathbf{p_v}$ with $|\mathbf{v}| < L$ is a nonnegative combination of $\mathbf{p_u}$ with $|\mathbf{u}| = L$. To see this, let $s := L - |\mathbf{v}|$, and note that

$$f_{\mathbf{v}} \mathbf{p_v} = F^{(\mathbf{v})}_{I,0} = \sum_{\mathbf{w} \in \mathcal{M}^s} F^{(\mathbf{vw})}_{I,0},$$

and each vector $F^{(\mathbf{vw})}_{I,0}$ belongs to $\mathcal{B}$. Hence $\mathbf{p_u} \in \mathcal{B}$ whenever $|\mathbf{u}| < L$. Combining all this shows that $\mathbf{p_u} \in \mathcal{B}$ for all $\mathbf{u} \in \mathcal{M}^*$. As a result, it follows that $\mathcal{C}_c \subseteq \mathcal{B}$.

While the cone $\mathcal{B}$ is polyhedral, it is not necessarily invariant under each $D^{(u)}$. For the purpose of constructing such an invariant cone, it is now shown that $\mathcal{B}$ is invariant under each $D^{(\mathbf{u})}$ whenever $|\mathbf{u}|$ is sufficiently long. By Lemma 9.4, it follows that for every vector $\mathbf{y}$, the vector $(1/f_{\mathbf{u}})D^{(\mathbf{u})}\mathbf{y}$ gets 'aligned' with $p_{\mathbf{u}}$ as $|\mathbf{u}|$ becomes large. Therefore it is possible to choose an integer s such that

$$\| \frac{\| \mathbf{p_u} \|}{\| D^{(\mathbf{u})}\mathbf{y} \|} D^{(\mathbf{u})}\mathbf{y} - \mathbf{p_u} \| \leq \epsilon \text{ whenever } |\mathbf{u}| \geq s,$$

whenever $\mathbf{y}$ equals one of the $2nr$ vectors $\mathbf{y}_{i,j}^{+}, \mathbf{y}_{i,j}^{-}$. Without loss of generality it may be assumed that $s \geq L$. In particular, the vectors $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{+}$ and $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{-}$, after normalization, are all within a distance of ϵ from $\mathbf{p_u}$, which in turn is within a distance of ϵ from some $\mathbf{x}_t$. By the triangle inequality, this implies that the normalized vectors corresponding to $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{+}$ and $D^{(\mathbf{u})}\mathbf{y}_{i,j}^{-}$ are all within a distance of 2ϵ from some $\mathbf{x}_t$, and hence belong to $\mathcal{B}$. In other words, we have shown that

$$D^{(\mathbf{u})}\mathcal{B} \subseteq \mathcal{B} \; \forall \mathbf{u} \text{ with } |\mathbf{u}| \geq s.$$

Now we are in a position to construct the desired polyhedral cone $\mathcal{P}$. Define

$$\mathcal{B}_i := \{D^{(\mathbf{u})}\mathcal{B} : |\mathbf{u}| = i\}, 1 \leq i \leq s-1.$$

Thus $\mathcal{B}_i$ is the set obtained by multiplying each vector in $\mathcal{B}$ by a matrix of the form $D^{(\mathbf{u})}$ where $\mathbf{u}$ has length precisely i. It is easy to see that, since $\mathcal{B}$ is polyhedral, so is each $\mathcal{B}_i$. Now define

$$\mathcal{P} := \text{Cone}\{\mathcal{B}, \mathcal{B}_1, \ldots, \mathcal{B}_{s-1}\}.$$

For this cone, we establish in turn each of the four claimed properties.

Property 1: By definition we have that $D^{(u)}\mathcal{B}_i \subseteq \mathcal{B}_{i+1} \; \forall u \in \mathcal{M}$, whenever $0 \leq i \leq s-2$, and we take $\mathcal{B}_0 = \mathcal{B}$. On the other hand, $D^{(u)}\mathcal{B}_{s-1} \subseteq \mathcal{B}$ as has already been shown. Hence $\mathcal{P}$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$.

Property 2: We have already seen that $\mathbf{p_u} \in \mathcal{B}$ for all $\mathbf{u} \in \mathcal{M}^*$. Hence $\mathcal{C}_c = \text{Cone}\{\mathbf{p_u} : \mathbf{u} \in \mathcal{M}^*\} \subseteq \mathcal{B} \subseteq \mathcal{P}$. To prove the other containment, note that by assumption, the sphere of radius 2ϵ centered at each cluster point $\mathbf{x}_i$ is contained in $\mathcal{C}_o$. Hence $\mathcal{B} \subseteq \mathcal{C}_o$. Moreover, $\mathcal{C}_o$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$. Hence $\mathcal{B}_i \subseteq \mathcal{C}_o$ for each $i \in \{1, \ldots, s-1\}$. Finally $\mathcal{P} \in \mathcal{C}_o$.

Property 3: Note that each $\mathbf{p_u}$ belongs to $\mathcal{B}$, which is in turn a subset of $\mathcal{P}$. In particular, $\phi = \mathbf{p}_\emptyset \in \mathcal{P}$.

Property 4: Since $\mathcal{P} \subseteq \mathcal{C}_o$, it follows that $(\mathcal{P})^p \supseteq (\mathcal{C}_o)^p$. Hence it is enough to show that $\theta^t \in (\mathcal{C}_o)^p$. But this is easy to establish. Let $\mathbf{y} \in \mathcal{C}_o$ be arbitrary. Then by the definition of $\mathcal{C}_o$ we have that

$$\theta D^{(\mathbf{u})}\mathbf{y} \geq 0 \; \forall \mathbf{u} \in \mathcal{M}^* \; \forall \mathbf{y} \in \mathcal{C}_o.$$

In particular, by taking $\mathbf{u}$ to be the empty string (leading to $D^{(\mathbf{u})} = I$), it follows that $\theta\mathbf{y} \geq 0 \; \forall \mathbf{y} \in \mathcal{C}_o$. Since $\mathbf{y}$ is arbitrary, this shows that $\theta^t \in (\mathcal{C}_o)^p$. ■

Proof of Theorem 9.2: The proof of the main theorem closely follows the material in [1], pp. 117-119. Let us 'recycle' the notation and let $\mathbf{y}_1, \ldots, \mathbf{y}_s$ denote generators of the polyhedral cone $\mathcal{P}$. In other words, $\mathcal{P}$ consists of all nonnegative combinations of the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_s$. Note that neither the integer s nor the generators need be uniquely defined, but this does not matter. Define the matrix

$$Y := [\mathbf{y}_1 | \ldots | \mathbf{y}_s] \in \mathbb{R}^{m^k \times s}.$$

Then it is easy to see that

$$\mathcal{P} = \{Y\mathbf{x} : \mathbf{x} \in \mathbb{R}^s_+\}.$$

Now we can reinterpret the four properties of Lemma 9.5 in terms of this matrix. Actually we need not bother about Property 2.

Property 1: Since $\mathcal{P}$ is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$, it follows that each $D^{(u)}\mathbf{y}_i$ is a nonnegative combination of $\mathbf{y}_1, \ldots, \mathbf{y}_s$. Hence there exist nonnegative matrices $G^{(u)} \in \mathbb{R}_+^{s \times m^k}, u \in \mathcal{M}$ such that

$$D^{(u)}Y = YG^{(u)}, \; \forall u \in \mathcal{M}.$$

Property 3: Since $\phi \in \mathcal{P}$, there exists a nonnegative vector $\mathbf{z} \in \mathbb{R}^s_+$ such that

$$\phi = Y\mathbf{z}.$$

Property 4: Since $\theta \in \mathcal{P}^p$, we have in particular that $\theta\mathbf{y}_i \geq 0$ for all i. Hence

$$\mathbf{h} := \theta Y \in \mathbb{R}^s_+.$$

Moreover, $\mathbf{h} \neq \mathbf{0}$, because $\theta\phi = \mathbf{hz} = 1$, the frequency of the empty string.

With these observations, we can rewrite the expression for the frequency of an arbitrary

string $\mathbf{u} \in \mathcal{M}^*$. We have

$$\begin{aligned} f_{\mathbf{u}} &= \theta D^{(u_1)} \cdots D^{(u_l)} \phi \\ &= \theta D^{(u_1)} \cdots D^{(u_l)} Y \mathbf{z} \\ &= \theta D^{(u_1)} \cdots D^{(u_{l-1})} Y G^{(u_l)} \mathbf{z} = \cdots \\ &= \theta Y G^{(u_1)} \cdots G^{(u_l)} \mathbf{z} \\ &= \mathbf{h} G^{(u_1)} \cdots G^{(u_l)} \mathbf{z} \end{aligned} \tag{9.7}$$

The formula (9.7) is similar in appearance to (7.1), but with one very important difference: *Every matrix and vector in (9.7) is nonnegative.* Therefore, in order to construct an irreducible HMM from the above formula, we need to ensure that the matrix $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is irreducible and row stochastic, that $\mathbf{h}$ satisfies $\mathbf{h} = \mathbf{h}Q$, and that $\mathbf{z} = \mathbf{e}_s$. This is achieved through a set of three reductions. Note that these reductions are the same as in [1], pp. 117-119.

Now for the first time we invoke the assumption that the process $\{\mathcal{Y}_t\}$ is α-mixing. From Theorem 8.1, this assumption implies that the matrix $S = \sum_{u \in \mathcal{M}} D^{(u)}$ has the 'strong Perron property,' namely: The spectral radius of S is one, and one is an eigenvalue of S; moreover, if λ is any eigenvalue of S besides one, then $|\lambda| < 1$. We also know that ϕ and θ are respectively a column eigenvector and a row eigenvector of S corresponding to the eigenvalue one.

Now let us return to the formula (9.7). Define $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ as before. Observe that Q is a nonnegative matrix; hence, by [7], Theorem 1.3.2, p. 6, it follows that the spectral radius $\rho(Q)$ is also an eigenvalue. Moreover, $\rho(Q)$ is at least equal to one, because

$$\mathbf{h}Q = \theta \sum_{u \in \mathcal{M}} Y G^{(u)} = \theta \left(\sum_{u \in \mathcal{M}} D^{(u)} \right) Y = \theta Y = \mathbf{h}.$$

Here we make use of the fact that θ is a row eigenvector of $\sum_{u \in \mathcal{M}} D^{(u)}$ corresponding to the eigenvalue one.

In what follows, we cycle through three steps in order to arrive at a situation where Q is irreducible and row stochastic. In each step we will be replacing the various matrices by other, smaller matrices that play the same role. To avoid notational clutter, the old and new matrices are denoted by the same symbols.

Step 1: If Q is irreducible, go to Step 3. If Q is reducible, permute rows and columns if necessary and partition Q as

$$Q = \begin{bmatrix} Q_{11} & Q_{12} \\ \mathbf{0} & Q_{22} \end{bmatrix},$$

where Q_{11} is irreducible and has dimension $(s-l)\times(s-l)$, and Q_{22} has dimension $l\times l$ for some $l<s$. (It is not assumed that Q_{22} is irreducible, since an irreducible partition of Q may have more than two 'blocks.') Since $Q=\sum_{u\in\mathcal{M}}G^{(u)}$ and each $G^{(u)}$ is nonnegative, if we partition each $G^{(u)}$ commensurately, then the block zero structure of Q will be reflected in each $G^{(u)}$. Now there are two possibilities: Either $\rho(Q_{11})=1$, or it is not. If $\rho(Q_{11})=1$, go to Step 2. If $\rho(Q_{11})\neq 1$, proceed as follows: Let $\lambda_1=\rho(Q_{11})\neq 1$. Choose a positive vector $\mathbf{x}_1\in\mathbb{R}_+^{s-l}$ such that $Q_{11}\mathbf{x}_1=\lambda_1\mathbf{x}_1$. (Note that, by [7], Theorem 2.2.10, p. 30, it is possible to choose a strictly positive eigenvector of Q_{11} corresponding to the eigenvalue $\rho(Q_{11})$, since Q_{11} is irreducible.) Then clearly $Q\mathbf{x}=\lambda_1\mathbf{x}$, where $\mathbf{x}=[\mathbf{x}_1^t\ \ \mathbf{0}^t]^t$. Since $\lambda_1\neq 1$, it follows that $\mathbf{hx}=0$. (Recall that a row eigenvector and a column eigenvector corresponding to different eigenvalues are orthogonal.) So if we partition $\mathbf{h}$ as $[\mathbf{h}_1\ \ \mathbf{h}_2]$, then $\mathbf{h}_1=\mathbf{0}$ since $\mathbf{x}_1$ is a positive vector. Now observe that each $G^{(u)}$ has the same block-triangular structure as Q. Hence, by a slight abuse of notation, let us define, for every string $\mathbf{u}\in\mathcal{M}^*$,

$$G^{(\mathbf{u})}=\begin{bmatrix} G_{11}^{(\mathbf{u})} & G_{12}^{(\mathbf{u})} \\ \mathbf{0} & G_{22}^{(\mathbf{u})} \end{bmatrix}.$$

Let us partition $\mathbf{z}$ commensurately. Because the first block of $\mathbf{h}$ is zero, it is easy to verify that, for every $\mathbf{u}\in\mathcal{M}^*$, we have

$$f_{\mathbf{u}}=\mathbf{h}G^{(\mathbf{u})}\mathbf{z}=\mathbf{h}_2G_{22}^{(\mathbf{u})}\mathbf{z}_2,$$

where $\mathbf{z}_2$ consists of the last l components of $\mathbf{z}$. Hence we can partition Y as $[Y_1|Y_2]$ where $Y_2\in\mathbb{R}^{r\times l}$ and make the following substitutions:

$$s\leftarrow l, Y\leftarrow Y_2, G^{(u)}\leftarrow G_{22}^{(u)}\ \forall u\in\mathcal{M}, \mathbf{h}\leftarrow\mathbf{h}_2, \mathbf{z}\leftarrow\mathbf{z}_2.$$

In this way, we have reduced the number of columns of Y from s to r, and (9.7) continues to hold. Now go back to Step 1.

Step 2: If we have reached this point, then Q is reducible, and if it is partitioned as above, we have $\rho(Q_{11})=1$. Choose a positive vector $\mathbf{x}_1$ such that $Q_{11}=\mathbf{x}_1$. Then $Q\mathbf{x}=\mathbf{x}$, where as before $\mathbf{x}=[\mathbf{x}_1^t\ \ \mathbf{0}^t]^t$. Next, note that

$$SY\mathbf{x}=\left(\sum_{u\in\mathcal{M}}D^{(u)}\right)Y\mathbf{x}=Y\left(\sum_{u\in\mathcal{M}}G^{(u)}\right)\mathbf{x}=YQ\mathbf{x}=Y\mathbf{x}.$$

Hence $Y\mathbf{x}$ is a column eigenvector of S corresponding to the eigenvalue one. However, from Theorem 8.1, the α-mixing property implies that S has a simple eigenvalue at one, with

corresponding column eigenvector $\phi = F_{I,0}$. Hence $F_{I,0}$ equals $Y\mathbf{x}$ times some scale factor, which can be taken as one without loss of generality (since both vectors are nonnegative). Partition Y as $[Y_1 \;\; Y_2]$ where $Y_1 \in \mathbb{R}^{r\times(s-l)}$. Then

$$F_{I,0} = [Y_1 \;\; Y_2] \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{0} \end{bmatrix} = Y_1\mathbf{x}_1.$$

Moreover, since each $G^{(u)}$ inherits the zero structure of Q, we have that

$$D^{(u)}[Y_1 \;\; Y_2] = [Y_1 \;\; Y_2] \begin{bmatrix} G_{11}^{(u)} & G_{12}^{(u)} \\ \mathbf{0} & G_{22}^{(u)} \end{bmatrix}.$$

In particular, we have that $D^{(u)}Y_1 = Y_1G_{11}^{(u)}$. This means that $F_{I,0}$ lies in the cone generated by the columns of Y_1, and that this cone is invariant under $D^{(u)}$ for each $u \in \mathcal{M}$. So if we define $\mathbf{h}_1 := \theta Y_1$, then because of the zero block in $\mathbf{x}$ it follows that

$$f_{\mathbf{u}} = \theta D^{(\mathbf{u})}F_{I,0} = \mathbf{h}_1G_{11}^{(\mathbf{u})}\mathbf{x}_1.$$

So now we can make the substitutions

$$s \leftarrow s-l, Y \leftarrow Y_1, G^{(u)} \leftarrow G_{11}^{(\mathbf{u})}, \mathbf{h} \leftarrow \mathbf{h}_1, \mathbf{z} \leftarrow \mathbf{x}_1.$$

With these substitutions we have the relationship (9.7) continues to hold. In the process, the number of columns of Y has been reduced from s to $s-l$. Moreover, the resulting matrix Q is the old Q_{11}, which is irreducible. Now go to Step 3.

Step 3: When we reach this stage, (9.7) continues to hold, but with two crucial additional features: Q is irreducible and $\rho(Q) = 1$. As before, let s denote the size of the matrix Q, and write $\mathbf{z} = [z_1 \dots z_s]^t$, where each z_i is positive. Define $Z = \text{Diag}\{z_1, \dots, z_s\}$. Now (9.7) can be rewritten as

$$f_{\mathbf{u}} = \mathbf{h}ZZ^{-1}G^{(u_1)}Z \cdot Z^{-1}G^{(u_2)}Z \dots Z^{-1}G^{(u_l)}Z \cdot Z^{-1}\mathbf{z}.$$

Thus (9.7) holds with the substitutions

$$G^{(u)} \leftarrow Z^{-1}G^{(u)}Z, \mathbf{h} \leftarrow \mathbf{h}Z, \mathbf{z} \leftarrow Z^{-1}\mathbf{z}.$$

In this process, Q gets replaced by $Z^{-1}QZ$. Now observe that

$$Z^{-1}QZ\mathbf{e}_s = Z^{-1}Q\mathbf{z} = Z^{-1}\mathbf{z} = \mathbf{e}_s.$$

In other words, the matrix $Z^{-1}QZ$ is row stochastic. It is obviously nonnegative and irreducible. Moreover, we have that $\mathbf{hz} = 1$ since it is the frequency of the empty string, which by definition equals one. Hence the row vector $\mathbf{h}Z^{-1}$ is row stochastic in that its entries add up to one. Hence, after we make the substitutions, (9.7) holds with the additional properties that (i) $Q := \sum_{u \in \mathcal{M}} G^{(u)}$ is row-stochastic, (ii) $\mathbf{h}$ is row-stochastic and satisfies $\mathbf{h} = \mathbf{h}Q$, and (iii) $\mathbf{z} = \mathbf{e}_s$. Now it follows from Lemma 9.1 that the process $\{\mathcal{Y}_t\}$ has a 'joint Markov process' HMM. Moreover, the matrix Q is irreducible.

Thus far it has been established that the stochastic process $\{\mathcal{Y}_t\}$ has an irreducible HMM. Moreover, this process is assumed to be α-mixing. So from Theorem 9.1, it finally follows that either the corresponding state transition matrix is aperiodic, or else the consistency conditions (9.3) hold. ■

Theorem 9.2 gives *sufficient* conditions for the existence of an irreducible HMM that satisfies some consistency conditions in addition. It is therefore natural to ask how close these sufficient conditions are to being necessary. The paper [1] also answers this question.

Theorem 9.3 *Given an irreducible HMM with n states and m outputs, define its period p. Rearrange the state transition matrix A as in Theorem 9.1, permute the matrices $M^{(u)}, u \in \mathcal{M}$ correspondingly, and define the blocks $M_i^{(u)}$ in analogy with the partition of A. Suppose in addition that there exists an index $q \leq s$ such that the following property holds: For every string $\mathbf{u} \in \mathcal{M}^q$ and every integer r between 1 and p, every column of the product $M_r^{(u_1)} M_{r+1}^{(u_2)} \dots M_{r+q-1}^{(u_q)}$ is either zero or else is strictly positive. In this computation, any subscript M_i is replaced by i mod p if $i > p$. With this property, the HMM is α-mixing and also ultra-mixing.*

For a proof, see [1], Lemma 2.

Thus we see that there is in fact a very small gap between the sufficiency condition presented in Theorem 9.2 and the necessary condition discovered earlier in [1]. If the sufficient conditions of Theorem 9.2 are satisfied, then there exists an irreducible HMM that also satisfies the consistency conditions (9.3). Conversely, if an irreducible HMM satisfies the consistency conditions (9.3) and one other technical condition, then it satisfies three out of the four hypotheses of Theorem 9.2, the only exception being the technical condition about the cluster points lying in the interior of the cone $\mathcal{C}_c$.

We conclude this section by discussing the nature of the 'technical' conditions in the hypotheses of Theorems 9.2 and 9.3. The idea is to show that, in a suitably defined topology, each of the conditions is satisfied by an 'open dense subset' of stochastic processes. Thus,

if the given process satisfies the condition, so does any sufficiently small perturbation of it, whereas if a given process fails to satisfy the condition, an arbitrarily small perturbation will cause the condition to hold.

Let us begin with the fourth hypothesis of Theorem 9.1. We follow [33] and define a topology on the set of all stationary stochastic processes assuming values in $\mathcal{M}$. Suppose we are given two stochastic processes assuming values in a common finite alphabet $\mathcal{M}$. Let $f_{\mathbf{u}}, g_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*$ denote the frequency vectors of the two stochastic processes. This is equivalent to specifying the joint distribution of l-tuples of each stochastic process, for every integer l. If we arrange all strings $\mathbf{u} \in \mathcal{M}^*$ in some appropriate lexical ordering (say first lexical), then each of $[f_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*], [g_{\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*]$ is a vector with a countable number of components, and each component lies between 0 and 1.[10] Let the symbols $\mathbf{f}, \mathbf{g}$, without any subscript, denote these vectors belonging to ℓ_∞. We might be tempted to compare the two stochastic processes by computing the norm $\| \mathbf{f} - \mathbf{g} \|_\infty$. The difficulty with this approach is that, as the length of the string $\mathbf{u}$ approaches infinity, the likelihood of that sequence will in general approach zero. Thus, in any 'reasonable' stochastic process, the difference $f_{\mathbf{u}} - g_{\mathbf{u}}$ will approach zero as $|\mathbf{u}| \to \infty$, but this tells us nothing about how close the two probability laws are. To get around this difficulty, for each $\mathbf{u} \in \mathcal{M}^*$, we define the vector $\mathbf{p}_{|\mathbf{u}} \in [0,1]^m$ as follows:

$$\mathbf{p}_{|\mathbf{u}} = \frac{1}{f_{\mathbf{u}}} \mathbf{f}_{\mathbf{u}v, v \in \mathcal{M}} = \left[\frac{f_{\mathbf{u}v}}{f_{\mathbf{u}}}, v \in \mathcal{M} \right].$$

Thus $\mathbf{p}_{|\mathbf{u}}$ is just the conditional distribution of the *next* symbol, given the past history $\mathbf{u}$. The advantage of $\mathbf{p}_{|\mathbf{u}}$ is that, even as $|\mathbf{u}|$ becomes large, the elements of this vector must still add up to one, and as a result they cannot all go to zero. With this convention, let us list all strings $\mathbf{u} \in \mathcal{M}^*$ in some appropriate lexical ordering (say first lexical), and for each $\mathbf{u}$ let us define the conditional distribution vectors $\mathbf{p}_{|\mathbf{u}}$ corresponding to $\{f_{\mathbf{u}}\}$, and the conditional distribution vectors $\mathbf{q}_{|\mathbf{u}}$ corresponding to the vector $\{\mathbf{g}_{\mathbf{u}}\}$. Finally, let us define the vectors

$$\tilde{\mathbf{p}} := [\mathbf{p}_{|\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*], \tilde{\mathbf{q}} := [\mathbf{q}_{|\mathbf{u}}, \mathbf{u} \in \mathcal{M}^*].$$

Thus both $\tilde{\mathbf{p}}, \tilde{\mathbf{q}}$ have a countable number of components, since $\mathcal{M}^*$ is a countable set. Thus the ℓ_∞ norm of the difference $\tilde{\mathbf{p}} - \tilde{\mathbf{q}}$ is a measure of the disparity between the two stochastic processes. This is essentially the distance measure introduced in [33]. With this measure, it is easy to see that the fourth hypothesis of Theorem 9.1 is truly technical: If a given stochastic

[10] Note that there is a lot of redundancy in this description of a stochastic process because, as we have already seen, the joint distribution of l-tuples can be uniquely determined from the joint distribution of s-tuples if $s > l$.

process satisfies the condition about the cluster points, then so will any sufficiently small perturbation of it, while if a given stochastic process fails to satisfy this condition, any sufficiently small perturbation of it will cause the condition to be satisfied.

Now let us turn to the condition in Theorem 9.3. Given two HMMs over a common state space, a natural metric is

$$\sum_{u \in \mathcal{M}} \| M_1^{(u)} - M_2^{(u)} \|,$$

where $\| \cdot \|$ is any reasonable matrix norm. Again, it is easy to see that the condition in Theorem 9.3 about the various columns being either identically zero or strictly positive is 'technical.' In fact, if for a HMM some elements of the matrices $M_r^{(u_1)} M_{r+1}^{(u_2)} \ldots M_{r+q-1}^{(u_q)}$ are zero, then by simply making an arbitrarily small perturbation in the matrices we can ensure that every entry is strictly positive.

10 Conclusions and Future Work

In this paper, we have reviewed a considerable body of literature pertaining to the complete realization problem for hidden Markov models. In addition, we have also presented some significant new results. In particular, a new notion called "ultra-mixing" has been introduced. It has been shown that if a finite Hankel rank process is both α-mixing and ultra-mixing, and if an additional technical condition is satisfied, then the process has an irreducible HMM and satisfies a consistency condition. There is a near converse: If a finite Hankel rank process has an irreducible HMM and satisfies a consistency condition, and also satisfies another technical condition, then the process is both α-mixing as well as ultra-mixing. By introducing suitable topologies on the set of all stochastic processes, and the set of all HMMs, it has been established that each of these additional conditions is truly 'technical,' in the sense that the condition holds on an open dense set of processes/HMMs.

Much work remains to be done. All of the work here is based on the assumption that the *complete statistics* of the process under study are known. In a practical application, such as in speech recognition or in computational biology, a much more likely scenario is that one is given a finite length sample path of the process under study, and is expected to construct a model for it. Since only a finite length sample path is available, one can only approximate the true probabilities of various strings via their frequencies of occurence. Hence it does not make sense to insist that a model should match these observed frequencies exactly. It would be much more sensible to match the observed frequencies only approximately, and use the

freedom so afforded to reduce the size of the state space. However, the theory for doing this is not as yet available. It should be noted that in the literature (see e.g., [3, 27]), one begins with a HMM *of known fixed order*, and tries to find the best estimate of the parameters of the HMM to fit the observed data. The much more natural approach of *choosing the model order based on the data* has not attracted much attention.

Even in the case where the complete statistics are known, there are some interesting problems that are still open. For instance, suppose one is given a stochastic process that satisfies all the hypotheses of Theorem 9.2. What is the *minimum number* of states needed to realize this process? Or, suppose one is given a HMM where the underlying state transition matrix is reducible. Is it possible to replace this HMM by another equivalent HMM (meaning that the output frequencies are preserved), where the new HMM has an irreducible state transition matrix? These are problems for future study.

Acknowledgements

The author thanks Prof. Probal Chaudhuri of the Indian Statistical Institute, Kolkata, and Prof. Rajeeva Karandikar, formerly of the Indian Statistical Institute, Delhi, and currently with Cranes Software, Bangalore, for several helpful discussions. He thanks Prof. Vincent Blondel for the references on undecidability issues. He thanks Prof. Isaac Meilijson of Tel Aviv University for drawing his attention to the paper [28]. Finally, he thanks the reviewers of the paper for their careful reading for several useful suggestions.

References

[1] B. D. O. Anderson, "The realization problem for hidden Markov models," *Mathematics of Control, Signals, and Systems*, 12(1), 80-120, 1999.

[2] B. D. O. Anderson, M. Deistler, L. Farina and L. Benvenuti, "Nonnegative realization of a system with a nonnegative impulse response," *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications*, 43, 134-142, 1996.

[3] P. Baldi and S. Brunak, *Bioinformatics: A Machine Learning Approach*, (Second Edition), MIT Press, Cambridge, MA, 2001.

[4] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, 37, 1554-1563, 1966.

[5] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains," *Annals of Mathematical Statistics*, 41(1), 164-171, 1970.

[6] L. Benvenuti and L. Farina, "A tutorial on the positive realization problem," *IEEE Transactions on Automatic Control*, 49, 651-664, 2004.

[7] A. Berman and R. J. Plemmons, *Nonnegative Matrices*, Academic Press, New York, 1979.

[8] D. Blackwell and L. Koopmans, "On the identifiability problem for functions of finite Markov chains," *Annals of Mathematical Statistics*, 28, 1011-1015, 1957.

[9] V. Blondel and V. Catarini, "Undecidable problems for probabilistic automata of fixed dimension," *Theory of Computing Systems*, **36**, 231-245, 2003.

[10] J. W. Carlyle, "Identification of state-calculable functions of finite Markov chains," *Annals of Mathematical Statistics*, 38, 201-205, 1967.

[11] J. W. Carlyle, "Stochastic finite-state system theory," in *System Theory*, L. Zadeh and E. Polak (Eds.), Chapter 10, McGraw-Hill, New York, 1969.

[12] S. E. Cawley, A. L. Wirth and T. P. Speed, "Phat – a gene finding program for *Plasmodium falciparum*," *Molecular & Biochemical Parasitology*, 118, 167-174, 2001.

[13] A. L. Delcher, D. Harmon, S. Kasif, O. White and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Research*, 27(23), 4636-4641, 1999.

[14] S. W. Dharmadhikari, "Functions of finite Markov chains," *Annals of Mathematical Statistics*, 34, 1022-1031, 1963.

[15] S. W. Dharmadhikari, "Sufficient conditions for a stationary process to be a function of a Markov chain," *Annals of Mathematical Statistics*, 34, 1033-1041, 1963.

[16] S. W. Dharmadhikari, "A characterization of a class of functions of finite Markov chains," *Annals of Mathematical Statistics*, 36, 524-528, 1965.

[17] S. W. Dharmadhikari, "A note on exchangeable processes with states of finite rank," *Annals of Mathematical Statistics*, 40(6), 2207-2208, 1969.

[18] S. W. Dharmadhikari and M. G. Nadkarni, "Some regular and non-regular functions of finite Markov chains," *Annals of Mathematical Statistics*, 41(1), 207-213, 1970.

[19] R. V. Erickson, "Functions of Markov chains," *Annals of Mathematical Statistics*, 41, 843-850, 1970.

[20] M. Fliess, "Series rationelles positives et processus stochastique," *Annales de l'Institut Henri Poincaré, Section B*, XI, 1-21, 1975.

[21] M. Fox and H. Rubin, "Functions of processes with Markovian states," *Annals of Mathematical Statistics*, 39, 938-946, 1968.

[22] E. J. Gilbert, "The identifiability problem for functions of Markov chains," *Annals of Mathematical Statistics*, 30, 688-697, 1959.

[23] A. Heller, "On stochastic processes derived from Markov chains," *Annals of Mathematics*, 36, 1286-1291, 1965.

[24] J. M. van den Hof, "Realization of continuous-time positive linear systems," *Systems and Control Letters*, 31, 243-253, 1997.

[25] J. M. van den Hof and J. H. van Schuppen, "Realization of positive linear systems using polyhedral cones," *Proceedings of the 33rd IEEE Conference on Decision and Control*, 3889-3893, 1994.

[26] H. Ito, S. Amari and K. Kobayashi, "Identifiability of hidden Markov information sources and their minimum degrees of freedom," *IEEE Transactions on Information Theory*, 38, 324-333, 1992.

[27] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, 1997.

[28] S. Kalikow, "Random Markov processes and uniform martingales," *Israel Journal of Mathematics*, 71(1), 33-54, 1990.

[29] A. Krogh, M. Brown, I. S. Mian, K. Sjölander and D. Haussler, "Hidden Markov models in computational biology: Applications to protein modeling," *J. Mol. Biol.*, 235, 1501-1531, 1994.

[30] A. Krogh, I. S. Mian and D. Haussler, "A hidden Markov model that finds genes in *E. coli* DNA," *Nucleic Acids Research*, 22(22), 4768-4778, 1994.

[31] L. Kronecker "Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen," *Monatsber. Königl. Preuss. Akad. Wiss. Berlin*, 535-600, 1881.

[32] W. H. Majoros and S. L. Salzberg, "An empirical analysis of training protocols for probabilistic gene finders," *BMC Bioinformatics*, available at `http://www.biomedcentral.com/1471-2105/5/206`, 21 December 2004.

[33] D. S. Ornstein and B. Weiss, "How sampling reveals a process," *Ann. Probab.*, 18(3), 905-930, 1990.

[34] G. Picci, "On the internal structure of finite-state stochastic processes," in *Recent Developments in Variable Structure Systems*, R. Mohler and A. Ruberti (Eds.), Lecture Notes in Economics and Mathematical Systems, No. 162, Springer-Verlag, Heidelberg, 1978.

[35] L. W. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, 77(2), 257-285, February 1989.

[36] G. Rozenberg and A. Salomaa, *Cornerstones in Undecidability*, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[37] S. L. Salzberg, A. L. Delcher, S. Kasif and O. White, "Microbial gene identification using interpolated Markov models," *Nucleic Acids Research*, 26(2), 544-548, 1998.

[38] E. Seneta, *Non-negative Matrices and Markov Chains*, (Second Edition), Springer-Verlag, New York, 1981.

[39] E. D. Sontag, "On certain questions of rationality and decidability," *Journal of Computer and System Science*, 11, 375-381, 1975.

[40] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*, Springer-Verlag, London, 2003.

[41] M. Vidyasagar, *Nonlinear Systems Analysis*, SIAM Publications, Philadelphia, PA, 2003.