Bounds on Convergence of Entropy Rate Approximations in Hidden Markov Processes

By

Nicholas F. Travers
B.S. Haverford College 2005

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA,

DAVIS

Approved:

Professor James Crutchfield (Chair)

Professor Benjamin Morris

Professor Janko Gravner

Committee in Charge

2013

Abstract

There is no general closed form expression for the entropy rate of a hidden Markov process. However, the finite length block estimates $h(t)$ often converge to the true entropy rate h quite rapidly. We establish exponential bounds on the rate of convergence of the block estimates for finite hidden Markov models under several different conditions, including exactness, unifilarity, and a flag-state condition. In the case of unifilar hidden Markov models, we also give exponential bounds on the L_1 and a.s. decay of the state uncertainty U_t.

Acknowledgements

I thank my advisor Jim Crutchfield for all his advice and guidance throughout my graduate career. I thank Ben Morris for several helpful discussions and, in particular, the idea for the follower model construction used for unifilar hidden Markov models. I thank Janko Gravner for his general support and also for his comments and suggestions for publication on papers containing earlier versions of the results given in this dissertation.

I thank my undergraduate real analysis professor Rob Manning who first introduced me to "higher mathematics" and is much of the reason that I decided, eventually, to come back and pursue my PhD. I thank Paul Klingsberg for helping to inspire my interest in mathematics from a young age and for the countless hours he spent working with me (and others).

Lastly, I thank my parents and my many friends at Davis, both in the graduate program and outside, whose support has helped me through difficult times and kept me going.

Contents

1 Introduction

Hidden Markov models (HMMs) are generalizations of Markov chains in which the underlying Markov state sequence (S_t) is observed through a noisy or lossy channel, leading to a (typically) non-Markovian output process (X_t). They were first introduced in the 50s as abstract mathematical models [1–3], but have since been applied quite successfully in a number of contexts for modeling, such as speech recognition [4–7] and bioinformatics [8–12].

One of the earliest major questions in the study of HMMs [3] was to determine the entropy rate of the output process:

$$\begin{aligned} h &= \lim_{t\to\infty} H(X_t|X_1, ..., X_{t-1}) \\ &= \lim_{t\to\infty} H(X_0|X_{-(t-1)}, ..., X_{-1}) \end{aligned}$$

Unlike for Markov chains, this actually turns out to be quite difficult for HMMs. Even in the finite case no general closed form expression is known, and it is widely believed that no such formula exists. A nice integral expression was provided in [3], but it is with respect to an invariant density that is not directly computable.

In practice, the entropy rate h is instead often approximated simply by the finite length block estimates:

$$\begin{aligned} h(t) &= H(X_t|X_1, ..., X_{t-1}) \\ &= H(X_0|X_{-(t-1)}, ..., X_{-1}) \end{aligned}$$

Thus, it is important to know about the rate of convergence of these estimates to ensure the quality of the approximation.

Moreover, even in cases where the entropy rate can be calculated exactly, such as for unifilar HMMs, the rate of convergence of the block estimates is still important for estimating other quantities such as the excess entropy [13], and for making finite data predictions. The entropy rate h is the observer's average uncertainty in prediction of the next output symbol X_0 after seeing the entire infinite past $X_{-\infty}^{-1} = ..., X_{-2}, X_{-1}$. However, for an observer wishing to make predictions from a finite observation sequence $X_{-t}^{-1} = X_{-t}, ..., X_{-1}$ the speed at which the block estimates $h(t)$ approach the entropy rate is also critical.

In the case of finite HMMs, where the internal state set $\mathcal{S}$ and output alphabet $\mathcal{X}$ are both finite, the block estimates $h(t)$ generally converge quite quickly, often exponentially. For instance, in [14] and [15] exponential bounds for the convergence rate are given for state-emitting HMMs with strictly positive transition probabilities in the underlying Markov chain. Also, in [16] exponential convergence is established using results from [17] (for both state-emitting and edge-emitting HMMs), with the assumption of strictly positive symbol emission probabilities and an aperiodic underlying Markov chain. However, without any positivity assumptions things become somewhat more complicated, and the proof methods used in these works do not carry through directly. Somewhat weaker conditions are assumed in (parts of) [18], but they still require the output process (X_t) to have full support, and, in general, we are not aware of any result that the block estimates $h(t)$ must converge exponentially for all finite HMMs.

The main objective of this dissertation is to establish exponential bounds on the convergence rate of the block estimates $h(t)$ for finite HMMs under alternative and weaker assumptions. In particular, we will consider three distinct conditions: exactness, unifiliarity, and a flag-state condition. Formal definitions will be given in Section 2 after introducing the requisite notation, but we summarize the concepts below:

- A HMM M is *exactly synchronizeable* or simply *exact* if it has some finite *synchronizing word* w such that an observer knows the current state of the HMM with certainty after seeing the output w.

- A HMM M is *unifilar* if the current state and next output symbol completely determine the next state.

- A HMM M is a *flag-state HMM* if each state i has some *flag word* w such that all states that can generate w may transition to i upon emitting w.

None of these conditions require positivity of the underlying state transition matrix or symbol emission probabilities, nor do they require the output process (X_t) to have full support. Moreover, the flag-state condition, when translated from the edge-emitting HMMs we focus on to state-emitting HMMs as in Appendix C, is strictly weaker than either (a) positivity of the state transition matrix or (b) aperiodicity of the Markov chain and positive emission probabilities. In fact, it is quite a weak condition, which we believe should in some sense "usually" hold, as will be described in Appendix B. The other two conditions are substantially stronger, but are relatively natural assumptions to make. Exactness or similar properties have been considered by several other authors as in [3, 16, 19, 20] and unifilarity is one of the key defining features of the ϵ-machine representation of a process [21, 22], which is a natural minimal predictive model.

The organization of the remainder of the dissertation is as follows:

- In Section 2 we establish notation and give formal definitions for our objects of study.
- In Section 3 we give some general information theoretic inequalities and estimates for Markov chains, which will be necessary for our proofs in later sections. We also introduce the block model M^n for a HMM M, which will be a key tool in the analysis in Sections 5 and 6.
- In Section 4 we establish results for exact HMMs. In particular, we present a construction known as the possibility machine that allows us to compute:

  $$\mathbb{P}(NSYN_t) = \mathbb{P}(X_0^{t-1} \text{ does not contain a sync word})$$

  and its rate of decay. We then prove a simple inequality, which shows that the entropy rate estimates $h(t)$ converge to h at least as fast as $\mathbb{P}(NSYN_t)$ converges to 0, thus establishing an upper bound on the rate of convergence for the estimates $h(t)$.
- In Section 5 we establish results for unifilar HMMs, including exponential convergence of the entropy rate block estimates $h(t)$, as well as exponential convergence of the state uncertainty U_t under the additional assumption of probabilistically distinguishable states. The proof methods are based on a construction known as the follower model.
- In Section 6 we establish bounds on convergence of the entropy rate estimates $h(t)$ for flag-state HMMs. The proof methods are based on the coupling techniques used in [14], but are somewhat more involved because of our weaker assumption.
- Finally, in Section 7 we summarize our results and give some concluding remarks.

2 Definitions and Notation

In this section we introduce the formal framework for our results. Definitions are provided for our various objects of study, and notation for the remainder of the work is established.

2.1 General

The set of real numbers is denoted by $\mathbb{R}$, the set of integers by $\mathbb{Z}$, the set of nonnegative integers by $\mathbb{Z}^+$, and the set of positive integers, or natural numbers, by $\mathbb{N}$. The cardinality of a set A is denoted $|A|$, and the length of a word (string) w is denoted in a similar fashion by $|w|$. $\mathcal{X}^*$ denotes the set of all finite words over an alphabet $\mathcal{X}$, including the empty word λ. For a sequence (a_n) and integers $n \leq m$, a_n^m denotes the finite subsequence $a_n, a_{n+1}, ..., a_m$. This notation is also extended in the natural way to the case $n = -\infty$ or $m = \infty$. If $n > m$, a_n^m denotes the empty string. For a finite set A and $i \in A$, $e_i = (0, ..., 1, ...0)$ denotes the standard basis vector in $\mathbb{R}^{|A|}$ with a 1 in the position of index i and 0s elsewhere. Frequently, the vector e_i will be interpreted as a probability measure on the set A, which places unit mass on the point i. In general, we will often not distinguish explicitly between probability vectors[1] and probability measures on finite sets.

2.2 Probability and Stochastic Processes

Generally, we will denote random variables by upper case English letters and their realizations with the corresponding lower case letters. The term *random variable* is used in a slightly generalized sense, in that not all random variables we consider will be real-valued. However, random variables that are not real-valued will usually be discrete, so they could be considered integer-valued by a relabeling of symbols.

If X is a random variable defined on a probability space with measure $\mathbb{P}$ then its distribution will be denoted by $\mathbb{P}(X)$. Similarly, if X and Y are two random variables on a common space with measure $\mathbb{P}$ and $\mathbb{P}(Y = y) > 0$, then $\mathbb{P}(X|Y = y)$ denotes the conditional distribution of X on the event $Y = y$. When the random variable Y is clear from context this may sometimes be abbreviated simply by $\mathbb{P}(X|y)$. Analogously, when clear from context, we may abbreviate $\mathbb{P}(X = x)$ as $\mathbb{P}(x)$ and $\mathbb{P}(X = x|Y = y)$ as $\mathbb{P}(x|y)$, for discrete random variables X, Y. By convention, we take $\mathbb{P}(x|y) = 0$ for all x, if $\mathbb{P}(y) = 0$.

[1] By a *probability vector* we mean a vector with nonnegative entries that sum to 1.

By a *stochastic process*, or simply a *process*, we will always mean a discrete time process, $\mathcal{P} = (X_t)_{t \in \mathbb{Z}}$ or $\mathcal{P} = (X_t)_{t \in \mathbb{Z}^+}$, over a finite alphabet $\mathcal{X}$. In the former case we refer to the process $\mathcal{P}$ as *two-sided*, and in the latter as *one-sided.* Frequently, we will consider the case of multiple processes (X_t) over a given alphabet $\mathcal{X}$, and in this case we associate a process $\mathcal{P}$ with its law $\mathbb{P}$.

2.2.1 Word Probabilities and the Process Language

If $\mathcal{P} = (X_t)$ is a stationary process with law $\mathbb{P}$ then we denote the stationary word probability for a word $w \in \mathcal{X}^*$ by $\mathbb{P}(w)$:

$$\mathbb{P}(w) \equiv \mathbb{P}(X_0^{|w|-1} = w) = \mathbb{P}(X_t^{t+|w|-1} = w)$$

and for a set of words A, all of some fixed length n, we define $\mathbb{P}(A)$ as the stationary probability of the set A:

$$\mathbb{P}(A) \equiv \sum_{w \in A} \mathbb{P}(w) = \mathbb{P}(X_0^{|w|-1} \in A)$$

Also, we denote the set of allowable words or *process language* by $\mathcal{L}(\mathcal{P})$ and the set of length-n allowable words by $\mathcal{L}_n(\mathcal{P})$:

$$\mathcal{L}(\mathcal{P}) \equiv \{w \in \mathcal{X}^* : \mathbb{P}(w) > 0\}$$
$$\mathcal{L}_n(\mathcal{P}) \equiv \{w \in \mathcal{L}(\mathcal{P}) : |w| = n\}$$

$\mathcal{L}_\infty(\mathcal{P})$, $\mathcal{L}_\infty^+(\mathcal{P})$, and $\mathcal{L}_\infty^-(\mathcal{P})$ denote, respectively, the set of allowable biinfinite sequences, allowable infinite futures, and allowable infinite pasts:

$$\mathcal{L}_\infty(\mathcal{P}) \equiv \{x_{-\infty}^{\infty} : x_n^m \in \mathcal{L}(\mathcal{P}), \text{ for each } n \leq m\}$$
$$\mathcal{L}_\infty^+(\mathcal{P}) \equiv \{x_0^{\infty} : x_n^m \in \mathcal{L}(\mathcal{P}), \text{ for each } 0 \leq n \leq m\}$$
$$\mathcal{L}_\infty^-(\mathcal{P}) \equiv \{x_{-\infty}^{-1} : x_n^m \in \mathcal{L}(\mathcal{P}), \text{ for each } n \leq m \leq -1\}$$

If the process $\mathcal{P}$ is clear from context, we may sometimes abbreviate $\mathcal{L}(\mathcal{P})$ and $\mathcal{L}_n(\mathcal{P})$ simply as $\mathcal{L}$ and $\mathcal{L}_n$.

2.2.2 Conditional Word Probabilities

If $\mathcal{P} = (X_t)$ is a stationary process over an alphabet $\mathcal{X}$ and $w, v \in \mathcal{X}^*$ are two words with $\mathbb{P}(v) > 0$, then $\mathbb{P}(w|v)$ denotes the conditional probability that the word v is followed by the word w:

$$\begin{aligned}\mathbb{P}(w|v) &\equiv \mathbb{P}(X_0^{|w|-1} = w | X_{-|v|}^{-1} = v) \\ &= \mathbb{P}(X_{-v}^{|w|-1} = vw)/\mathbb{P}(X_{-|v|}^{-1} = v) \\ &= \mathbb{P}(vw)/\mathbb{P}(v)\end{aligned}$$

For a word v with $\mathbb{P}(v) = 0$, we also define by convention:

$$\mathbb{P}(w|v) \equiv 0 \ , \quad \text{for all } w$$

Conditioning on infinite pasts $x_{-\infty}^{-1}$ instead of finite pasts $v = x_{-t}^{-1}$ is slightly more subtle because the probability of each infinite past $x_{-\infty}^{-1}$ is normally zero, but it may be accomplished by taking limits of finite length conditionals. Formally, we define for each past $x_{-\infty}^{-1} \in \mathcal{X}^{-\mathbb{N}}$ and word $w \in \mathcal{X}^*$:

$$\mathbb{P}(w|x_{-\infty}^{-1}) \equiv \begin{cases} \lim_{t\to\infty} \mathbb{P}(w|x_{-t}^{-1}) \ , \ \text{if } x_{-\infty}^{-1} \text{ is regular} \\ 0 \ , \ \text{else (by convention)} \end{cases} \tag{1}$$

where a past $x_{-\infty}^{-1}$ is said to be *regular* if:

1. $x_{-\infty}^{-1} \in \mathcal{L}_{\infty}^{-}(\mathcal{P})$, and
2. $\lim_{t\to\infty} \mathbb{P}(w|x_{-t}^{-1})$ exists, for each $w \in \mathcal{X}^*$.

It follows from the martingale convergence theorem that a.e. past $x_{-\infty}^{-1}$ is regular, so the definition (1) is meaningful in the limiting sense for a.e. past.

Remark. *It can also be shown that, for any word w, the function $\mathbb{P}(w|\cdot) : \mathcal{X}^{-\mathbb{N}} \to \mathbb{R}$ defines a version of the formal conditional probability $\mathbb{P}(X_0^{|w|-1} = w|X_{-\infty}^{-1})$. That is, for any subset A of infinite pasts $x_{-\infty}^{-1}$, which is measurable with respect to the σ-algebra generated by finite cylinder* *sets, we have:*

$$\mathbb{P}\left(X_0^{|w|-1} = w \ , \ X_{-\infty}^{-1} \in A\right) = \int_A \mathbb{P}(w|x_{-\infty}^{-1}) d\mathbb{P}(x_{-\infty}^{-1})$$

However, this fact will not be necessary for later developments.

2.3 Information Theory

We review below some basic definitions from discrete information theory, which will be necessary for formulating our results. For a more comprehensive overview the reader is referred to [23].

Throughout this section, and the remainder of the dissertation, we adopt the following standard information theoretic conventions for logarithms (of any base):

$$\begin{aligned} 0 \cdot \log(0) &\equiv \lim_{\xi \to 0^+} \xi \cdot \log(\xi) = 0 \\ 0 \cdot \log(1/0) &\equiv \lim_{\xi \to 0^+} \xi \cdot \log(1/\xi) = 0 \\ \log(1/0) &\equiv \lim_{\xi \to 0^+} \log(1/\xi) = \infty \end{aligned}$$

Note that with these conventions the functions $\xi \log(\xi)$ and $\xi \log(1/\xi)$ are both continuous on $[0,1]$.

2.3.1 Entropy and Related Measures

Definition 1. *The* entropy $H(\mathbb{P})$ *of a probability measure* $\mathbb{P}$ *on a finite set* $\mathcal{X}$ *is:*

$$H(\mathbb{P}) \equiv -\sum_{x \in \mathcal{X}} \mathbb{P}(x) \log_2 \mathbb{P}(x)$$

The entropy of a discrete random variable X *distributed according to the measure* $\mathbb{P}$ *is:*

$$H(X) \equiv H(\mathbb{P})$$

Definition 2. *Let* X, Y *be discrete random variables taking values in finite alphabets* $\mathcal{X}, \mathcal{Y}$*, and let* $\mathbb{P}(X,Y)$ *be their joint distribution. Then the* joint entropy $H(X,Y)$ *is:*

$$H(X,Y) \equiv -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} \mathbb{P}(x,y) \log_2 \mathbb{P}(x,y)$$

and the conditional entropy $H(X|Y)$ *is:*

$$\begin{aligned} H(X|Y) &\equiv \sum_{y \in \mathcal{Y}} \mathbb{P}(y) \cdot H(X|Y=y) \\ &= -\sum_{y \in \mathcal{Y}} \mathbb{P}(y) \cdot \sum_{x \in \mathcal{X}} \mathbb{P}(x|y) \log_2 \mathbb{P}(x|y) \end{aligned}$$

Similarly, for discrete random variables $X_1, ..., X_n$ *with joint distribution* $\mathbb{P}(X_1, ..., X_n) = \mathbb{P}(X_1^n)$ *the* joint entropy $H(X_1, ..., X_n) = H(X_1^n)$ *and conditional entropy* $H(X_n|X_1, ..., X_{n-1}) = H(X_n|X_1^{n-1})$ *are defined as:*

$$\begin{aligned} H(X_1^n) &\equiv -\sum_{x_1^n} \mathbb{P}(x_1^n) \log_2 \mathbb{P}(x_1^n) \ , \ \textit{and} \\ H(X_n|X_1^{n-1}) &\equiv -\sum_{x_1^{n-1}} \mathbb{P}(x_1^{n-1}) \cdot H(X_n|X_1^{n-1} = x_1^{n-1}) \ . \end{aligned}$$

Definition 3. *The* relative entropy *or* Kullback-Leibler divergence $D(\mathbb{P}||\mathbb{P}')$ *between two probability measures* $\mathbb{P}$ *and* $\mathbb{P}'$ *on a finite set* $\mathcal{X}$ *is defined as:*

$$D(\mathbb{P}||\mathbb{P}') \equiv \sum_{x \in \mathcal{X}} \mathbb{P}(x) \log_2 \left(\frac{\mathbb{P}(x)}{\mathbb{P}'(x)} \right)$$

Intuitively, the entropy $H(X)$ may be thought of as the uncertainty in predicting X or, equivalently, the amount of information obtained by observing X. The joint entropy $H(X,Y)$ is the uncertainty in predicting the pair (X,Y), and the conditional entropy $H(X|Y)$ is the (average) uncertainty in predicting X given knowledge of Y. Similarly, the joint entropy $H(X_1^n)$ is the uncertainty in predicting the n-tuple of random variables $(X_1, ..., X_n)$, and the conditional entropy $H(X_n|X_1^{n-1})$ is the (average) uncertainty in predicting X_n given knowledge of $X_1, ..., X_{n-1}$. The Kullback-Leibler divergence $D(\mathbb{P}||\mathbb{P}')$ may be thought of as a distance measure between the two probability distributions $\mathbb{P}$ and $\mathbb{P}'$, but it is not a true metric; it is not symmetric and does not obey the triangle inequality.

We summarize below some basic but important properties of these information theoretic quantities, which will be useful in our proofs later on. For proofs of the properties themselves the reader is referred to [23].

1. $0 \leq D(\mathbb{P}||\mathbb{P}') \leq \infty$.

2. $0 \leq H(X) \leq \log_2 |\mathcal{X}|$.

3. $0 \leq H(X|Y) \leq H(X)$ and $0 \leq H(X|Y,Z) \leq H(X|Y)$.

4. $0 \leq H(X,Y) \leq H(X) + H(Y)$. More generally, $0 \leq H(X_1^n) \leq \sum_{t=1}^n H(X_t)$.

5. The joint entropy is symmetric in X and Y, but the conditional entropy (in general) is not: $H(X,Y) = H(Y,X)$, but $H(X|Y) \neq H(Y|X)$. Similar statements hold with X replaced by X_n and Y replaced with X_1^{n-1}.

6. $H(X,Y) = H(Y)+H(X|Y) = H(X)+H(Y|X)$. More generally, $H(X_1^n) = \sum_{t=1}^n H(X_t|X_1^{t-1})$.

7. Special Cases: $H(X) = 0$ if and only if $\mathbb{P}(x) = 1$ for some $x \in \mathcal{X}$, and $H(X) = \log_2 |\mathcal{X}|$ if and only if $\mathbb{P}(X)$ is the uniform distribution on $\mathcal{X}$. $H(X|Y) = H(X)$ if and only if X and Y are independent. Similarly, $H(X,Y) = H(X) + H(Y)$ if and only if X and Y are independent. $D(\mathbb{P}||\mathbb{P}') = 0$ if and only if $\mathbb{P} = \mathbb{P}'$.

8. $H(\mathbb{P})$ is a continuous and concave function of the measure $\mathbb{P}$.

2.3.2 The Entropy Rate and its Block Estimates

Definition 4. *For a stationary process* (X_t) *the* entropy rate h *is the (averaged) asymptotic uncertainty in prediction of the next symbol given observation of all previous symbols:*

$$h \equiv \lim_{t\to\infty} H(X_t|X_1^{t-1})$$

Existence of the limit is a consequence of property 3 above, which ensures that the sequence of *block estimates*:

$$h(t) \equiv H(X_t|X_1^{t-1}) = H(X_0|X_{-t+1}^{-1})$$

is monotonically decreasing and, thus, must converge to some limiting value $h \geq 0$. However, the rate at which these estimates converge depends very much on the particular process. In the remainder of this dissertation, we will be interested primarily in establishing conditions for exponential convergence in hidden Markov processes and bounds on the rate of convergence:

$$\alpha = \limsup_{t \to \infty} \ \{h(t) - h\}^{1/t}$$

The following lemma gives two simple alternative expressions for the entropy rate h. Both are conceptually nice equivalences, but will also be technically useful later in establishing bounds on the rate of convergence of the block estimates $h(t)$ for various types of HMMs.

Lemma 1. *For a stationary process $(X_t)_{t \in \mathbb{Z}}$:*

$$h = \lim_{t \to \infty} H(X_1^t)/t \tag{2}$$

and:

$$h = \int H(X_0 | x_{-\infty}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) \tag{3}$$

where:

$$H(X_0|x_{-\infty}^{-1}) \equiv \begin{cases} -\sum_x \mathbb{P}(x|x_{-\infty}^{-1}) \log_2 \mathbb{P}(x|x_{-\infty}^{-1}) \ , \ \textit{if } x_{-\infty}^{-1} \textit{ is regular} \\ 0 \ , \ \textit{else (by convention)} \end{cases}$$

Proof. (2) follows from the decomposition of property 6:

$$H(X_1^t) = \sum_{\tau=1}^{t} H(X_\tau | X_1^{\tau-1}) = \sum_{\tau=1}^{t} h(\tau)$$

and the fact that the block estimates $h(\tau)$ limit to h, by definition.

(3) follows from the dominated convergence theorem. For every regular past $x_{-\infty}^{-1}$, $H(X_0|x_{-t}^{-1}) \to$

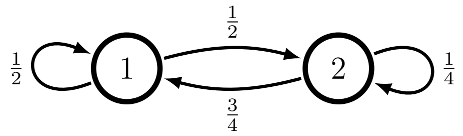
$H(X_0|x_{-\infty}^{-1})$, and $H(X_0|x_{-t}^{-1})$ is bounded by $\log_2|\mathcal{X}|$, for all t. Thus:

$$\begin{aligned}
h &\equiv \lim_{t\to\infty} h(t)\\
&= \lim_{t\to\infty} H(X_0|X_{-t}^{-1})\\
&= \lim_{t\to\infty} \sum_{w\in\mathcal{L}_t} H(X_0|X_{-t}^{-1}=w)\cdot\mathbb{P}(w)\\
&= \lim_{t\to\infty} \sum_{w\in\mathcal{L}_t} \int_{Z_w} H(X_0|x_{-t}^{-1})d\mathbb{P}(x_{-\infty}^{-1})\\
&= \lim_{t\to\infty} \int H(X_0|x_{-t}^{-1})d\mathbb{P}(x_{-\infty}^{-1})\\
&= \int H(X_0|x_{-\infty}^{-1})d\mathbb{P}(x_{-\infty}^{-1})
\end{aligned}$$

where Z_w is the set of regular pasts $x_{-\infty}^{-1}$ such that $x_{-|w|}^{-1}=w$. □

2.4 Markov Chains

We will consider here only finite-state Markov chains, because our results are only for finite HMMs. We review below the basic definition and some important properties. For a more thorough introduction the reader is referred to reference [24].



$$\mathcal{T} = \begin{pmatrix} 1/2 & 1/2 \\ 3/4 & 1/4 \end{pmatrix}$$

Figure 1: A 2 state Markov model. The graphical representation is presented on the left and the transition matrix $\mathcal{T}$ on the right. The state set is $\mathcal{S}=\{1,2\}$.

Definition 5. *A (finite)* Markov model *is a pair $(\mathcal{S},\mathcal{T})$ where:*

- *$\mathcal{S}$ is a finite set of states.*
- *$\mathcal{T}$ is an $|\mathcal{S}|\times|\mathcal{S}|$ stochastic*[2] *transition matrix.*

[2] I.e., nonnegative entries and all rows sum to 1.

Visually, a Markov model can be depicted as a directed graph with labeled edges, as in Figure 1. The vertices are the states and, for each pair of states i, j with $\mathcal{T}_{ij} > 0$, there is a directed edge from i to j labeled with the transition probability $\mathcal{T}_{ij}$. From the current state S_t the next state S_{t+1} is determined by selecting an outgoing edge from S_t according to these probabilities.

More formally, for a fixed state $i \in \mathcal{S}$ the *Markov process* or *Markov chain* generated by the Markov model $(\mathcal{S}, \mathcal{T})$ from initial state i is the random state sequence $(S_t)_{t \in \mathbb{Z}^+}$ with distribution $\mathbb{P}_i$ defined by the relations:

$$\begin{cases} \mathbb{P}_i(S_0 = i) = 1 \\ \mathbb{P}_i(S_{t+1} = k | S_t = j, S_0^{t-1} = s_0^{t-1}) = \mathcal{T}_{jk} \quad , \quad t \geq 0, j, k \in \mathcal{S}, s_0^{t-1} \in \mathcal{S}^t \end{cases}$$

By linearity this definition may be extended to an arbitrary initial state distribution ρ. That is, the *Markov process* generated by the model $(\mathcal{S}, \mathcal{T})$ from an initial state distribution ρ is the state sequence $(S_t)_{t \in \mathbb{Z}^+}$ with distribution $\mathbb{P}_\rho$ defined by:

$$\mathbb{P}_\rho(\cdot) = \sum_i \rho_i \cdot \mathbb{P}_i(\cdot)$$

By induction on t, it can easily be shown that the distribution of the random variable S_t according to the measure $\mathbb{P}_\rho$ may be calculated simply by applying the t_{th} power of the transition matrix $\mathcal{T}$ to the initial distribution ρ:

$$\mathbb{P}_\rho(S_t) = \rho \mathcal{T}^t \ , \ t \geq 0$$

2.4.1 Transience, Recurrence, and Irreducibility

For a Markov model $(\mathcal{S}, \mathcal{T})$, state j is said to be *accessible* from state i if there exists a path in the associated graph from i to j or, equivalently, if $\mathcal{T}_{ij}^t > 0$ for some $t \geq 0$. A state i is said to *transient* if there exists some other state j, such that j is accessible from i, but i is not accessible from j. If a state is not transient it is said to be *recurrent.* If each state j is accessible from each other state i the Markov model is said to be *irreducible.* In this case, of course, all states are recurrent.

Moreover, for any irreducible Markov model $(\mathcal{S}, \mathcal{T})$ it can be shown that:

1. There is a unique *stationary distribution* π satisfying $\pi = \pi \mathcal{T}$,

2. $\pi_i > 0$, for each $i \in \mathcal{S}$, and

3. The Markov chain $(S_t)_{t \in \mathbb{Z}^+}$ generated by the model $(\mathcal{S}, \mathcal{T})$ with S_0 chosen according to π is a stationary and ergodic process.

In the following, many of the Markov models we consider will be irreducible and, in this case, we will denote the (unique) stationary measure on state sequences $(S_t)_{t \in \mathbb{Z}^+}$ simply by $\mathbb{P}$:

$$\mathbb{P}(\cdot) \equiv \mathbb{P}_\pi(\cdot)$$

2.4.2 Periodicity and Aperiodicity

The *period* p of an irreducible Markov model $(\mathcal{S}, \mathcal{T})$ is the greatest common divisor of the lengths of all loops in the associated graph:

$$p \equiv \gcd\{n : \mathcal{T}^n_{ii} > 0, \text{ for some } i \in \mathcal{S}\}$$

This terminology is motivated by the following fact: For a chain of period p the state set $\mathcal{S}$ can always be partitioned into p disjoint and non-empty state equivalence classes $\mathcal{E}_0, ..., \mathcal{E}_{p-1}$ such that all states in class $\mathcal{E}_k$ can transition only to states in class $\mathcal{E}_{(k+1 \mod p)}$. Thus, the flow on equivalence classes is periodic:

$$... \ \mathcal{E}_0, ..., \mathcal{E}_{p-1}, \mathcal{E}_0, ..., \mathcal{E}_{p-1}, \mathcal{E}_0, ..., \mathcal{E}_{p-1} \ ...$$

If $p = 1$ the Markov model $(\mathcal{S}, \mathcal{T})$, or the associated Markov chain it generates, is said to be *aperiodic*. In this case, there is always exponential convergence to the stationary distribution π from any initial distribution ρ. That is:

$$\|\mathbb{P}_\rho(S_t) - \pi\|_1 = \|\rho \mathcal{T}^t - \pi\|_1 \leq K\alpha^t \ , \ t \in \mathbb{N}$$

for some constants $K > 0$ and $0 < \alpha < 1$ that depend on the model $(\mathcal{S}, \mathcal{T})$, but not on t or the initial distribution ρ.

However, if $p > 1$ then convergence to the stationary distribution from all initial conditions is impossible. Any initial distribution with support only on states in class $\mathcal{E}_k$ can have support only on states in class $\mathcal{E}_{(k+t \mod p)}$ at time t.

2.4.3 Multi-Step Chains

For a Markov model $(\mathcal{S}, \mathcal{T})$ the associated *n-step model* $(\mathcal{S}, \mathcal{B})$ is the Markov model with state set $\mathcal{S}$ and transition matrix $\mathcal{B}$ defined by:

$$\mathcal{B}_{ij} = \mathcal{T}_{ij}^n$$

If the original model $(\mathcal{S}, \mathcal{T})$ is irreducible, then the n-step model $(\mathcal{S}, \mathcal{B})$ is also irreducible for all n relatively prime to the period p of the original model. However, if n is not relatively prime to p then the n-step model can never be irreducible.

In the special case $n = p$, the n-step model $(\mathcal{S}, \mathcal{B})$ has exactly p disjoint irreducible components, corresponding to the equivalences classes $\mathcal{E}_0, ..., \mathcal{E}_{p-1}$. That is, the graph associated to the n-step model has p disjoint strongly connected components with state sets $\mathcal{E}_0, ..., \mathcal{E}_{p-1}$, and for each equivalence class $\mathcal{E}_k$ the $|\mathcal{E}_k| \times |\mathcal{E}_k|$ matrix $\mathcal{B}_k$ defined by:

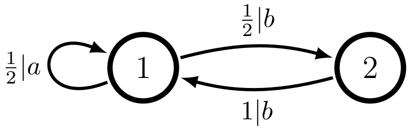
$$(\mathcal{B}_k)_{ij} = \mathcal{B}_{ij} \ , \ i, j \in \mathcal{E}_k$$

is stochastic.

Moreover, in this case, the irreducible Markov model $(\mathcal{E}_k, \mathcal{B}_k)$ associated to each equivalence class $\mathcal{E}_k$ is aperiodic. This fact can be quite useful in translating results for aperiodic chains to chains of general period, as we will do below in Section 3 with large deviation bounds.

2.5 Hidden Markov Models

As noted above, we will consider only the case of finite hidden Markov models where both the internal set of states $\mathcal{S}$ and set of output symbols $\mathcal{X}$ are finite. There are two primary types: state-emitting and edge-emitting. The state-emitting variety is perhaps the simpler of the two, and also the more commonly studied. However, our results will be derived in the edge-emitting setting, because some of the conditions we study are more natural in this context. For a translation of the main theorems to the state-emitting setting, the reader is referred to Appendix C.



$$T^{(a)} = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \end{pmatrix}$$

$$T^{(b)} = \begin{pmatrix} 0 & 1/2 \\ 1 & 0 \end{pmatrix}$$

Figure 2: A 2 state HMM known as the *Even Machine*. The HMM has internal states $\mathcal{S} = \{1, 2\}$ and output alphabet $\mathcal{X} = \{a, b\}$. The process it generates is called the *Even Process* [25] because there are always an even number of b's between consecutive a's. The graphical representation is presented on the left, with the corresponding transition matrices on the right. In the graphical representation edges are labeled $p|x$ for the transition probability $p = \mathcal{T}_{ij}^{(x)}$ and symbol x.

Definition 6. *A (finite, edge-emitting)* hidden Markov model *is a 3-tuple* $(\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ *where:*

- $\mathcal{S}$ *is a finite set of states.*
- $\mathcal{X}$ *is a finite alphabet of output symbols.*
- $\mathcal{T}^{(x)}, x \in \mathcal{X}$ *are* $|\mathcal{S}| \times |\mathcal{S}|$ *sub-stochastic*[3] *symbol-labeled transition matrices whose sum* $\mathcal{T}$ *is stochastic.*

Like Markov models, a hidden Markov model (HMM) can be depicted visually as a directed graph with labeled edges, as in Figure 2. The vertices are the states, and for each i, j, x with $\mathcal{T}_{ij}^{(x)} > 0$ there is directed edge from i to j labeled with the symbol x and transition probability $\mathcal{T}_{ij}^{(x)}$. The sum of the probabilities on all outgoing edges from each state is 1. The operation of the HMM is as follows: From the current state S_t the HMM picks an outgoing edge E_t according to their probabilities, generates the symbol X_t labeling this edge, and then follows the edge to the next state S_{t+1}.

More formally, for an initial state i we have measure $\mathbb{P}_i$ on joint state-symbol sequences $(S_t, X_t)_{t \in \mathbb{Z}^+}$ defined by:

$$\begin{cases} \mathbb{P}_i(S_0 = i) = 1 \\ \mathbb{P}_i(S_{t+1} = k, X_t = x | S_t = j, S_0^{t-1} = s_0^{t-1}, X_0^{t-1} = x_0^{t-1}) = \mathcal{T}_{jk}^{(x)} \ , \\ \quad \text{for all } j, k \in \mathcal{S}, x \in \mathcal{X}, t \geq 0, \text{ and } (s_0^{t-1}, x_0^{t-1}) \in \mathcal{S}^t \times \mathcal{X}^t \end{cases} \tag{4}$$

[3] I.e., nonnegative entries and row sums less than or equal to 1.

If the initial state is chosen according to a distribution ρ rather than as a fixed state i we have, as for Markov chains, measure $\mathbb{P}_\rho$ defined by linearity:

$$\mathbb{P}_\rho(\cdot) = \sum_i \rho_i \cdot \mathbb{P}_i(\cdot) \tag{5}$$

From these definitions it follows, of course, that the state sequence $(S_t)_{t \in Z^+}$ is a Markov chain with transition matrix $\mathcal{T} = \sum_x \mathcal{T}^{(x)}$. However, we will be interested not only in the state sequence (S_t), but also (in fact primarily) in the associated sequence of output symbols (X_t), which is not generally Markovian. The idea is that an observer of the HMM may directly observer this sequence of outputs, but not the hidden internal states.

2.5.1 Basic Assumptions

To avoid some trivial, but exceptional, cases, in the remainder of this dissertation we will always make the following two assumptions for any HMM:

1. Each symbol $x \in \mathcal{X}$ can actually be generated with positive probability. That is, for each $x \in \mathcal{X}$, $\mathcal{T}_{ij}^{(x)} > 0$ for some $i, j \in \mathcal{S}$. If this is not the case, one can restrict the alphabet to $\mathcal{X}/\{x\}$.

2. The state set $\mathcal{S}$ and output alphabet $\mathcal{X}$ both always have cardinality at least 2. If this is not the case, all stated results are essentially trivial, but some of the definitions and constructions used in the proofs may be ambiguous or undefined.

These assumptions will no longer be stated explicitly, but are henceforth intended as part of our definition of HMM.

2.5.2 Inherited Properties and Stationary Measures

A HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is said to be, respectively, irreducible or aperiodic if the underlying Markov model $(\mathcal{S}, \mathcal{T})$ is irreducible or aperiodic. Similarly state j of a HMM is said to be accessible from state i if j is accessible from i in the underlying Markov model, and j is said to be transient (respectively recurrent) if j is transient (respectively recurrent) in the underlying Markov model.

For an irreducible HMM, we denote by $\mathbb{P}$ the (unique) stationary, ergodic measure on joint state-symbol sequences $(S_t, X_t)_{t \in \mathbb{Z}^+}$ given by choosing S_0 according to the stationary distribution π for the underlying Markov chain:

$$\mathbb{P}(\cdot) \equiv \mathbb{P}_\pi(\cdot)$$

By stationarity, we may uniquely extend this one-sided, stationary measure $\mathbb{P}$ to a two-sided (stationary) measure on bi-infinite state-symbol sequences $(S_t, X_t)_{t \in \mathbb{Z}}$. From the two-sided measure $\mathbb{P}$ we may then define two-sided measures $\mathbb{P}_i$ and $\mathbb{P}_\rho$ by:

$$\mathbb{P}_i(\cdot) = \mathbb{P}(\cdot | S_0 = i) \text{ and } \mathbb{P}_\rho(\cdot) = \sum_i \rho_i \cdot \mathbb{P}_i(\cdot)$$

consistent with the corresponding one-sided measures $\mathbb{P}_i$ and $\mathbb{P}_\rho$ defined above by (4) and (5). In the following development most HMMs we consider will be irreducible, and in this case we will use $\mathbb{P}$, $\mathbb{P}_i$, and $\mathbb{P}_\rho$ to refer to both one-sided measures and the corresponding two-sided measures interchangeably.

If not otherwise specified, we will assume random variables for an irreducible HMM are distributed according to the stationary measure $\mathbb{P}$. So, for example, an expression like $H(S_t | X_0^{t-1})$ represents the conditional entropy in the state S_t given the output sequence X_0^{t-1} when the distribution over joint sequences $(S_t, X_t)_{t \in \mathbb{Z}}$ is given by the measure $\mathbb{P}$. The entropy rate h and block estimate $h(t)$ of an irreducible hidden Markov model are, by definition, the entropy rate and block estimate of its stationary output process $\mathcal{P} = (X_t)_{t \in \mathbb{Z}}$.

2.5.3 Additional Notation

In Section 2.5.4, below, we will define several other important properties of HMMs, which do not have real analogs for ordinary Markov models. Before doing so, however, we introduce some notation here that will be needed for many of the definitions or later in our proofs. Throughout $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is assumed to be a fixed HMM.

1. We denote the minimum nonzero transition probability of the HMM M by $p_{\min}$ and the

maximum transition probability by $p_{\max}$:

$$p_{\min} \equiv \min\{\mathcal{T}^{(x)}_{ij} : \mathcal{T}^{(x)}_{ij} > 0\} \ \text{ and } \ p_{\max} \equiv \max\{\mathcal{T}^{(x)}_{ij}\}$$

In the case M is irreducible, we denote also by $r_{\min}$ and $r_{\max}$ the minimum and maximum ratios of the stationary state probabilities:

$$r_{\min} \equiv \min_{i,j} \pi_i/\pi_j \ \text{ and } \ r_{\max} \equiv \max_{i,j} \pi_i/\pi_j$$

2. For a word $w \in \mathcal{X}^*$, we denote by $\mathbb{P}_i(w)$ the probability of generating w starting in state i, and by $\mathbb{P}_\rho(w)$ the probability of generating w with the initial state chosen according to the distribution ρ:

$$\mathbb{P}_i(w) \equiv \mathbb{P}_i(X_0^{|w|-1} = w) \ \text{ and } \ \mathbb{P}_\rho(w) \equiv \mathbb{P}_\rho(X_0^{|w|-1} = w)$$

In the case M is irreducible, $\mathbb{P}(w)$ is the probability of the word w for the stationary output process of the HMM $(X_t)_{t \in \mathbb{Z}}$:

$$\mathbb{P}(w) \equiv \mathbb{P}_\pi(w)$$

Also, for a set of words A of some fixed length n:

$$\mathbb{P}_i(A) \equiv \sum_{w \in A} \mathbb{P}_i(w) \ , \ \mathbb{P}_\rho(A) \equiv \sum_{w \in A} \mathbb{P}_\rho(w) \ , \ \text{and} \ \mathbb{P}(A) \equiv \sum_{w \in A} \mathbb{P}(w)$$

3. For a state $i \in \mathcal{S}$, $\mathcal{L}(i)$, $\mathcal{L}_n(i)$, and $\mathcal{L}^+_\infty(i)$ are, respectively, the language, length-n restriction, and set of allowable infinite futures of state i:

$$\begin{aligned}\mathcal{L}(i) &\equiv \{w \in \mathcal{X}^* : \mathbb{P}_i(w) > 0\} \\ \mathcal{L}_n(i) &\equiv \{w \in \mathcal{L}(i) : |w| = n\} \\ \mathcal{L}^+_\infty(i) &\equiv \{x_0^\infty : x_0^t \in \mathcal{L}(i), \ \text{for all } t \geq 0\}\end{aligned}$$

Also, $\mathcal{L}(M)$, $\mathcal{L}_n(M)$, and $\mathcal{L}_\infty^+(M)$ are the language, length-n restriction, and set of allowable infinite futures of the HMM M:

$$\mathcal{L}(M) \equiv \bigcup_i \mathcal{L}(i) \ , \ \mathcal{L}_n(M) \equiv \bigcup_i \mathcal{L}_n(i) \ , \ \mathcal{L}_\infty^+(M) \equiv \bigcup_i \mathcal{L}_\infty^+(i)$$

4. For $i \in \mathcal{S}$ and $w \in \mathcal{X}^*$, $\delta_i(w)$ is the set of states j, which state i can transition to upon emitting the word w:

$$\delta_i(w) \equiv \{j \in \mathcal{S} : \mathbb{P}_i(X_0^{|w|-1} = w, S_{|w|} = j) > 0\} \tag{6}$$

Also, for $A \subseteq \mathcal{S}$:

$$\delta_A(w) \equiv \bigcup_{i \in A} \delta_i(w) \tag{7}$$

5. Finally, for a word $w \in \mathcal{X}^*$, $\phi_i(w)$ is the distribution over the current state induced by observing w from initial state i and $\phi_\rho(w)$ is the distribution induced by observing w with the initial state chosen according to ρ:

$$\phi_i(w) \equiv \mathbb{P}_i(S_{|w|} | X_0^{|w|-1} = w)$$
$$\phi_\rho(w) \equiv \mathbb{P}_\rho(S_{|w|} | X_0^{|w|-1} = w)$$

In the case M is irreducible, $\phi(w)$ is the distribution over the current state induced by observing w with initial state chosen according to the stationary distribution π:

$$\phi(w) \equiv \phi_\pi(w)$$

In this case, we also denote the uncertainty in the induced distribution $\phi(w)$ by $u(w)$, and the uncertainty in the induced distribution $\phi(X_0^{t-1})$ or *state uncertainty* at time t by U_t. The *expected state uncertainty*, $\mathbb{E}(U_t)$, is the expected value of the random variable U_t with respect to the stationary measure π.

$$\begin{aligned} u(w) &\equiv H(\phi(w)) \\ U_t &\equiv H(\phi(X_0^{t-1})) = u(X_0^{t-1}) \\ \mathbb{E}(U_t) \equiv \mathbb{E}_\pi(U_t) &= \sum_{w \in \mathcal{L}_t(M)} \mathbb{P}(w) u(w) \end{aligned}$$

Note that with the conventions introduced in Section 2.2 $\phi_i(w)$, $\phi_\rho(w)$, and $\phi(w)$ are each the null distribution consisting of all zeros if, respectively, $\mathbb{P}_i(w) = 0$, $\mathbb{P}_\rho(w) = 0$, or $\mathbb{P}(w) = 0$. Thus, with our conventions for logarithms, $u(w) = -\sum_{i \in \mathcal{S}} \phi(w)_i \log_2(\phi(w)_i) = 0$, for any word $w \notin \mathcal{L}(M)$ (where $\phi(w)_i$ is the i_{th} component of the probability vector $\phi(w)$).

2.5.4 Additional Properties

With the requisite notation established, we now proceed to the definitions of some key HMM properties, including exactness, unifilarity, and the flag-state property, which will be the focus of Sections 4, 5, and 6, respectively. Properties are listed below, followed by several comments about relations between them, testability, commonality, and so forth.

Properties

- A HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is *unifilar* if for each state i and symbol x there is at most 1 outgoing edge from state i labeled with symbol x in the associated graph:

$$|\delta_i(x)| \leq 1 \ , \quad \text{for each } i \in \mathcal{S}, x \in \mathcal{X}$$

- A word $w \in \mathcal{L}(M)$ is a *synchronizing word* or *sync word* for a HMM M if the observer knows the state of the HMM with certainty after seeing w:

$$\delta_{\mathcal{S}}(w) = \{j\} \ , \quad \text{for some } j \in \mathcal{S}$$

  A HMM is *exactly synchronizeable* or simply *exact* if it has some finite synchronizing word w. That is, if it is possible for an observer to synchronize exactly to the state of the HMM after observing a finite sequence of outputs.

- A word $w \in \mathcal{L}(M)$ is a *flag word* for state j of a HMM M if each state i that can generate the word w may transition to state j on w:

  $$j \in \delta_i(w) \ , \quad \text{for all } i \text{ with } \mathbb{P}_i(w) > 0$$

  This implies, in particular, that the observer always believes state j is possible as the current state of the HMM after seeing w, regardless of all previous output:

  $$j \in \delta_{\mathcal{S}}(vw) \ , \quad \text{for all } v \text{ with } vw \in \mathcal{L}(M)$$

  Thus, we think of the word w as signaling or flagging to the observer that state j is possible.

  If a state j has some flag word w, then we say it is a *flag state.* If each state j of a HMM M is a flag state, then we say M is a *flag-state HMM* (or *FS HMM*). If each state j has a length-1 flag word (or *flag symbol*) we say M is *FS1.*

- States i, j of a HMM M are said to be *probabilistically distinguishable* (or *PD*) if there is some word $w \in \mathcal{X}^*$ with:

  $$\mathbb{P}_i(w) \neq \mathbb{P}_j(w)$$

  If each pair of distinct state i, j is PD then we say that the M is a *PD HMM* or that M has *probabilistically distinguishable states.* If for each pair of distinct states i, j there is a symbol $x \in \mathcal{X}$ (length-1 word) such that:

  $$\mathbb{P}_i(x) \neq \mathbb{P}_j(x)$$

  then we say that the HMM is *PD1.*

- States i, j of a HMM M are said to be *incompatible* if there exists some length $n \in \mathbb{N}$ such that the sets of allowed words of length n that can be generated from the two states i and j

(and hence the sets of allowed words of any length $m > n$) have no overlap:

$$\mathcal{L}_n(i) \cap \mathcal{L}_n(j) = \emptyset$$

- States i, j of a HMM M are said to be *path-mergeable* if there exists some state k and word $w \in \mathcal{X}^*$ such that both i and j can transition to k on w:

$$k \in \delta_i(w) \cap \delta_j(w)$$

Comments

1. All of these properties are explicitly testable for a given HMM M. Unifilarity is immediate. Exactness can be tested using the possibility machine construction given in Section 4. Tests for the other properties are given in Appendix A.

2. In Appendix B we will show that an irreducible HMM is a flag-state HMM if every distinct pair of states i, j is either path-mergeable or incompatible. This implies, as will be discussed there further, that the flag-state condition is quite weak, and should in some sense "usually" be satisfied.

   The proofs given below for flag-state HMMs are not as conceptually appealing as those for either exact or unifilar HMMs, and the bounds on convergence of the entropy rate estimates $h(t)$ will usually not be as good for flag-state HMMs as for the other cases, either. Nevertheless, the results for flag-state models are important, because they are the most generally applicable.

3. In our earlier works [26, 27] the term *ϵ-machine* was used for a (finite) unifilar, PD HMM. More generally, ϵ-machines are minimal unifilar presentations of stationary processes whose states consist of equivalence classes of infinite pasts $x_{-\infty}^{-1}$ with same conditional probability distribution over the random future X_0^{∞} [21, 22]. In the finite case the definition given in [26, 27] is equivalent to the more general definition in terms of equivalences classes of pasts. However, the state set of an ϵ-machine is not always finite even if the process alphabet $\mathcal{X}$ is, and the process alphabet $\mathcal{X}$ need not be finite either for the ϵ-machine presentation to be well defined.

In later sections of this dissertation we will not appeal directly to this representation of a stationary process in terms of equivalence classes of pasts $x_{-\infty}^{-1}$ or use the term ϵ-machine explicitly. Nevertheless, it is worth noting that much of our interest in unifilar HMMs is motivated by the ϵ-machine construction, and that this is one of the main reasons we feel that unifilarity is a natural property to study.

It is also worth noting that, while we will not present it this way because it is simpler not to, the proof technique for bounding convergence of the entropy rate estimates in exact HMMs was inspired by the ϵ-machine representation. Indeed for a (finite) exact HMM the ϵ-machine always has (at most) a countable set of states, and any word w that is synchronizing for the original HMM M is also synchronizing for the ϵ-machine of the stationary output process $\mathcal{P} = (X_t)_{t \in \mathbb{Z}}$ generated by M. This fact can be used to show that for a (finite) exact HMM, $h(t+1) - h \leq \mathbb{P}(NSYN_t) \cdot \log_2 |\mathcal{X}|$, for all t, where $NSYN_t$ is the set of words $w \in \mathcal{L}_t(M)$ that do not contain any synchronizing subword. The same bound will be established by alternate and more direct means in Section 4, but the realization first came about by considering the ϵ-machine representation.

4. The following are a number of immediate relations and implications concerning the various properties and definitions. None of them are at all difficult to prove, but they are worth stating explicitly, because otherwise they may not occur to the reader to consider. In our proofs in later sections these implications will frequently be used implicitly without direct reference.

   (i) If w is a synchronizing word for a HMM M, then so is vw for any word v with $vw \in \mathcal{L}(M)$.

   (ii) If M is unifilar then extensions of synchronizing words are also synchronizing words. That is, wv is a sync word for any sync word w and word v with $wv \in \mathcal{L}(M)$.

   (iii) If w is a flag word for a state j, then so is vw for any word v with $vw \in \mathcal{L}(M)$.

   (iv) If w is a flag word for a state j and $v \in \mathcal{L}(j)$, then wv is flag word for each state $i \in \delta_j(v)$.

   (v) It follows from (i) that for an exact, irreducible HMM M there is some sync word $w_i \in \mathcal{L}(i)$ for each state i. In particular, if $w \in \mathcal{L}(j)$ is a sync word for some state j, then for each i the word $w_i = v_i w \in \mathcal{L}(i)$ is a sync word, where v_i is any word such that

$j \in \delta_i(v_i)$. Existence of the v_is is, of course, guaranteed by irreducibility.

(vi) It follows from (iv) that if some state j of an irreducible HMM M has a flag word w, then each state i also has some flag word w_i. In particular, for each i, $w_i = wv_i$ is a flag word for state i, where v_i is any word such that $i \in \delta_j(v_i)$. Thus, an irreducible HMM is a flag-state HMM if it has any state j with a flag word.

(vii) For an exact, irreducible HMM M any sync word w will eventually be generated in a.e. infinite output sequence x_0^∞. Thus, by (i), for $\mathbb{P}$ a.e. $x_0^\infty \in \mathcal{L}_\infty^+(M)$ there exists some finite time $t \in \mathbb{N}$ such that x_0^t is sync word. If the HMM M is also unifilar then, by (ii), we know that, in addition, for any such output sequence x_0^∞, x_0^τ is a sync word for all $\tau \geq t$.

(viii) The sync word definition is equivalent to the following: $w \in \mathcal{L}(M)$ is a sync word if there is some state $j \in \mathcal{S}$ such that:

$$\delta_i(w) = \{j\} \ , \quad \text{for all } i \text{ with } \mathbb{P}_i(w) > 0 \tag{8}$$

On the other hand, by definition, $w \in \mathcal{L}(M)$ is flag word for a state j if:

$$j \in \delta_i(w) \ , \quad \text{for all } i \text{ with } \mathbb{P}_i(w) > 0 \tag{9}$$

Thus, being a sync word is a strictly stronger property than being a flag word. After seeing a sync word w the observer knows the HMM must be in some given state j, regardless of previous output, whereas after seeing a flag word w the observer knows that state j is always possible, regardless of previous output, but other states may be possible as well. Of course, this means that exactness is a strictly stronger property than the flag-state property, so our results for flag-state HMMs also hold for exact HMMs. However, in the specific case of exact HMMs the bounds we obtain will be much better.

(ix) For a unifilar HMM, $|\delta_i(w)| \leq 1$ for each state i and word w. Thus, in light of (8) and (9), a flag word for a unifilar HMM is necessarily a sync word. So a unifilar, flag-state HMM is necessarily exact.

2.6 A Remark on Additional Notation in Later Sections

All notation introduced above in Section 2 will be used consistently throughout this dissertation. However, the reader should note that additional notation introduced later in Sections 4, 5, and 6 for various constructions involving exact, unifilar, and flag-state HMMs is taken to be internal to those sections and not to transfer between. Indeed, the same symbols will be used to represent different objects in Sections 4, 5, and 6.

3 Preliminaries

In this section we prove some important information theoretic inequalities and lemmas on large deviations and hitting times for Markov chains, which will be useful in the following sections for establishing bounds on convergence of the entropy rate block estimates $h(t)$. We also introduce the block model M^n of a HMM M, which will be useful in establishing convergence of the entropy estimates and related quantities for unifilar and flag-state HMMs. Several of the results presented here are probably well known, but we were not generally able to find references for them except where explicitly stated. So, proofs are given for completeness when references are not.

3.1 Information Theoretic Inequalities

We present two inequalities relating the difference in entropies of two distributions, $|H(\mu) - H(\nu)|$, to their difference in total variational norm, $\|\mu - \nu\|_{TV} \equiv \frac{1}{2}\|\mu - \nu\|_1$. The first is a lower bound and the second an upper bound.

Lemma 2. *Let $\mu = (\mu_1, ..., \mu_N)$ be a probability measure on the finite set $\{1, ..., N\}$. Let $k_{\max} =$ argmax μ_k (with lowest index k in case of a tie) and let $\nu = e_{k_{\max}}$. If $\|\mu - \nu\|_{TV} \equiv \epsilon \leq \min\{\frac{1}{e}, \frac{1}{2^{N-1}}\}$ then:*

$$H(\mu) = |H(\mu) - H(\nu)| \geq \epsilon$$

Proof. In the case $\epsilon = 0$ the inequality is trivial; both sides are 0. Thus, let us assume $\epsilon > 0$, which implies, of course, that $N \geq 2$. In this case, since $\sum_{k \neq k_{\max}} \mu_k = \epsilon > 0$, there must be some $k \neq k_{max}$ with $\mu_k \geq \epsilon/(N-1)$. So we have:

$$H(\mu) = \sum_j -\mu_j \log_2 \mu_j \geq -\mu_k \log_2 \mu_k \geq -\left(\frac{\epsilon}{N-1}\right) \log_2 \left(\frac{\epsilon}{N-1}\right) \tag{10}$$

The last inequality follows from the fact that $f(x) = -x \log_2(x)$ is increasing on the interval $[0, 1/e]$ and $\frac{\epsilon}{N-1} \leq \mu_k \leq \epsilon \leq 1/e$. The claim follows from (10) since for all $0 < \epsilon \leq \frac{1}{2^{N-1}}$:

$$-\left(\frac{\epsilon}{N-1}\right) \log_2 \left(\frac{\epsilon}{N-1}\right) = \left(\frac{\epsilon}{N-1}\right) \log_2 \left(\frac{N-1}{\epsilon}\right) \geq \left(\frac{\epsilon}{N-1}\right) \log_2 \left(\frac{1}{\epsilon}\right) \geq \epsilon$$

□

Lemma 3. *Let $\mu = (\mu_1, ..., \mu_N)$ and $\nu = (\nu_1, ..., \nu_N)$ be two probability measures on the finite set $\{1, ..., N\}$. If $\|\mu - \nu\|_{TV} \equiv \epsilon \leq 1/e$ then:*

$$|H(\mu) - H(\nu)| \leq N\epsilon \log_2(1/\epsilon)$$

Proof. If $\|\mu - \nu\|_{TV} = \epsilon$ then:

$$|\mu_k - \nu_k| \leq \epsilon, \text{ for all } k$$

Thus:

$$\begin{aligned}
|H(\mu) - H(\nu)| &= \left| \sum_k \nu_k \log_2(\nu_k) - \mu_k \log_2(\mu_k) \right| \\
&\leq \sum_k |\nu_k \log_2(\nu_k) - \mu_k \log_2(\mu_k)| \\
&\leq N \cdot \max_{\epsilon' \in [0,\epsilon]} \max_{x \in [0,1-\epsilon']} |(x + \epsilon') \log_2(x + \epsilon') - x \log_2(x)| \\
&= N \cdot \max_{\epsilon' \in [0,\epsilon]} \epsilon' \log_2(1/\epsilon') \\
&= N \cdot \epsilon \log_2(1/\epsilon)
\end{aligned}$$

The last two equalities, for $0 \leq \epsilon' \leq \epsilon \leq 1/e$, may be verified using single variable calculus techniques for maximization. □

3.2 Large Deviation and Hitting Times Estimates for Markov Chains

We present two large deviation estimates for irreducible Markov chains, and also an important lemma on hitting times in Markov chains (Lemma 6). The primary large deviation estimate for irreducible, aperiodic chains (Lemma 4) is given, essentially, in reference [28]. The subsequent lemma presented here is simply an asymptotic extension dropping the aperiodicity hypothesis.

Lemma 4. *Let $(\mathcal{S}, \mathcal{T})$ be an irreducible, aperiodic Markov model with stationary distribution π, and let $(S_t)_{t \in \mathbb{Z}^+}$ be the associated Markov chain it generates. Also, let $f : \mathcal{S} \to \mathbb{R}$, let $Y_t = f(S_t)$,* *and let $Y_t^{avg} = \frac{1}{t}\sum_{n=0}^{t-1} Y_n$. Then for any initial state $i \in \mathcal{S}$ and $\epsilon > 0$:*

$$\mathbb{P}_i(Y_t^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) \leq \exp\left(-\frac{(t\epsilon - 4 \cdot \frac{m}{q}\|f\|_\infty)^2}{8t \cdot \frac{m^2}{q^2}\|f\|_\infty^2}\right) \text{ , for } t \text{ sufficiently large} \quad (11)$$

where $\|f\|_\infty \equiv \max_{i\in\mathcal{S}} |f(i)|$ is the sup norm of f, $\mathbb{E}_\pi(f) \equiv \sum_{i\in\mathcal{S}} \pi_i f(i)$ is the expected value of f with respect to the stationary measure π, and $m \in \mathbb{N}$, $0 < q < 1$ are any constants such that there exists a probability distribution ρ on $\mathcal{S}$ satisfying:

$$\mathbb{P}_i(S_m \in \cdot) \geq q \cdot \rho(\cdot) \text{ , for each } i \in \mathcal{S} \quad (12)$$

Remark. *Note that since $\mathbb{P}_i(S_t = j) \to \pi_j$, as $t \to \infty$, for any irreducible, aperiodic Markov chain there is always some such triple ρ, q, m satisfying (12). In particular, if we take $\rho = e_j$ and $q = \pi_j/2$, then there must be some m such that (12) is satisfied. The bound (11), however, holds for any m, q, ρ satisfying (12). To obtain an ideal bound on the large deviation rate for a given Markov chain one may attempt to maximize over possible m, q.*

Proof. A very similar statement is given in reference [28] for Markov chains on a general state space with some m, q, ρ satisfying the condition (12). However, there is a slight error in the proof[4]. When this error is corrected the bound changes from the one presented in [28] to the one given above (11). □

Lemma 5. *Let $(\mathcal{S}, \mathcal{T})$ be an irreducible Markov model of period $p \geq 1$, let $(\mathcal{S}, \mathcal{B})$ be the associated p-step model defined in Section 2.4.3, and let $(\mathcal{E}_k, \mathcal{B}_k)$ be the (irreducible, aperiodic) restriction of the model $(\mathcal{S}, \mathcal{B})$ to states in the k_{th} periodic equivalence class $\mathcal{E}_k$:*

$$(\mathcal{B}_k)_{ij} = \mathcal{B}_{ij} \text{ , } i, j \in \mathcal{E}_k$$

[4] In the proof the authors state (in our notation) that:

$$|\mathbb{E}_i\{Y_t - \mathbb{E}_\pi(f)\}| \leq \|f\|_\infty \cdot (1-q)^{\lfloor t/m \rfloor} \text{ , for each } i \in \mathcal{S}, t \in \mathbb{N}$$

where $\mathbb{E}_i(\cdot) \equiv \mathbb{E}(\cdot|S_0 = i)$. This should instead be:

$$|\mathbb{E}_i\{Y_t - \mathbb{E}_\pi(f)\}| \leq 2\|f\|_\infty \cdot (1-q)^{\lfloor t/m \rfloor} \text{ , for each } i \in \mathcal{S}, t \in \mathbb{N}$$

Also, as in the previous lemma, let $f : \mathcal{S} \to \mathbb{R}$ and define $Y_t = f(S_t)$, $Y_t^{avg} = \frac{1}{t}\sum_{n=0}^{t-1} Y_n$, and $\mathbb{E}_\pi(f) = \sum_i \pi_i f(i)$, where π is the stationary distribution for the original model $(\mathcal{S}, \mathcal{T})$. Then, for each $i \in \mathcal{S}$ and $\epsilon > 0$:

$$\limsup_{t\to\infty} \ \mathbb{P}_i(|Y_t^{avg} - \mathbb{E}_\pi(f)| \geq \epsilon)^{1/t} \leq \beta$$

where $\beta = \beta(\epsilon) \in (0,1)$ is defined by:

$$\beta \equiv \max_{0\leq k\leq p-1} \beta_k^{1/p}$$

with $\beta_k = \beta_k(\epsilon) \in (0,1)$ given by:

$$\beta_k \equiv \exp\left(-\frac{\epsilon^2 q_k^2}{8m_k^2\|f\|_\infty^2}\right)$$

Here m_k, q_k are the values of m, q satisfying the relation (12) for the (irreducible, aperiodic) Markov model $(\mathcal{E}_k, \mathcal{B}_k)$.

Proof. If we define $g = -f$ and $Z_t = g(S_t)$, then $\|g\|_\infty = \|f\|_\infty$ and:

$$\begin{aligned}\mathbb{P}_i(|Y_t^{avg} - \mathbb{E}_\pi(f)| \geq \epsilon) &= \mathbb{P}_i(Y_t^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) + \mathbb{P}_i(Y_t^{avg} - \mathbb{E}_\pi(f) \leq -\epsilon) \\ &= \mathbb{P}_i(Y_t^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) + \mathbb{P}_i(Z_t^{avg} - \mathbb{E}_\pi(g) \geq \epsilon)\end{aligned}$$

Thus, it suffices to show that the one sided bound:

$$\limsup_{t\to\infty} \ \mathbb{P}_i(Y_t^{avg} - \mathbb{E}_\pi(f) \geq \epsilon)^{1/t} \leq \beta$$

holds for any f with a given infinity norm.

In the case $p = 1$ this follows directly from the previous lemma. For $p > 1$ it can be proved by considering length-p blocks. Let us define:

$$Y_{k,t}^{avg} \equiv \frac{1}{t}\sum_{n=0}^{tp-1} \mathbb{1}\{S_n \in \mathcal{E}_k\} \cdot f(S_n)$$

as the empirical average of f evaluated at the first t states from equivalence class $\mathcal{E}_k$ in the length tp state sequence S_0^{tp-1} and let:

$$\mathbb{E}_{\pi^k}(f) \equiv \sum_{i \in \mathcal{E}_k} \pi_i^k \cdot f(i)$$

be the expected value of f with respect to the normalized stationary measure π^k on states in equivalence class $\mathcal{E}_k$ given by $\pi_i^k = p \cdot \pi_i$, $i \in \mathcal{E}_k$. Then:

$$\begin{aligned}
\mathbb{P}_i(Y_{tp}^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) &\leq \mathbb{P}_i(\exists k : Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon) \\
&\leq \sum_{k=0}^{p-1} \mathbb{P}_i(Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon) \\
&\leq p \cdot \max_k \mathbb{P}_i(Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon)
\end{aligned}$$

or, equivalently:

$$\mathbb{P}_i(Y_\tau^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) \leq p \cdot \max_k \mathbb{P}_i(Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon)$$

for any $\tau = tp, t \in \mathbb{N}$. If τ is not an integer multiple of p things are slightly more complicated. However, since $\mathcal{S}$ is finite, f is bounded. So, we still have:

$$\begin{aligned}
\mathbb{P}_i(Y_\tau^{avg} - \mathbb{E}_\pi(f) \geq \epsilon) &\leq \mathbb{P}_i(\exists k : Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon') \\
&\leq p \cdot \max_k \mathbb{P}_i(Y_{k,t}^{avg} - \mathbb{E}_{\pi^k}(f) \geq \epsilon') \qquad (13)
\end{aligned}$$

for any $\epsilon' < \epsilon$ and all sufficiently large τ, where $t = \lfloor \tau/p \rfloor$.

For any $\beta' > \max_k \beta_k(\epsilon')^{1/p}$ we know by the previous lemma that the quantity on the right hand side of (13) is bounded above by $(\beta')^\tau$, for all sufficiently large τ. Since each $\beta_k = \beta_k(\epsilon')$ is a continuous function of ϵ', β' can be chosen arbitrarily close to β by taking ϵ' arbitrarily close to ϵ. Therefore:

$$\limsup_{\tau \to \infty} \; \mathbb{P}_i(Y_\tau^{avg} - \mathbb{E}_\pi(f) \geq \epsilon)^{1/\tau} \leq \beta$$

as required. □

To state the final lemma on hitting times we need to introduce some terminology.

- If $i, j \in \mathcal{S}$ and B is a subset of $\mathcal{S}$ we say j *is accessible from* i *without passing through* B if there exists a state sequence s_0^t, $t \geq 0$ such that $s_0 = i$, $s_t = j$, $\mathbb{P}_i(S_0^t = s_0^t) > 0$, and $s_n \notin B$ for each $0 \leq n \leq t$.
- If $j \in \mathcal{S}$ and A and B are subsets of $\mathcal{S}$ we say that j *is accessible from* A *without passing through* B, if there exists $i \in A$ such that j is accessible from i without passing through B.
- If $i \in \mathcal{S}$ and A is a subset of $\mathcal{S}$ we say A *is accessible from* i if there exists $j \in A$ such that j is accessible from i.

Lemma 6. *Let $(\mathcal{S}, \mathcal{T})$ be a Markov model, not necessarily irreducible. Let A and B be disjoint non-empty subsets of $\mathcal{S}$, and let $\widetilde{\mathcal{S}} \subset \mathcal{S}/B$ be the set of states that are accessible from A without passing through B. Define $\widetilde{\mathcal{T}}$ to be the $|\widetilde{\mathcal{S}}| \times |\widetilde{\mathcal{S}}|$ substochastic matrix, which is the restriction of the original matrix $\mathcal{T}$ to the states in $\widetilde{\mathcal{S}}$:*

$$\widetilde{\mathcal{T}}_{ij} = \mathcal{T}_{ij} \ , \ i, j \in \widetilde{\mathcal{S}}$$

Also, let α be the spectral radius of $\widetilde{\mathcal{T}}$, and let T be the first hitting time for the set B:

$$T \equiv \inf\{t : S_t \in B\}$$

Then:

1. *For any initial distribution ρ on $\mathcal{S}$ with support A* [5]*:*

$$\lim_{t \to \infty} \mathbb{P}_\rho(T > t)^{1/t} = \alpha$$

2. *For any $A' \subset A$ and initial distribution ρ with support A':*

$$\limsup_{t \to \infty} \ \mathbb{P}_\rho(T > t)^{1/t} \leq \alpha$$

[5] That is, $\rho_i > 0$ if and only if $i \in A$.

3. If B is accessible from each $i \in \widetilde{\mathcal{S}}$ then:

$$\alpha < 1$$

Remark. *Note, in particular, that if A is the set of transient states of a Markov model and B is the set of recurrent states then $\widetilde{\mathcal{S}} = A$, and $\alpha < 1$ by part 3 of the lemma, since some recurrent state must be accessible from each transient state. Moreover, in this case, we have by part 2 of the lemma that for each transient state i:*

$$\limsup_{t \to \infty} \mathbb{P}_i(T > t)^{1/t} \leq \alpha$$

where T is the hitting time for the set of recurrent states B. This is probably, in general, the most useful application of this lemma for Markov chains, and it is the way the lemma will be applied in Section 5 below. However, in Section 4 we will need to use it in the more general form as stated above.

Proof of 1. By mutual induction on t it is easily seen that for each $t \in \mathbb{N}$:

(i) $\mathbb{P}_\rho(S_t = j | T > t) = \frac{(\widetilde{\rho}\widetilde{\mathcal{T}}^t)_j}{\|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1}$, for each $j \in \widetilde{\mathcal{S}}$, and

(ii) $\mathbb{P}_\rho(T > t) = \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1$,

where $\widetilde{\rho}$ is the distribution on $\widetilde{\mathcal{S}}$ defined by $\widetilde{\rho}_i = \rho_i, i \in \widetilde{\mathcal{S}}$. Thus, we must show that:

$$\lim_{t \to \infty} \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1^{1/t} = \alpha$$

Now, of course, we have:

$$\lim_{t \to \infty} \|\widetilde{\mathcal{T}}^t\|_1^{1/t} = \alpha$$

where:

$$\|\widetilde{\mathcal{T}}^t\|_1 \equiv \max\{\|v\widetilde{\mathcal{T}}^t\|_1 : \|v\|_1 = 1\}$$

is the operator norm of $\widetilde{\mathcal{T}}^t$ with respect to $\|\cdot\|_1$ on vectors. So, the upper bound is immediate:

$$\limsup_{t\to\infty} \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1^{1/t} \leq \limsup_{t\to\infty} \|\widetilde{\mathcal{T}}^t\|_1^{1/t} = \alpha$$

To prove the lower bound let us a fix a normalized eigenvector v of $\widetilde{\mathcal{T}}$ whose associated eigenvalue a is of maximum modulus. That is:

$$v\widetilde{\mathcal{T}} = av \ , \ \|v\|_1 = 1 \ , \text{ and } |a| = \alpha$$

Also, let u be the vector defined by $u_i = |v_i|$, $i \in \widetilde{\mathcal{S}}$. Then, since $\widetilde{\mathcal{T}}^t$ has nonnegative entries, we have by the triangle inequality:

$$\|u\widetilde{\mathcal{T}}^t\|_1 \geq \|v\widetilde{\mathcal{T}}^t\|_1 = \|a^t \cdot v\|_1 = |a|^t \cdot \|v\|_1 = \alpha^t$$

and since $\widetilde{\mathcal{T}}^t$ and u both have nonnegative entries:

$$\|u\widetilde{\mathcal{T}}^t\|_1 = \sum_{i\in\widetilde{\mathcal{S}}} u_i \|e_i\widetilde{\mathcal{T}}^t\|_1$$

Therefore, since each $u_i \leq 1$, we know that for each $t \in \mathbb{N}$ there is some $j = j(t) \in \widetilde{\mathcal{S}}$ with:

$$\|e_j\widetilde{\mathcal{T}}^t\|_1 \geq \alpha^t / |\widetilde{\mathcal{S}}|$$

Now, by assumption, the distribution $\widetilde{\rho}$ has support A and each state $j \in \widetilde{\mathcal{S}}$ is accessible from A without passing through B. So, for each $j \in \widetilde{\mathcal{S}}$ there must be some $m_j \in \mathbb{N}$ such that:

$$\left(\widetilde{\rho}\widetilde{\mathcal{T}}^{m_j}\right)_j \equiv p_j > 0$$

Let:

$$p \equiv \min_{j\in\widetilde{\mathcal{S}}} p_j \qquad \text{and} \qquad m \equiv \max_{j\in\widetilde{\mathcal{S}}} m_j$$

Then for any $t > m$ we have:

$$\begin{aligned}
\|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1 &\overset{(a)}{=} \|\widetilde{\rho}\widetilde{\mathcal{T}}^{m_j}\widetilde{\mathcal{T}}^{t-m_j}\|_1 \\
&\overset{(b)}{\geq} \left\|\left((\widetilde{\rho}\widetilde{\mathcal{T}}^{m_j})_j\ e_j\right)\widetilde{\mathcal{T}}^{t-m_j}\right\|_1 \\
&= \|p_j e_j \widetilde{\mathcal{T}}^{t-m_j}\|_1 \\
&\overset{(c)}{\geq} p_j \cdot \|e_j\widetilde{\mathcal{T}}^{t-m_j}\widetilde{\mathcal{T}}^{m_j}\|_1 \\
&\geq p \cdot \|e_j\widetilde{\mathcal{T}}^t\|_1 \\
&\overset{(d)}{\geq} p \cdot \left(\alpha^t/|\widetilde{\mathcal{S}}|\right) \qquad (14)
\end{aligned}$$

where $j = j(t)$. The decomposition in Step (a) is possible since $t > m \geq m_j$. The inequality in Step (b) holds since all vectors and matrices on both sides of the inequality have nonnegative entries. The inequality in Step (c) holds because $\widetilde{\mathcal{T}}$ is substochastic, so $\|\widetilde{\mathcal{T}}\|_1$, and hence $\|\widetilde{\mathcal{T}}^{m_j}\|_1$, are both at most 1. Finally, the inequality in Step (d) holds because of our choice $j = j(t)$.

Since p and $|\widetilde{\mathcal{S}}|$ are both positive constants independent of t it follows from the inequality (14) that:

$$\liminf_{t\to\infty} \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1^{1/t} \geq \alpha$$

which proves the claim.

□

Proof of 2. As in the previous case, it is easily seen by induction on t that:

$$\mathbb{P}_\rho(T > t) = \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1\ , \quad \text{for each } t \in \mathbb{N}$$

where $\widetilde{\rho}$ is the distribution on $\widetilde{\mathcal{S}}$ defined by $\widetilde{\rho}_i = \rho_i, i \in \widetilde{\mathcal{S}}$. Thus:

$$\limsup_{t\to\infty} \mathbb{P}_\rho(T > t)^{1/t} = \limsup_{t\to\infty} \|\widetilde{\rho}\widetilde{\mathcal{T}}^t\|_1^{1/t} \leq \limsup_{t\to\infty} \|\widetilde{\mathcal{T}}^t\|_1^{1/t} = \alpha$$

□

Proof of 3. If ρ is any distribution with support A, then by part 1 of this lemma:

$$\lim_{t\to\infty} \mathbb{P}_\rho(T > t)^{1/t} = \alpha \tag{15}$$

On the other hand, for any such distribution ρ:

$$\begin{aligned}
\mathbb{P}_\rho\left(T > |\widetilde{\mathcal{S}}| \cdot t\right) &= \prod_{n=1}^{t} \mathbb{P}_\rho\left(T > |\widetilde{\mathcal{S}}| \cdot n \;\middle|\; T > |\widetilde{\mathcal{S}}| \cdot (n-1)\right) \\
&\le \prod_{n=1}^{t} \max_{i\in\widetilde{\mathcal{S}}} \; \mathbb{P}_i(T > |\widetilde{\mathcal{S}}|) \\
&= q^t
\end{aligned} \tag{16}$$

for all $t \in \mathbb{N}$, where $q \equiv \max_{i\in\widetilde{\mathcal{S}}} \; \mathbb{P}_i(T > |\widetilde{\mathcal{S}}|)$. Together (15) and (16) imply:

$$\alpha \le q^{1/|\widetilde{\mathcal{S}}|}$$

The claim follows since $q < 1$ if B is accessible from each state $i \in \widetilde{\mathcal{S}}$. (Any state $i \in \widetilde{\mathcal{S}}$ that can reach B must be able to do so by a path of length at most $|\widetilde{\mathcal{S}}|$ steps.) □

3.3 Block Models

Definition 7. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible HMM. The n-*block model *(or simply* block model*) M^n is the HMM $(\mathcal{S}, \mathcal{W}, \{\mathcal{B}^{(w)}\})$ where $\mathcal{W} = \mathcal{L}_n(M)$ is the set of length-n allowable words for M and the block transition matrices $\mathcal{B}^{(w)}$ are defined by:*

$$\mathcal{B}^{(w)} \equiv \prod_{t=0}^{n-1} \mathcal{T}^{(x_t)} \ , \ w = x_0^{n-1} \in \mathcal{L}_n(M)$$

Note that the state transition matrix $\mathcal{B} \equiv \sum_{w\in\mathcal{W}} \mathcal{B}^{(w)}$ for the block model is the same state transition matrix $\mathcal{B}$ as for the n-step Markov model $(\mathcal{S}, \mathcal{B})$ defined in Section 2.4.3, and that for each $i, j \in \mathcal{S}$ and $w \in \mathcal{L}_n(M)$, $\mathcal{B}_{ij}^{(w)} = \mathbb{P}_i(X_0^{n-1} = w, S_n = j)$. Also, note that in the case $n = 1$, the block model M^1 is simply the original HMM M.

An example of this construction is presented below in Figure 3.

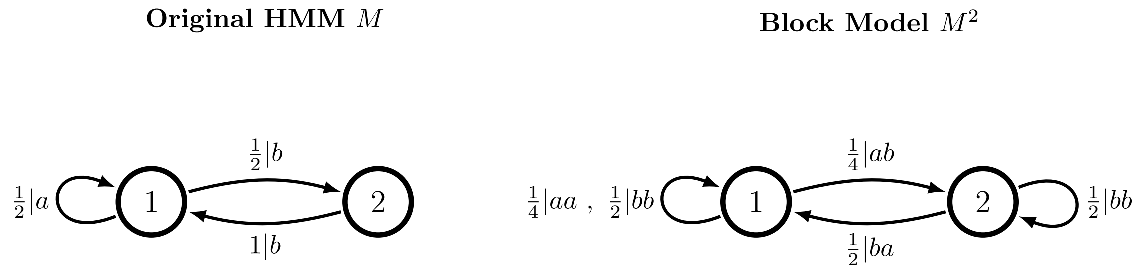


Figure 3: The block model construction M^2 for the Even Machine M of Figure 2.

The following lemma states some of the basic properties of block models. Properties 1 and 3 are immediate from the definition. Property 2 follows from the Chinese remainder theorem.

Lemma 7. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible HMM, and let $M^n = (\mathcal{S}, \mathcal{W}, \{\mathcal{B}^{(w)}\})$ be the corresponding n-block model.*

1. *The stationary distribution π for M is also stationary for M^n, i.e. $\pi = \pi\mathcal{B}$.*

2. *If n is relatively prime to $p = per(M)$, then M^n is also irreducible. Thus, in this case, π is the unique stationary distribution for M^n.*

3. *For any initial distribution ρ on $\mathcal{S}$, $w_0^{t-1} \in \mathcal{W}^t$, and $b_0^t \in \mathcal{S}^{t+1}$:*

$$\mathbb{P}_\rho^n(W_0^{t-1} = w_0^{t-1}, B_0^t = b_0^t) = \mathbb{P}_\rho(X_0^{nt-1} = w_0^{t-1} \textit{ and } S_{n\tau} = b_\tau, \tau = 0, 1, ..., t)$$

   where B_t and W_t are, respectively, the state and output symbol random variables for the block model at time t, and $\mathbb{P}_\rho^n$ is the distribution of these random variables when the initial state B_0 of the block model M^n is chosen according to the distribution ρ.

Property 3 is the primary reason for introducing block models. Since the joint process (B_t, W_t) generated by the block model M^n is simply a scaled version of the joint process (S_t, X_t) of the original HMM M, generated over length-n blocks of output symbols, questions about the original HMM may often be addressed by studying the block model rather than the original HMM directly. This is useful since the block model may have nicer (or stronger) properties than the original HMM, which simplify the analysis. In particular:

(i) If M is an irreducible, PD HMM and $\ell, n \in \mathbb{N}$ are defined by:

$$\begin{aligned} \ell &\equiv \max_{i \neq j} \ \min\{m : \mathbb{P}_i(X_0^{m-1}) \neq \mathbb{P}_j(X_0^{m-1})\} \\ n &\equiv \min\{m \geq \ell : m \text{ is relatively prime to } p = per(M)\} \end{aligned} \tag{17}$$

then the block model M^n is an irreducible, PD1 HMM.

(ii) If M is an irreducible, FS HMM and $\ell, n \in \mathbb{N}$ are defined by:

$$\begin{aligned} \ell &\equiv \max_{j \in \mathcal{S}} \ |w_j| \ , \quad \text{where } w_j \text{ is the flag word for state } j \\ n &\equiv \min\{m \geq \ell : m \text{ is relatively prime to } p = per(M)\} \end{aligned} \tag{18}$$

then the block model M^n is an irreducible, FS1 HMM (by point (iii) in Section 2.5.4).

Of course, the block model also always retains most essential properties. In particular:

- If M is exact, then so is M^n for any $n \in \mathbb{N}$.
- If M is unifilar, then so is M^n for any $n \in \mathbb{N}$.

The following lemma relates the speed of convergence of the entropy rate estimates and the decay of the expected state uncertainty for the block model to that of the original model. It will be essential for our proofs in Sections 5 and 6 below.

Lemma 8. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible HMM with associated block model $M^n = (\mathcal{S}, \mathcal{W}, \{\mathcal{B}^{(w)}\})$, for some $n \geq 2$ and relatively prime to $per(M)$. Let $(S_t, X_t)_{t \in \mathbb{Z}}$ and $(B_t, W_t)_{t \in \mathbb{Z}}$ be, respectively, the stationary joint state-symbol processes for M and M^n. Also, let h, $h(t)$, and $U_t = H(\phi(X_0^{t-1}))$ be the entropy rate, length-t entropy rate estimate, and state uncertainty at time t for the original HMM M, and let g, $g(t)$, and $V_t = H(\phi(W_0^{t-1}))$ be the entropy rate, length-t entropy rate estimate, and state uncertainty at time t for the block model M^n. Then:*

1. $g = nh$
2. $\limsup_{t \to \infty} (h(t) - h)^{1/t} = \left[\limsup_{t \to \infty} (g(t) - g)^{1/t}\right]^{1/n}$
3. $\limsup_{t \to \infty} \mathbb{E}(U_t)^{1/t} = \left[\limsup_{t \to \infty} \mathbb{E}(V_t)^{1/t}\right]^{1/n}$

Proof. The proofs of all three claims are based on applications of part 3 of Lemma 7. We will denote all such applications by (*).

The proof of Point 1 is a direct computation:

$$g = \lim_{t\to\infty} \frac{H(W_0^{t-1})}{t} \overset{(*)}{=} \lim_{t\to\infty} \frac{H(X_0^{nt-1})}{t} = \lim_{t\to\infty} \frac{H(X_0^{nt-1})}{nt} \cdot n = nh$$

Point 3 follows from the relation:

$$\mathbb{E}(U_{nt}) \overset{(*)}{=} \mathbb{E}(V_t)$$

and the fact that:

$$\mathbb{E}(U_t) = H(S_t|X_0^{t-1}) = H(S_0|X_{-t}^{-1})$$

is monotonically decreasing. Point 2 follows from the relation:

$$\begin{aligned} g(t+1) - g &= H(W_t|W_0^{t-1}) - nh \\ &\overset{(*)}{=} H(X_{nt}^{n(t+1)-1}|X_0^{nt-1}) - nh \\ &= \sum_{\tau=nt}^{n(t+1)-1} H(X_\tau|X_0^{\tau-1}) - nh \\ &= \sum_{\tau=nt}^{n(t+1)-1} (h(\tau+1) - h) \end{aligned}$$

and the fact that:

$$h(t) = H(X_t|X_1^{t-1}) = H(X_0|X_{-(t-1)}^{-1})$$

is monotonically decreasing. □

4 Results for Exact HMMs

In this section we prove an exponential upper bound for the speed of convergence of the entropy rate estimates $h(t)$ in exact HMMs. This is an extension of our earlier work with Crutchfield [26], wherein similar methods were used to derive an analogous result, but under the additional assumption of unifilarity.

Throughout this section we will denote by SYN_t the set of length-t words in the process language $\mathcal{L}(M)$ containing a synchronizing word, and by $NSYN_t$ the set of length-t words in the process language that do not contain a synchronizing word:

$$SYN_t \equiv \{w \in \mathcal{L}_t(M) : \text{ some subword of } w \text{ is a sync word for } M\}$$
$$NSYN_t \equiv \{w \in \mathcal{L}_t(M) : \text{ no subword of } w \text{ is a sync word for } M\}$$

The basic structure of the arguments is as follows:

- First, we introduce an auxilliary HMM known as the possibility machine, which allows us to easily compute the probability $\mathbb{P}(NSYN_t)$ that the first t output symbols do not contain a synchronizing word and how it scales.
- Then, we prove an upper bound on the differences $h(t+1) - h$ in terms of $\mathbb{P}(NSYN_t)$, which guarantees that the error in the block estimates $h(t)$ decays to 0 at least as fast as $\mathbb{P}(NSYN_t)$.

In some simple examples where the speed of convergence of $h(t)$ can actually be computed analytically[6] we know that $\mathbb{P}(NSYN_t)$ and the differences $h(t) - h$, in fact, decay at the same rate, so our upper bound on the rate of convergence of the block estimates $h(t)$ is tight.We believe this will often be the case, though we are not usually able to compute the actual speed of convergence of the estimates $h(t)$ by other means to verify this claim[7].

[6] For instance, the Even Machine of Figure 2 and Random Random XOR Machine [30].

[7] And, indeed, one can construct examples where the true speed of convergence is strictly faster than our upper bound by taking some non-zero transition probabilities very close to zero.

4.1 The Possibility Machine

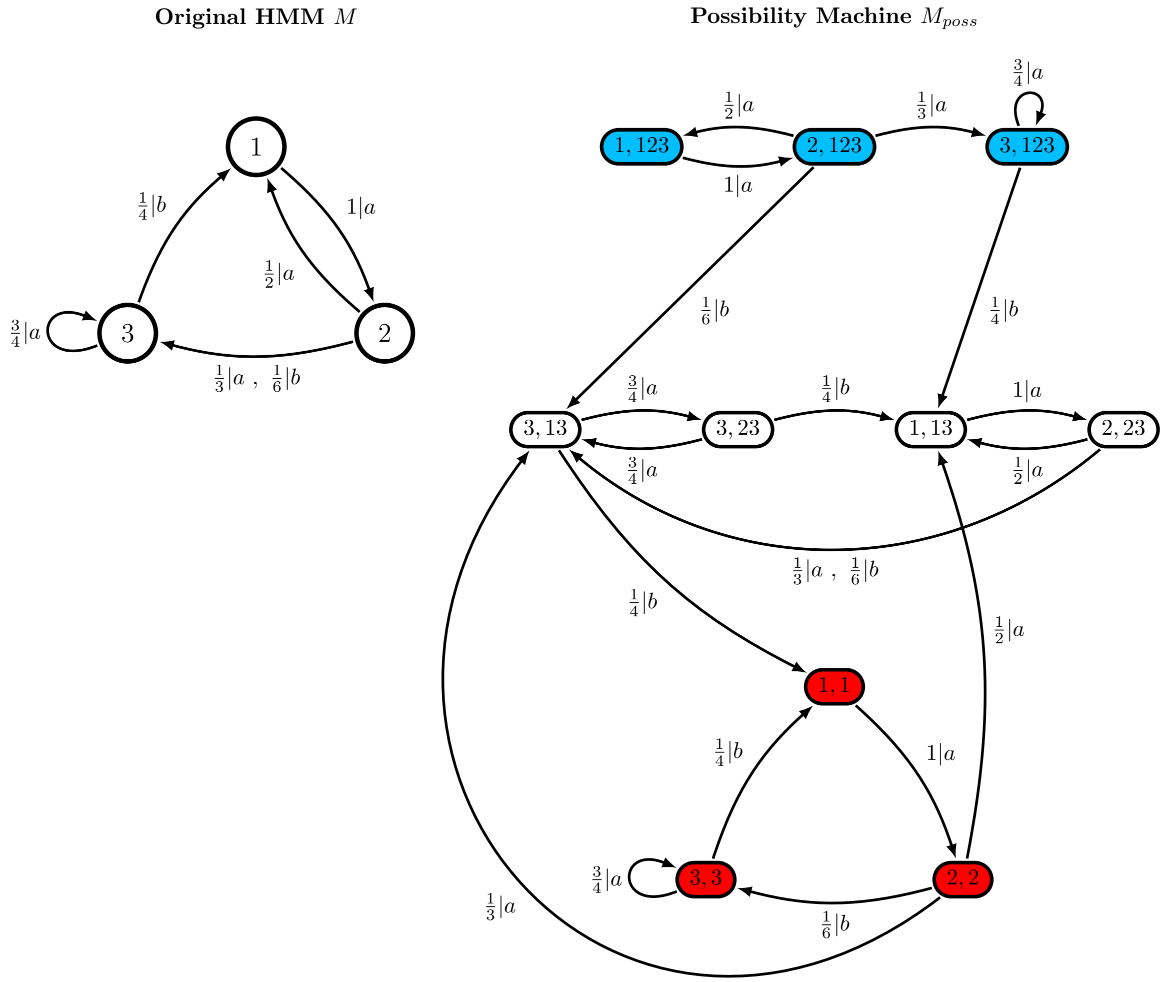


Figure 4: The possibility machine construction for an exact, 3 state, 2 symbol HMM. In the possibility machine initial states are colored blue and sync states are colored red. All other states are white. States names have also been abbreviated. So, for example, the state $(1, \{1, 2, 3\})$ is denoted simply by $1, 123$.

Definition 8. *For a HMM* $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ *the* extended possibility machine $\overline{M}_{poss}$ *is the HMM* $(\overline{\mathcal{R}}, \mathcal{X}, \{\overline{\mathcal{Q}}^{(x)}\})$ *where:*

- $\overline{\mathcal{R}} = \{(i, A) : i \in \mathcal{S}, A \text{ is a subset of } \mathcal{S}\}$
- *The transition matrices* $\overline{\mathcal{Q}}^{(x)}$ *are defined by:*

$$\overline{\mathcal{Q}}^{(x)}_{(i,A)(j,B)} = \mathcal{T}^{(x)}_{ij} \cdot \mathbb{1}\{\delta_A(x) = B\}$$

A state $(i, A) \in \overline{\mathcal{R}}$ *is said to be a* sync state *if* $A = \{i\}$ *and an* initial state *if* $A = \mathcal{S}$. *The* possibility machine M_{poss} *is the restriction of* $\overline{M}_{poss}$ *to its initial states and other states accessible from them. More precisely,* $M_{poss} = (\mathcal{R}, \mathcal{X}, \{\mathcal{Q}^{(x)}\})$ *where:*

- $\mathcal{R} = \{r \in \overline{\mathcal{R}} : r \textit{ is initial or is accessible from some initial state } r'\}$
- *The transition matrices* $\mathcal{Q}^{(x)}$ *are defined by:*

$$\mathcal{Q}^{(x)}_{rr'} = \overline{\mathcal{Q}}^{(x)}_{rr'} \ , \ r, r' \in \mathcal{R}$$

An example of this construction is given in Figure 4. We will denote the output sequence of the possibility machine by $(Y_t)_{t \in \mathbb{Z}^+}$ and its internal state sequence as $(R_t)_{t \in \mathbb{Z}^+} = (T_t, P_t)_{t \in \mathbb{Z}^+}$, where $T_t \in \mathcal{S}$ and P_t is a subset of $\mathcal{S}$. The interpretation is as follows:

- Y_t is the t_{th} output symbol of the original HMM M.
- T_t is the true state of the HMM M at time t.
- P_t is the set of states that the observer believes it is possible for the HMM M to be in after observing the output Y_0^{t-1}.

Initially, all states are possible to the observer ($S_0 \sim \pi$), so the initial states of the possibility machine are those of the form $(i, \mathcal{S})$. On the other hand, the observer becomes synchronized when the only possible state is the true state, so the sync states are those of the form $(i, \{i\})$.

The following lemma justifies and formalizes this interpretation.

Lemma 9. *Let* $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ *be an irreducible HMM with stationary distribution* π. *Let* $M_{poss} = (\mathcal{R}, \mathcal{X}, \{\mathcal{Q}^{(x)}\})$ *be the associated possibility machine, and let* μ *be the distribution on its state set* $\mathcal{R}$ *defined by:*

$$\mu_{(i,A)} = \begin{cases} \pi_i, \ , \ \textit{if } A = \mathcal{S} \\ 0 \ , \ \textit{else} \end{cases}$$

Then the distribution over joint sequences $(T_t, Y_t)_{t \in \mathbb{Z}^+}$ *with the initial possibility machine state* R_0 *chosen according to* μ *is the same as the stationary distribution over state-symbol sequences*

$(S_t, X_t)_{t \in \mathbb{Z}^+}$ *for the original HMM M given by choosing S_0 according to π:*

$$\mathbb{P}^{poss}_{\mu}(T_0^\infty, Y_0^\infty) = \mathbb{P}_\pi(S_0^\infty, X_0^\infty) \equiv \mathbb{P}(S_0^\infty, X_0^\infty) \tag{19}$$

Moreover, if $R_0 \sim \mu$ then (with probability 1):

$$\delta_{\mathcal{S}}(Y_0^{t-1}) = P_t \ , \ \textit{for each } t \in \mathbb{N} \tag{20}$$

In particular, Y_0^{t-1} is a synchronizing word for M if and only if the set of possible states P_t consists only of the true state T_t [8]*. Or, equivalently, if and only if R_t is a sync state.*

$$Y_0^{t-1} \textit{ is a sync word } \iff P_t = \{T_t\} \iff R_t \textit{ is a sync state}$$

Remark. *Note, in particular, that this lemma implies the original HMM M is exact (has sync words) if and only if the possibility machine M_{poss} has sync states. So, the possibility machine construction is a way of testing a HMM for exactness.*

Proof of (20). For any $w = x_0^{t-1} \in \mathcal{X}^*$ it is easily seen (by induction on t) that:

$$\delta_{\mathcal{S}}(w) = \mathcal{S}_t$$

where the sets $\mathcal{S}_0, ..., \mathcal{S}_t$ are defined by the relations:

$$\mathcal{S}_0 = \mathcal{S} \text{ and } \mathcal{S}_{n+1} = \delta_{\mathcal{S}_n}(x_n) \ , \ n = 0, ..., t-1$$

The claim follows from this fact, since the initial set of possible states P_0 is always $\mathcal{S}$ if R_0 is chosen according to μ, and by construction of the possibility machine:

$$P_{n+1} = \delta_{P_n}(Y_n)$$

□

[8] Note that by the first claim (19), we must have $T_t \in \delta_{T_0}(Y_0^{t-1})$. Thus, by the second claim (20), $T_t \in P_t = \delta_{\mathcal{S}}(Y_0^{t-1})$, so it is impossible that the output sequence Y_0^{t-1} synchronizes an observer of the original HMM M to any state other than the T_t.

Proof of (19). We show by induction on t that:

$$\mathbb{P}_{\mu}^{poss}(T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) = \mathbb{P}(S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1})$$

for all $t \geq 0$ and $s_0^t \in \mathcal{S}^{t+1}, x_0^{t-1} \in \mathcal{X}^t$.

Basis Step - Show for $t = 0$.

For each $s_o \in \mathcal{S}$ we have:

$$\mathbb{P}_{\mu}^{poss}(T_0 = s_0) = \mu_{(s_o,\mathcal{S})} = \pi_{s_o} = \mathbb{P}(S_0 = s_o)$$

Inductive Step - Assume for t, and show for $t+1$.

If $\mathbb{P}(S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1}) = \mathbb{P}_{\mu}^{poss}(T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) = 0$, then clearly also $\mathbb{P}(S_0^{t+1} = s_0^{t+1}, X_0^t = x_0^t) = \mathbb{P}_{\mu}^{poss}(T_0^{t+1} = s_0^{t+1}, Y_0^t = x_0^t) = 0$. Thus, let us assume $\mathbb{P}(S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1}) = \mathbb{P}_{\mu}^{poss}(T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) > 0$. In this case, we may condition on the event $\{T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}\}$, and we find:

$$\begin{aligned}
&\mathbb{P}_{\mu}^{poss}(T_{t+1} = s_{t+1}, Y_t = x_t | T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) \\
&\quad \overset{(*)}{=} \mathbb{P}_{\mu}^{poss}(T_{t+1} = s_{t+1}, Y_t = x_t | T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}, P_t = A) \\
&\quad = \mathbb{P}_{\mu}^{poss}(T_{t+1} = s_{t+1}, Y_t = x_t | R_t = (s_t, A)) \\
&\quad = \sum_{B \subseteq \mathcal{S}} \mathbb{P}_{\mu}^{poss}(R_{t+1} = (s_{t+1}, B), Y_t = x_t | R_t = (s_t, A)) \\
&\quad = \sum_{B \subseteq \mathcal{S}} \mathcal{T}_{s_t s_{t+1}}^{(x_t)} \cdot \mathbb{1}\{\delta_A(x_t) = B\} \\
&\quad = \mathcal{T}_{s_t s_{t+1}}^{(x_t)} \\
&\quad = \mathbb{P}(S_{t+1} = s_{t+1}, X_t = x_t | S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1})
\end{aligned} \tag{21}$$

where $A \equiv \delta_{\mathcal{S}}(x_0^{t-1})$ and (*) follows from the relation (20) proved above, which guarantees that $P_t = A$ with probability 1 on the event $\{T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}\}$. Combining (21) with the inductive hypothesis yields:

$$\begin{aligned}
&\mathbb{P}_{\mu}^{poss}(T_0^{t+1} = s_0^{t+1}, Y_0^t = x_0^t) \\
&\quad = \mathbb{P}_{\mu}^{poss}(T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) \cdot \mathbb{P}_{\mu}^{poss}(T_{t+1} = s_{t+1}, Y_t = x_t | T_0^t = s_0^t, Y_0^{t-1} = x_0^{t-1}) \\
&\quad = \mathbb{P}(S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1}) \cdot \mathbb{P}(S_{t+1} = s_{t+1}, X_t = x_t | S_0^t = s_0^t, X_0^{t-1} = x_0^{t-1}) \\
&\quad = \mathbb{P}(S_0^{t+1} = s_0^{t+1}, X_0^t = x_0^t)
\end{aligned}$$

proving the claim. □

4.2 Scaling of $\mathbb{P}(NSYN_t)$

Combining Lemma 9 with Lemma 6 we now obtain the scaling rate of $\mathbb{P}(NSYN_t)$.

Proposition 1. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an exact, irreducible HMM with associated possibility machine $M_{poss} = (\mathcal{R}, \mathcal{X}, \{\mathcal{Q}^{(x)}\})$. Let $\widetilde{\mathcal{R}} \subset \mathcal{R}$ be the set of possibility machine states, which are accessible from the set of initial states without passing through the set of sync states (See Section 3.2 for definitions). Finally, let $\widetilde{\mathcal{Q}}$ be the restriction of the state transition matrix $\mathcal{Q} = \sum_x \mathcal{Q}^{(x)}$ for M_{poss} to the state set $\widetilde{\mathcal{R}}$:*

$$\widetilde{\mathcal{Q}}_{rr'} = \mathcal{Q}_{rr'} \ , \ r, r' \in \widetilde{\mathcal{R}}$$

Then:

$$\lim_{t \to \infty} \mathbb{P}(NSYN_t)^{1/t} = \alpha$$

where $0 \leq \alpha < 1$ is the spectral radius of $\widetilde{\mathcal{Q}}$.

Proof. By Lemma 9 we know that if $R_0 \sim \mu$ then:

(i) The distribution of the output sequence Y_0^{t-1} for the possibility machine is the same as the distribution of the output sequence X_0^{t-1} for the original HMM M (with $S_0 \sim \pi$).

(ii) Y_0^{t-1} is a sync word if and only if R_t is a sync state.

Thus:

$$\begin{aligned}\mathbb{P}(NSYN_t) &= \mathbb{P}(X_0^{t-1} \notin SYN_t)\\ &= \mathbb{P}(X_0^{n-1} \text{ is not a sync word, for all } 1 \leq n \leq t)\\ &\overset{(i)}{=} \mathbb{P}_\mu^{poss}(Y_0^{n-1} \text{ is not a sync word, for all } 1 \leq n \leq t)\\ &\overset{(ii)}{=} \mathbb{P}_\mu^{poss}(R_n \text{ is not a sync state, for all } 1 \leq n \leq t)\\ &= \mathbb{P}_\mu^{poss}(T > t)\end{aligned}$$

where $T \equiv \inf\{t : R_t \text{ is a sync state}\}$ is the first hitting time for the set of sync states.

Therefore, by part 1 of Lemma 6, we have:

$$\lim_{t \to \infty} \mathbb{P}(NSYN_t)^{1/t} = \alpha$$

Moreover, by part 3 of this same lemma, we know α must be less than 1, since some sync state is accessible from each $r \in \widetilde{\mathcal{R}}$. In particular, since M is exact, we know that for each $i \in \mathcal{S}$ there is some word $w_i \in \mathcal{L}(i)$, which synchronizes the observer to some state $j_i \in \mathcal{S}$. So, for any possibility machine state $r = (i, A)$, $\delta_r^{poss}(w_i)$ is the singleton set consisting of the sync state $(j_i, \{j_i\})$.[9] □

Remark. *Looking at the proof of Lemma 6 it is clear that for any finite t we have the exact formula:*

$$\mathbb{P}(NSYN_t) = \mathbb{P}_\mu^{poss}(T > t) = \|\widetilde{\mu}\widetilde{\mathcal{Q}}^t\|_1$$

where $\widetilde{\mu}$ is the distribution on $\widetilde{\mathcal{R}}$ defined by $\widetilde{\mu}_r = \mu_r, r \in \widetilde{\mathcal{R}}$. For estimates with finite and relatively small t, so that the computation of $\widetilde{\mu}\widetilde{\mathcal{Q}}^t$ is not too computationally intensive, this may be more useful than the asymptotic scaling rate α.

The following upper bound on the decay of the expected state uncertainty $\mathbb{E}(U_t)$ in an exact, unifilar HMM is a simple consequence of Proposition 1. It is far easier to prove than analogous results derived in Section 5 for nonexact, unifilar HMMs.

[9] This follows from the more general statement $\delta_{(i,A)}^{poss}(w) = \{(j, \delta_A(w)) : j \in \delta_i(w)\}$, which is easily proved by induction on the length of w.

Proposition 2. *For an exact, unifilar, irreducible HMM M:*

$$\limsup_{t \to \infty} \ \mathbb{E}(U_t)^{1/t} \leq \alpha$$

where $0 \leq \alpha < 1$ is the spectral radius of the matrix $\widetilde{Q}$, as defined in Proposition 1.

Proof. If M is unifilar then extensions of synchronizing words are also synchronizing words so:

$$w \in SYN_t \Longrightarrow w \text{ is a sync word } \Longrightarrow u(w) = 0$$

Thus:

$$\mathbb{E}(U_t) = \sum_{w \in \mathcal{L}_t(M)} \mathbb{P}(w) u(w) \leq \mathbb{P}(NSYN_t) \cdot \log_2 |\mathcal{S}|$$

The claim follows since:

$$\lim_{t \to \infty} \mathbb{P}(NSYN_t)^{1/t} = \alpha$$

by Proposition 1. □

4.3 Sync Bound on Convergence of the Entropy Rate Estimates

Using the results of the previous section for the decay of $\mathbb{P}(NSYN_t)$ we now obtain an upper bound on the speed of convergence of the entropy rate block estimates $h(t)$. The key estimate is Lemma 11 below, which shows that the difference $h(t+1) - h$ is upper bounded by a constant times $\mathbb{P}(NSYN_t)$. Before proceeding to prove this estimate, however, we will first need to introduce one simple additional lemma concerning conditioning on infinite pasts, which contain synchronizing words.

Lemma 10. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an exact, irreducible HMM, and let w be a sync word for M. If $x_{-\infty}^{-1}$ is a regular past with $x_{-n}^{-m} = w$ for some $1 \leq m \leq n$ then:*

$$H(X_0 | x_{-t}^{-1}) = H(X_0 | x_{-\infty}^{-1}) \ , \ \textit{for all } t \geq n$$

Proof. Assume that w synchronizes the observer to state i. Then, for any $t \geq n$:

$$\mathbb{P}(S_{-m+1} = i | X_{-t}^{-1} = x_{-t}^{-1}) = \mathbb{P}(S_{-m+1} = i | X_{-n}^{-m} = x_{-n}^{-m}) = 1$$

So, for each $t \geq n$ and $x \in \mathcal{X}$:

$$\begin{aligned} \mathbb{P}(x|x_{-t}^{-1}) &\equiv \mathbb{P}(X_0 = x | X_{-t}^{-1} = x_{-t}^{-1}) \\ &= \mathbb{P}(X_0 = x | X_{-t}^{-1} = x_{-t}^{-1}, S_{-m+1} = i) \\ &= \mathbb{P}(X_0 = x | X_{-m+1}^{-1} = x_{-m+1}^{-1}, S_{-m+1} = i) \qquad (22) \end{aligned}$$

Since the right hand side of (22) is a constant independent of t, we know that for each x:

$$\begin{aligned} \mathbb{P}(x|x_{-\infty}^{-1}) &\equiv \lim_{t \to \infty} \mathbb{P}(x|x_{-t}^{-1}) \\ &= \mathbb{P}(x|x_{-t}^{-1}) \ , \text{ for each } t \geq n \end{aligned}$$

Thus, for each $t \geq n$ we have:

$$\begin{aligned} H(X_0|x_{-\infty}^{-1}) &\equiv -\sum_x \mathbb{P}(x|x_{-\infty}^{-1}) \log_2 \mathbb{P}(x|x_{-\infty}^{-1}) \\ &= -\sum_x \mathbb{P}(x|x_{-t}^{-1}) \log_2 \mathbb{P}(x|x_{-t}^{-1}) \\ &= H(X_0|x_{-t}^{-1}) \end{aligned}$$

□

Lemma 11. *For an exact, irreducible HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$:*

$$h(t+1) - h \leq \mathbb{P}(NSYN_t) \cdot \log_2 |\mathcal{X}| \ , \ \textit{for all } t \in \mathbb{N}$$

Proof. Let Z_w be the set of regular infinite pasts $x_{-\infty}^{-1}$ terminating in the word w:

$$Z_w \equiv \{x_{-\infty}^{-1} : x_{-\infty}^{-1} \text{ is regular and } x_{-|w|}^{-1} = w\}$$

Then, for any fixed $t \in \mathbb{N}$:

$$\begin{aligned} h(t+1) &\equiv H(X_{t+1}|X_1^t) \\ &= H(X_0|X_{-t}^{-1}) \\ &= \sum_{w \in \mathcal{L}_t(M)} H(X_0|X_{-t}^{-1} = w) \cdot \mathbb{P}(w) \\ &= \sum_{w \in \mathcal{L}_t(M)} \int_{Z_w} H(X_0|x_{-t}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) \end{aligned}$$

Thus, using the integral formula for the entropy rate (3) and the previous lemma we have:

$$\begin{aligned} h(t+1) - h &= \sum_{w \in \mathcal{L}_t(M)} \int_{Z_w} H(X_0|x_{-t}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) - \int H(X_0|x_{-\infty}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) \\ &= \sum_{w \in \mathcal{L}_t(M)} \int_{Z_w} H(X_0|x_{-t}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) - \sum_{w \in \mathcal{L}_t(M)} \int_{Z_w} H(X_0|x_{-\infty}^{-1}) d\mathbb{P}(x_{-\infty}^{-1}) \\ &= \begin{cases} \sum_{w \in NSYN_t} \int_{Z_w} \left(H(X_0|x_{-t}^{-1}) - H(X_0|x_{-\infty}^{-1}) \right) d\mathbb{P}(x_{-\infty}^{-1}) + \\ \sum_{w \in SYN_t} \int_{Z_w} \left(H(X_0|x_{-t}^{-1}) - H(X_0|x_{-\infty}^{-1}) \right) d\mathbb{P}(x_{-\infty}^{-1}) \end{cases} \\ &\leq \sum_{w \in NSYN_t} \mathbb{P}(w) \cdot \log_2 |\mathcal{X}| \ + \sum_{w \in SYN_t} \mathbb{P}(w) \cdot 0 \\ &= \mathbb{P}(NSYN_t) \cdot \log_2 |\mathcal{X}| \end{aligned}$$

□

Theorem 1. *For an exact, irreducible HMM M:*

$$\limsup_{t \to \infty} \{h(t) - h\}^{1/t} \leq \alpha$$

where $0 \leq \alpha < 1$ is the spectral radius of the matrix $\widetilde{Q}$, as defined in Proposition 1.

Proof. This follows immediately from the decay rate of $\mathbb{P}(NSYN_t)$ given in Proposition 1 and Lemma 11. □

5 Results for Unifilar HMMs

In this section we establish exponential bounds on the pointwise and L_1 decay of the state uncertainty U_t for PD, unifilar HMMs. Then, as a consequence of the L_1 bounds, we derive exponential bounds on the speed of convergence of the entropy rate block estimates $h(t)$, for general unifilar HMMs, by a means of a PD reduction.

In many places we will assume, also, that the HMM is nonexact, so we do not have to treat the exact case separately where the analysis is slightly different. In the case of an exact, unifilar HMM the pointwise decay rate of U_t is simply $\alpha = 0$, since $\mathbb{P}$ a.e. x_0^∞ contains a sync word. Also, in this case, we have already exponential bounds on the L_1 decay of state uncertainty and convergence of the entropy estimates given by Proposition 2 and Theorem 1. These bounds are at least as strong as what we would obtain using the alternate methods for unifilar HMMs given below.

Similar statements about exponential decay of the state uncertainty U_t and exponential convergence of the entropy estimates $h(t)$, for nonexact, unifilar HMMs were given in our earlier work with Crutchfield [27]. However, the methods of proof presented here are quite different and, we feel, intuitively somewhat nicer. The new proof methods also give the pointwise decay rate of U_t exactly, which the old methods do not.

The central idea behind the proofs is the construction of an auxiliary HMM known as the *follower model*. The basic concept was originally suggested to us by Morris [31], though the details are somewhat more involved than they may have initially appeared at the time, because the follower model is not always irreducible (even ignoring trivial or non-accessible components).

5.1 The Follower Model

For a unifilar HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ we define the extended state set $\overline{\mathcal{S}}$ and transition function $\overline{\delta}_i(w)$, $i \in \overline{\mathcal{S}}, w \in \mathcal{X}^*$ by:

$$\overline{\mathcal{S}} \equiv \mathcal{S} \cup \{d\}$$

and:

$$\overline{\delta}_i(w) \equiv \begin{cases} j \ , & \text{if } i \in \mathcal{S} \text{ and } \delta_i(w) = \{j\} \\ d \ , & \text{if } i \in \mathcal{S} \text{ and } \delta_i(w) = \emptyset \\ d \ , & \text{if } i = d \end{cases} \tag{23}$$

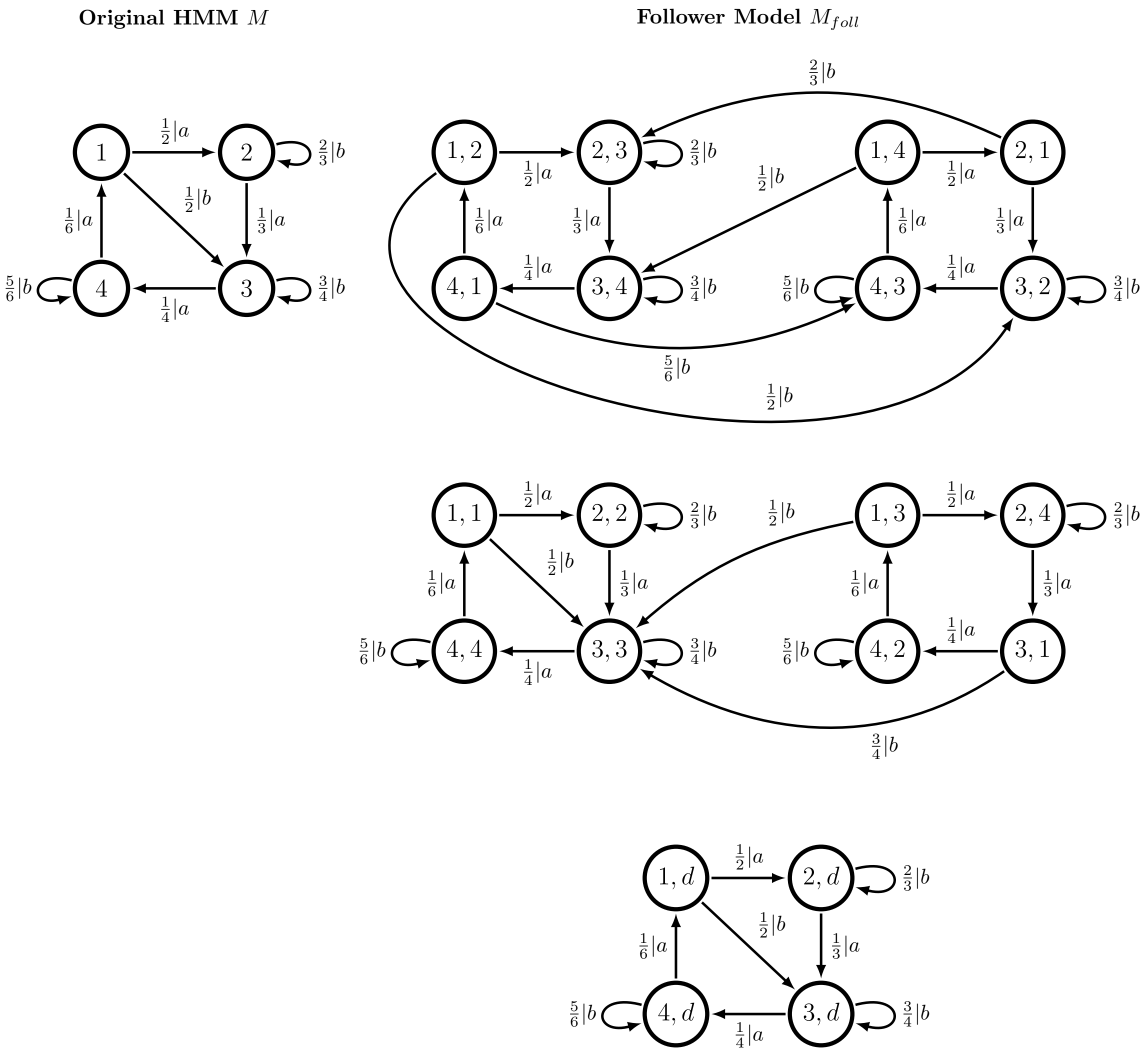


Figure 5: The follower model construction for a 4 state, 2 symbol, unifilar HMM. In the follower model parentheses have been omitted from the state labels. So, for example, the state $(2, 3)$ is denoted simply by $2, 3$.

where $\delta_i(w)$ is the standard HMM transition function given by (6). By unifilarity, for any $i \in \mathcal{S}$ and $w \in \mathcal{X}^*$, $\delta_i(w)$ is always either $\emptyset$ or $\{j\}$, for some $j \in \mathcal{S}$. Thus, $\overline{\delta}_i(w)$ is always well defined. We will refer to d as the *dead state*. The meaning of this terminology will become clear in the discussion below.

Definition 9. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible, unifilar HMM. The corresponding* follower model *M_{foll} is the unifilar HMM $(\mathcal{S} \times \overline{\mathcal{S}}, \mathcal{X}, \{\mathcal{Q}^{(x)}\})$ with transition matrices $\mathcal{Q}^{(x)}$ defined by:*

$$\mathcal{Q}^{(x)}_{(i,\bar{i})(j,\bar{j})} = \mathcal{T}^{(x)}_{ij} \cdot \mathbb{1}_{\{\delta_{\bar{i}}(x)=\bar{j}\}} \ , \ i, j \in \mathcal{S} \ \textit{and} \ \bar{i}, \bar{j} \in \overline{\mathcal{S}}$$

An example of this construction is presented in Figure 5. We will denote the output sequence of the follower model M_{foll} as $(Y_t)_{t \in \mathbb{Z}^+}$ and its internal state sequence as $(R_t, \overline{R}_t)_{t \in \mathbb{Z}^+}$, $R_t \in \mathcal{S}$ and $\overline{R}_t \in \overline{\mathcal{S}}$. The distribution of these random variables with initial state $(R_0, \overline{R}_0) = (i, \bar{i})$ is denoted by $\mathbb{P}^{foll}_{(i,\bar{i})}$.

R_t is thought of as the *leader state* at time t and $\overline{R}_t$ as the *follower state*. The next output symbol Y_t and next leader state R_{t+1} are selected from the current leader state R_t in the normal way. Then, the next follower state $\overline{R}_{t+1}$ is determined simply by following the outgoing edge in the original HMM M from state $\overline{R}_t$ labeled with symbol Y_t. If there is no such edge, then $\overline{R}_t$ transitions to the dead state d. The dead state also transitions to itself on all output symbols. Thus, once the follower path dies, it remains dead for all future time.

Formally, we will adopt the convention:

$$\mathbb{P}_d(w) \equiv 0 \ , \ w \in \mathcal{X}^*$$

for word probabilities from the dead state, so that for each $i \in \overline{\mathcal{S}}$ and $w, v \in \mathcal{X}^*$:

$$\mathbb{P}_i(wv) = \mathbb{P}_i(w) \cdot \mathbb{P}_{\bar{\delta}_i(w)}(v)$$

The following lemma (applying this convention) is a direct consequence of the follower model construction and, indeed, the primary reason for introducing it.

Lemma 12. *If M is an irreducible, uniflar HMM with associated follower model M_{foll} then:*

1. *For each $(i, \bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ the distribution over output sequences Y_0^∞ from state $(i, \bar{i})$ for M_{foll} is the same as the distribution over output sequences X_0^∞ from state i for M:*

$$\mathbb{P}^{foll}_{(i,\bar{i})}(Y_0^\infty) = \mathbb{P}_i(X_0^\infty)$$

2. *For each initial state $(i, \bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ and $t \in \mathbb{N}$ the following all hold with probability 1 with*

respect to the measure $\mathbb{P}^{foll}_{(i,\bar{i})}$ [10]*:*

$$\begin{aligned}
\overline{\delta}_i(Y_0^{t-1}) &= R_t \\
\overline{\delta}_{\bar{i}}(Y_0^{t-1}) &= \overline{R}_t \\
\mathbb{P}_i(Y_0^{t-1}) &= \prod_{n=0}^{t-1} \mathbb{P}_{R_n}(Y_n) \\
\mathbb{P}_{\bar{i}}(Y_0^{t-1}) &= \prod_{n=0}^{t-1} \mathbb{P}_{\overline{R}_n}(Y_n)
\end{aligned}$$

Note, in particular, that point 2 implies that for each $1 \leq \tau \leq t-1$:

$$\begin{aligned}
\mathbb{P}_i(Y_0^{t-1}) &= \mathbb{P}_i(Y_0^{\tau-1}) \cdot \mathbb{P}_{R_\tau}(Y_\tau^{t-1}) \ , \quad \text{and} \\
\mathbb{P}_{\bar{i}}(Y_0^{t-1}) &= \mathbb{P}_{\bar{i}}(Y_0^{\tau-1}) \cdot \mathbb{P}_{\overline{R}_\tau}(Y_\tau^{t-1}) \ .
\end{aligned}$$

5.1.1 Recurrent Decomposition

In general, the follower model is not irreducible. However, the associated graph always has a finite number of recurrent, strongly components[11] $\mathcal{C}_k, k \in K$, and each of these components itself defines an irreducible (unifilar) HMM M^k_{foll}. We will say a component $\mathcal{C}_k$, or the associated HMM M^k_{foll}, is *trivial* if either:

(a) $\bar{i} = i$, for some $(i, \bar{i})$ in $\mathcal{C}_k$, or

(b) $\bar{i} = d$, for some $(i, \bar{i})$ in $\mathcal{C}_k$

and nontrivial otherwise. Also, we will denote by $\mathcal{R}$ the set of all recurrent states for the follower HMM M_{foll}:

$$\mathcal{R} \equiv \bigcup_{k \in K} \bigcup_{(i,\bar{i}) \in \mathcal{C}_k} (i, \bar{i})$$

[10] Here, $\mathbb{P}_i(Y_0^{t-1})$ is the random variable (defined on the follower HMM space) such that $\mathbb{P}_i(Y_0^{t-1}) = \mathbb{P}_i(x_0^{t-1})$ on the event that the follower HMM output sequence Y_0^{t-1} is equal to the symbol sequence $x_0^{t-1} \in \mathcal{X}^t$. Similarly, $\mathbb{P}_{R_n}(Y_n)$ is the random variable such that $\mathbb{P}_{R_n}(Y_n) = \mathbb{P}_j(x)$ on the event $\{R_n = j, Y_n = x\}$. $\mathbb{P}_{\bar{i}}(Y_0^{t-1})$ and $\mathbb{P}_{\overline{R}_n}(Y_n)$ are defined analogously. $\overline{\delta}_i(Y_0^{t-1})$ is the $\overline{\mathcal{S}}$-valued random variable such that $\overline{\delta}_i(Y_0^{t-1}) = \overline{\delta}_i(x_0^{t-1})$ on the event $Y_0^{t-1} = x_0^{t-1}$, and $\overline{\delta}_{\bar{i}}(Y_0^{t-1})$ is the $\overline{\mathcal{S}}$-valued random variable such that $\overline{\delta}_{\bar{i}}(Y_0^{t-1}) = \overline{\delta}_{\bar{i}}(x_0^{t-1})$ on the event $Y_0^{t-1} = x_0^{t-1}$.

[11] A strongly connected component C of a directed graph G is said to be *recurrent* if there are no outgoing edges from vertices in C to vertices not in C.

An example of this recurrent decomposition is presented below in Figure 6.

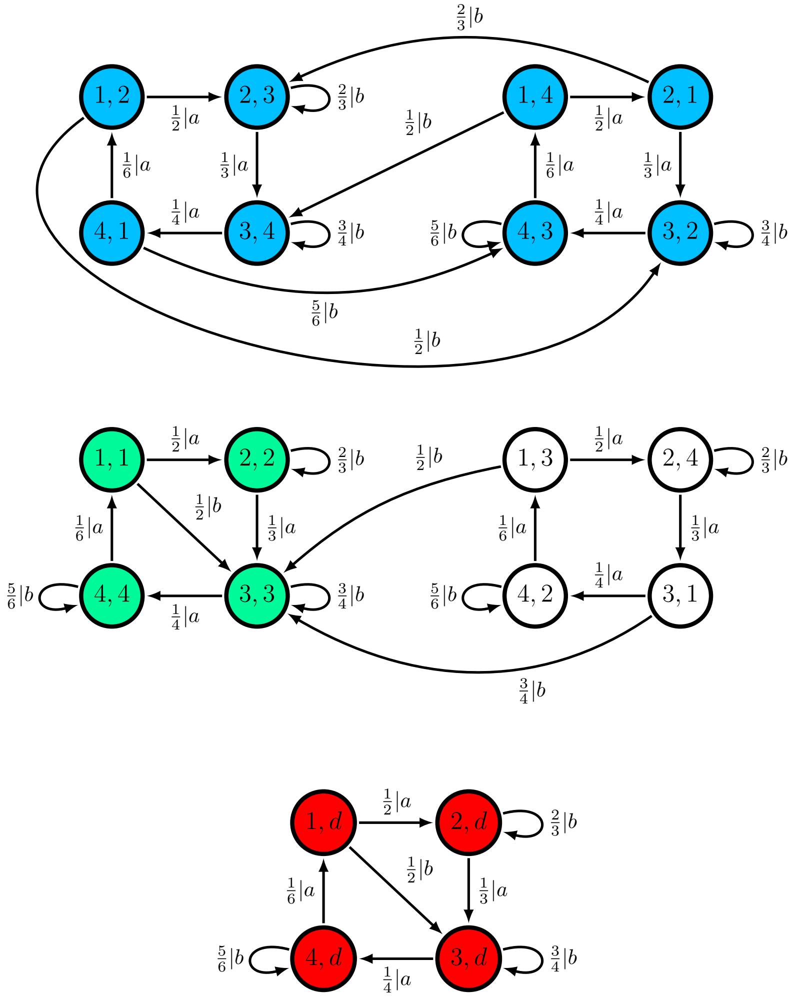


Figure 6: The recurrent decomposition for the follower model M_{foll} given in Figure 5. There are 3 recurrent components: The dead component colored in red, the matched component colored in green, and one nontrivial component colored in blue.

The following are some important properties of the recurrent decomposition, which are illustrated by the example above, but true more generally. All of them are direct consequences of the follower model construction and irreducibility of the original HMM M.

(i) There are always exactly two trivial components for M_{foll}: the *dead component* $\mathcal{C}_d$ consisting of all states of the form (i, d) such that $i \in \mathcal{S}$ and the *matched component* $\mathcal{C}_m$ consisting of all states of the form (i, i) such that $i \in \mathcal{S}$.

(ii) There may or may not also be nontrivial recurrent components. Indeed, there are nontrivial components if and only if M is nonexact.

(iii) If there are nontrivial components, then for each nontrivial component $\mathcal{C}_k$ and state $(i, \bar{i}) \in \mathcal{C}_k$, $\mathbb{P}_{\bar{i}}(x) > 0$ for any symbol x with $\mathbb{P}_i(x) > 0$. (Otherwise, $(i, \bar{i})$ would transition to some state in the dead component on symbol x.)

(iv) If there are nontrivial components, then for each nontrivial component $\mathcal{C}_k$ and state $i \in \mathcal{S}$ there exists some $\bar{i} \in \mathcal{S}$ (possibly more than 1) such that $(i, \bar{i}) \in \mathcal{C}_k$. (Indeed, if $(j, \bar{j}) \in \mathcal{C}_k$ for some nontrivial $\mathcal{C}_k$ and w is any word with $\overline{\delta}_j(w) = i$, then the follower model state $(j, \bar{j})$ transitions to $(i, \bar{i}) \in \mathcal{C}_k$ on w, where $\bar{i} = \overline{\delta}_{\bar{j}}(w)$.)

Also, note that, by Lemma 12, the following equivalences always hold with probability 1 with respect to the measure $\mathbb{P}^{foll}_{(i,\bar{i})}$, for each initial state $(i, \bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$:

$$\overline{\delta}_i(Y_0^{t-1}) = \overline{\delta}_{\bar{i}}(Y_0^{t-1}) \Longleftrightarrow (R_t, \overline{R}_t) \in \mathcal{C}_m$$
$$\mathbb{P}_{\bar{i}}(Y_0^{t-1}) = 0 \Longleftrightarrow (R_t, \overline{R}_t) \in \mathcal{C}_d$$

5.1.2 Follower Edge Chain

Since M_{foll} is unifilar each edge z of M_{foll} can be represented (uniquely) as a 3-tuple $z = (i, \bar{i}, x)$, where $(i, \bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ and x is a symbol with $\mathbb{P}_i(x) > 0$. $z = (i, \bar{i}, x)$ is the outgoing edge from state $(i, \bar{i})$ labeled with symbol x.

We will denote the edge set of M_{foll} as $\mathcal{Z}$ and its edge sequence by $(Z_t)_{t \in \mathbb{Z}^+}$:

$$Z_t \equiv (R_t, \overline{R}_t, Y_t)$$

Also, we will denote the edge set of a component HMM M^k_{foll} by $\mathcal{Z}_k$ and the stationary distribution for the irreducible edge chain (Z_t) associated to the component HMM M^k_{foll} by μ^k:

$$\mu^k_{(i,\bar{i},x)} = \nu^k_{(i,\bar{i})} \cdot \mathbb{P}_i(x) \ , \ \text{for } z = (i, \bar{i}, x) \in \mathcal{Z}_k$$

where ν^k is the stationary state distribution of M^k_{foll}.

5.2 A Useful Lemma

We give below a simple, but useful, lemma for the difference between the induced distributions $\phi(x_0^{t-1})$ and $\phi_i(x_0^{t-1})$ in a unifilar HMM. The lemma and its corollaries will be used in Sections 5.3 and 5.4 for establishing bounds on the decay rate of the state uncertainty U_t.

Lemma 13. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible, unifilar HMM. Then for each $i \in \mathcal{S}$ and $x_0^{t-1} \in \mathcal{L}(i)$:*

$$\|\phi(x_0^{t-1}) - \phi_i(x_0^{t-1})\|_{TV} = \sum_{j \in \mathcal{S}_i^c(x_0^{t-1})} \frac{\pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\mathbb{P}(x_0^{t-1})} = \frac{\sum_{j \in \mathcal{S}_i^c(x_0^{t-1})} \pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\sum_{j \in \mathcal{S}} \pi_j \cdot \mathbb{P}_j(x_0^{t-1})}$$

where:

$$\mathcal{S}_i(x_0^{t-1}) \equiv \{j \in \mathcal{S} : \overline{\delta}_j(x_0^{t-1}) = \overline{\delta}_i(x_0^{t-1})\} \quad \textit{and} \quad \mathcal{S}_i^c(x_0^{t-1}) \equiv \mathcal{S}/\mathcal{S}_i(x_0^{t-1})$$

Proof. By the law of total probability:

$$\begin{aligned}
\phi(x_0^{t-1}) &\equiv \mathbb{P}(S_t | X_0^{t-1} = x_0^{t-1}) \\
&= \sum_{j \in \mathcal{S}} \mathbb{P}(S_0 = j | X_0^{t-1} = x_0^{t-1}) \cdot \mathbb{P}(S_t | S_0 = j, X_0^{t-1} = x_0^{t-1}) \\
&= \sum_{j \in \mathcal{S}} \frac{\mathbb{P}(S_0 = j, X_0^{t-1} = x_0^{t-1})}{\mathbb{P}(X_0^{t-1} = x_0^{t-1})} \cdot \phi_j(x_0^{t-1}) \\
&= \sum_{j \in \mathcal{S}} \frac{\pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\mathbb{P}(x_0^{t-1})} \cdot \phi_j(x_0^{t-1}) \\
&= \sum_{j \in \mathcal{S}_i(x_0^{t-1})} \frac{\pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\mathbb{P}(x_0^{t-1})} \cdot \phi_i(x_0^{t-1}) \ + \sum_{j \in \mathcal{S}_i^c(x_0^{t-1})} \frac{\pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\mathbb{P}(x_0^{t-1})} \cdot \phi_j(x_0^{t-1})
\end{aligned}$$

The claim follows since, by unifilarity, the two distributions $\phi_i(x_0^{t-1})$ and $\phi_j(x_0^{t-1})$ can have no common support if they are not equal (i.e., if $j \in \mathcal{S}_i^c(x_0^{t-1})$). □

Corollary (1). *For each $i \in \mathcal{S}$ and $x_0^{t-1} \in \mathcal{L}(i)$:*

$$\|\phi(x_0^{t-1}) - \phi_i(x_0^{t-1})\|_{TV} \leq (|\mathcal{S}| - 1|) \cdot r_{\max} \cdot \max_{j \in \mathcal{S}_i^c(x_0^{t-1})} \frac{\mathbb{P}_j(x_0^{t-1})}{\mathbb{P}_i(x_0^{t-1})} \tag{24}$$

Proof. Since $|S_i^c(x_0^{t-1})| \leq |\mathcal{S}| - 1$ we have:

$$\begin{aligned}
\|\phi(x_0^{t-1}) - \phi_i(x_0^{t-1})\|_{TV} &= \frac{\sum_{j \in \mathcal{S}_i^c(x_0^{t-1})} \pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\sum_{j \in \mathcal{S}} \pi_j \cdot \mathbb{P}_j(x_0^{t-1})} \\
&\leq \frac{\sum_{j \in \mathcal{S}_i^c(x_0^{t-1})} \pi_j \cdot \mathbb{P}_j(x_0^{t-1})}{\pi_i \cdot \mathbb{P}_i(x_0^{t-1})} \\
&\leq (|\mathcal{S}| - 1) \cdot r_{\max} \cdot \max_{j \in \mathcal{S}_i^c(x_0^{t-1})} \frac{\mathbb{P}_j(x_0^{t-1})}{\mathbb{P}_i(x_0^{t-1})}
\end{aligned}$$

□

Corollary (2). *For each $i \in \mathcal{S}$ and $x_0^\infty \in \mathcal{L}_\infty^+(i)$:*

$$\left(\frac{p_{\min}}{p_{\max}}\right) \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV} \leq \|\phi(x_0^{t+1}) - \phi_i(x_0^{t+1})\|_{TV} \leq \left(\frac{p_{\max}}{p_{\min}}\right) \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV} \tag{25}$$

for all sufficiently large t.

Proof. For given $i \in \mathcal{S}$ and $x_0^\infty \in \mathcal{L}_\infty^+(i)$ let:

$$\begin{aligned}
A &\equiv \{j \in \mathcal{S} : x_0^\infty \in \mathcal{L}_\infty^+(j)\} \\
B &\equiv \{j \in A : \overline{\delta}_j(x_0^t) \neq \overline{\delta}_i(x_0^t), \text{ for any } t\}
\end{aligned}$$

Also, let $t_0 \geq 0$ be defined by:

$$t_0 \equiv \min\{t : \mathbb{P}_j(x_0^t) = 0, \forall j \notin A \text{ and } \overline{\delta}_j(x_0^t) = \overline{\delta}_i(x_0^t), \forall j \in A/B\}$$

Then, by the lemma, we have for each $t \geq t_0$:

$$\|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV} = \frac{\sum_{j \in \mathcal{S}_i^c(x_0^t)} \pi_j \cdot \mathbb{P}_j(x_0^t)}{\sum_{j \in \mathcal{S}} \pi_j \cdot \mathbb{P}_j(x_0^t)} = \frac{\sum_{j \in B} \pi_j \mathbb{P}_j(x_0^t)}{\sum_{j \in A} \pi_j \mathbb{P}_j(x_0^t)}$$

Therefore, since $\mathbb{P}_j(x_0^{t+1}) = \mathbb{P}_j(x_0^t) \cdot \mathbb{P}_{\overline{\delta}_j(x_0^t)}(x_{t+1})$ satisfies:

$$p_{\min} \cdot \mathbb{P}_j(x_0^t) \leq \mathbb{P}_j(x_0^{t+1}) \leq p_{\max} \cdot \mathbb{P}_j(x_0^t) \text{ , for each } j \in A, \, t \geq 0$$

we have for all $t \geq t_0$ that:

$$\begin{aligned}
\|\phi(x_0^{t+1}) - \phi_i(x_0^{t+1})\|_{TV} &= \frac{\sum_{j \in B} \pi_j \mathbb{P}_j(x_0^{t+1})}{\sum_{j \in A} \pi_j \mathbb{P}_j(x_0^{t+1})} \\
&\leq \frac{\sum_{j \in B} \pi_j (p_{\max} \cdot \mathbb{P}_j(x_0^t))}{\sum_{j \in A} \pi_j (p_{\min} \cdot \mathbb{P}_j(x_0^t))} \\
&= \left(\frac{p_{\max}}{p_{\min}}\right) \cdot \frac{\sum_{j \in B} \pi_j \mathbb{P}_j(x_0^t)}{\sum_{j \in A} \pi_j \mathbb{P}_j(x_0^t)} \\
&= \left(\frac{p_{\max}}{p_{\min}}\right) \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV}
\end{aligned}$$

and:

$$\begin{aligned}
\|\phi(x_0^{t+1}) - \phi_i(x_0^{t+1})\|_{TV} &= \frac{\sum_{j \in B} \pi_j \mathbb{P}_j(x_0^{t+1})}{\sum_{j \in A} \pi_j \mathbb{P}_j(x_0^{t+1})} \\
&\geq \frac{\sum_{j \in B} \pi_j (p_{\min} \cdot \mathbb{P}_j(x_0^t))}{\sum_{j \in A} \pi_j (p_{\max} \cdot \mathbb{P}_j(x_0^t))} \\
&= \left(\frac{p_{\min}}{p_{\max}}\right) \cdot \frac{\sum_{j \in B} \pi_j \mathbb{P}_j(x_0^t)}{\sum_{j \in A} \pi_j \mathbb{P}_j(x_0^t)} \\
&= \left(\frac{p_{\min}}{p_{\max}}\right) \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV}
\end{aligned}$$

□

5.3 Pointwise Decay of the State Uncertatinty U_t under PD1 Assumption

Throughout Section 5.3 we assume that $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is a nonexact, PD1, irreducible, unifilar HMM with associated follower HMM M_{foll}, and for each $i, j \in \mathcal{S}$ we define $D_{i,j} \in (0, \infty]$ by:

$$D_{i,j} \equiv D(\mathbb{P}_i(X_0) || \mathbb{P}_j(X_0))$$

Also, for each nontrivial component $\mathcal{C}_k$ of M_{foll} we define the constants $0 < \gamma_k < \infty$ and $0 < \eta_k < 1$ by[12]:

$$\gamma_k \equiv \sum_{(i,\bar{i}) \in \mathcal{C}_k} \nu^k_{(i,\bar{i})} \cdot D_{i,\bar{i}} \quad \text{and} \quad \eta_k \equiv 2^{-\gamma_k}$$

Finally, we define $0 < \alpha < 1$ by:

$$\alpha \equiv \max\{\eta_k : \mathcal{C}_k \text{ is nontrivial}\} \tag{26}$$

and the index k^* by:

$$k^* \equiv \text{argmax}\{\eta_k : \mathcal{C}_k \text{ is nontrivial}\}$$

chosen arbitrarily among maximizing k in the case of a tie.

We will show below in Proposition 3 that the state uncertainty U_t for the HMM M a.s. decays at exponential rate α. Before doing so, however, we will need to establish a few lemmas. The first characterizes the asymptotic behavior from the nontrivial recurrent components of the follower model, and the second characterizes the pointwise decay rate of the difference between the induced distributions $\phi(X_0^{t-1})$ and $\phi_i(X_0^{t-1})$.

Lemma 14. *For each nontrivial component $\mathcal{C}_k$ of M_{foll} and state $(i, \bar{i}) \in \mathcal{C}_k$:*

$$\lim_{t \to \infty} \left(\frac{\mathbb{P}_{\bar{i}}(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} \right)^{1/t} = \eta_k \ , \ \mathbb{P}^{foll}_{(i,\bar{i})} \ a.s.$$

Proof. Let $f : \mathcal{Z}_k \to \mathbb{R}$ be defined by[13]:

$$f(j, \bar{j}, x) \equiv \log_2 \left(\frac{\mathbb{P}_j(x)}{\mathbb{P}_{\bar{j}}(x)} \right)$$

[12] Note that $\gamma_k < \infty$ for each nontrivial $\mathcal{C}_k$, since $D_{i,\bar{i}} < \infty$ for each nontrivial $\mathcal{C}_k$ and $(i, \bar{i}) \in \mathcal{C}_k$ by point (iii) above on the recurrent decomposition of M_{foll}.

[13] Note that since $\mathcal{C}_k$ is nontrivial, $\mathbb{P}_{\bar{j}}(x) > 0$ for each edge $z = (j, \bar{j}, x) \in \mathcal{Z}_k$. Thus, $f(z)$ is finite for each $z \in \mathcal{Z}_k$.

Then, by Lemma 12, we know that for each $t \in \mathbb{N}$:

$$\begin{aligned}
\log_2 \left(\frac{\mathbb{P}_i(Y_0^{t-1})}{\mathbb{P}_{\bar{i}}(Y_0^{t-1})} \right) &= \log_2 \left(\frac{\prod_{n=0}^{t-1} \mathbb{P}_{R_n}(Y_n)}{\prod_{n=0}^{t-1} \mathbb{P}_{\overline{R}_n}(Y_n)} \right) \\
&= \sum_{n=0}^{t-1} \log_2 \left(\frac{\mathbb{P}_{R_n}(Y_n)}{\mathbb{P}_{\overline{R}_n}(Y_n)} \right) \\
&= \sum_{n=0}^{t-1} f(Z_n)
\end{aligned}$$

if the initial follower model state is $(R_0, \overline{R}_0) = (i, \bar{i})$. Thus, by the strong law for Markov chains [24, Theorem 4.16] we have that $\mathbb{P}^{foll}_{(i,\bar{i})}$ a.s. :

$$\begin{aligned}
\lim_{t \to \infty} \frac{1}{t} \log_2 \left(\frac{\mathbb{P}_i(Y_0^{t-1})}{\mathbb{P}_{\bar{i}}(Y_0^{t-1})} \right) &= \mathbb{E}_{\mu^k}(f) \\
&\equiv \sum_{(j,\bar{j},x) \in \mathcal{Z}_k} \mu^k_{(j,\bar{j},x)} \cdot f(j, \bar{j}, x) \\
&= \sum_{(j,\bar{j},x) \in \mathcal{Z}_k} \left(\nu^k_{(j,\bar{j})} \cdot \mathbb{P}_j(x) \right) \cdot \log_2 \left(\frac{\mathbb{P}_j(x)}{\mathbb{P}_{\bar{j}}(x)} \right) \\
&= \sum_{(j,\bar{j}) \in \mathcal{C}_k} \nu^k_{(j,\bar{j})} \cdot D_{j,\bar{j}} \\
&= \gamma_k
\end{aligned}$$

The claim follows since $\lim_{t \to \infty} \frac{1}{t} \log_2 \left(\frac{\mathbb{P}_i(Y_0^{t-1})}{\mathbb{P}_{\bar{i}}(Y_0^{t-1})} \right) = \gamma_k$ implies:

$$\lim_{t \to \infty} \left(\frac{\mathbb{P}_{\bar{i}}(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} \right)^{1/t} = 2^{-\gamma_k} = \eta_k$$

□

Lemma 15. *For each state $i \in \mathcal{S}$:*

$$\lim_{t \to \infty} \|\phi(X_0^{t-1}) - \phi_i(X_0^{t-1})\|_{TV}^{1/t} = \alpha \ , \ \mathbb{P}_i \ a.s.$$

Proof. For each $i, j \in \mathcal{S}$ the hitting time for the recurrent state set $\mathcal{R}$:

$$T \equiv \inf\{t \geq 0 : (R_t, \overline{R}_t) \in \mathcal{R}\}$$

is a.s. finite with respect to the measure $\mathbb{P}^{foll}_{(i,j)}$, and once a recurrent component is hit we can characterize the asymptotic behavior by Lemmas 12 and 14.

In particular, if $(R_0, \overline{R}_0) = (i, j)$ and $(R_T, \overline{R}_T) \in \mathcal{C}_k$, for some nontrivial $\mathcal{C}_k$, then:

$$\lim_{t\to\infty} \left(\frac{\mathbb{P}_j(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} \right)^{1/t} = \lim_{t\to\infty} \left(\frac{\mathbb{P}_j(Y_0^{T-1})}{\mathbb{P}_i(Y_0^{T-1})} \cdot \frac{\mathbb{P}_{\overline{R}_T}(Y_T^{t-1})}{\mathbb{P}_{R_T}(Y_T^{t-1})} \right)^{1/t} = \eta_k$$

with probability 1. On the other hand, if $(R_0, \overline{R}_0) = (i, j)$ and $(R_T, \overline{R}_T) \in \mathcal{C}_d$ then:

$$\mathbb{P}_j(Y_0^{t-1}) = \frac{\mathbb{P}_j(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} = 0 \ , \text{ for all } t \geq T$$

and if $(R_0, \overline{R}_0) = (i, j)$ and $(R_T, \overline{R}_T) \in \mathcal{C}_m$ then:

$$\overline{\delta}_j(Y_0^{t-1}) = \overline{\delta}_i(Y_0^{t-1}) \ , \text{ for all } t \geq T$$

Thus, for each $i, j \in \mathcal{S}$, we know that with probability 1 with respect to the measure $\mathbb{P}^{foll}_{(i,j)}$ either:

(a) $\lim_{t\to\infty} \left(\frac{\mathbb{P}_j(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} \right)^{1/t} \leq \alpha$, or

(b) $\overline{\delta}_j(Y_0^{t-1}) = \overline{\delta}_i(Y_0^{t-1})$, for all sufficiently large t.

Since there are only finitely many states $j \in \mathcal{S}$ and $\mathbb{P}_i(X_0^\infty) = \mathbb{P}^{foll}_{(i,j)}(Y_0^\infty)$ for each $i, j \in \mathcal{S}$ (by Lemma 12) it follows that:

$$\mathbb{P}_i(G_i^1) = 1 \ , \quad \text{for each } i \in \mathcal{S}$$

where G_i^1 is the set of symbol sequences $x_0^\infty \in \mathcal{L}^+_\infty(i)$ such that for each state $j \in \mathcal{S}$ either:

(a') $\lim_{t\to\infty} \left(\frac{\mathbb{P}_j(x_0^{t-1})}{\mathbb{P}_i(x_0^{t-1})} \right)^{1/t} \leq \alpha$, or

(b') $\overline{\delta}_j(x_0^{t-1}) = \overline{\delta}_i(x_0^{t-1})$, for all sufficiently large t.

Now, by point (iv) for the recurrent decomposition, we know that for each $i \in \mathcal{S}$ there exists some $j_i \in \mathcal{S}$ such that $(i, j_i) \in C_{k^*}$. And, by Lemma 14, we have:

$$\lim_{t \to \infty} \left(\frac{\mathbb{P}_{j_i}(Y_0^{t-1})}{\mathbb{P}_i(Y_0^{t-1})} \right)^{1/t} = \eta_{k^*} = \alpha \ , \ \mathbb{P}_{i,j_i}^{foll} \text{ a.s.}$$

Thus, since $\mathbb{P}_i(X_0^\infty) = \mathbb{P}_{(i,j_i)}^{foll}(Y_0^\infty)$, we know:

$$\mathbb{P}_i(G_i^2) = 1 \ , \quad \text{for each } i \in \mathcal{S}$$

where:

$$G_i^2 \equiv \left\{ x_0^\infty \in \mathcal{L}_\infty^+(i) : \lim_{t \to \infty} \left(\frac{\mathbb{P}_{j_i}(x_0^{t-1})}{\mathbb{P}_i(x_0^{t-1})} \right)^{1/t} = \alpha \right\}$$

Therefore, for each i, the set of "good" symbol sequences $G_i \equiv G_i^1 \cap G_i^2$ is a probability 1 set with respect to the measure $\mathbb{P}_i$. The claim follows from Lemma 13, which implies:

$$\lim_{t \to \infty} \|\phi(x_0^{t-1}) - \phi_i(x_0^{t-1})\|_{TV}^{1/t} = \alpha \ , \quad \text{for each } x_0^\infty \in G_i$$

□

Proposition 3. *The state uncertainty U_t a.s. decays at exponential rate α:*

$$\lim_{t \to \infty} U_t^{1/t} = \alpha \ , \ \mathbb{P} \ a.s.$$

Proof. By unifilarity, we know that for each $i \in \mathcal{S}$, $x_0^{t-1} \in \mathcal{L}(i)$ the distribution $\phi_i(x_0^{t-1})$ has support only a single state $\overline{\delta}_i(x_0^{t-1})$. Thus, $H(\phi_i(x_0^{t-1})) = 0$ and we may express the state uncertainty $u(x_0^{t-1})$ as:

$$u(x_0^{t-1}) \equiv H(\phi(x_0^{t-1})) = |H(\phi(x_0^{t-1})) - H(\phi_i(x_0^{t-1}))|$$

Using this relation and the norm bounds for entropy differences given by Lemmas 2 and 3, it follows from the previous lemma that:

$$\lim_{t\to\infty} u(x_0^{t-1})^{1/t} = \lim_{t\to\infty} \|\phi(x_0^{t-1}) - \phi_i(x_0^{t-1})\|_{TV}^{1/t} = \alpha \ , \quad \text{for } \mathbb{P}_i \text{ a.e. } x_0^\infty$$

The claim follows since $\mathbb{P} = \sum_i \pi_i \mathbb{P}_i$ is a weighted average of the $\mathbb{P}_i$ measures. □

5.4 L_1 Decay of the State Uncertatinty U_t under PD1 Assumption

Throughout Section 5.4 we assume that $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is a nonexact, PD1, irreducible, unifilar HMM with associated follower HMM M_{foll} and define the constants $\gamma_k > 0$ as in the previous section. Our goal is to establish bounds on the L_1 rate of decay of the state uncertainty U_t, as opposed to the pointwise (almost sure) decay rate analyzed above. The proofs rely heavily on the follower model construction, as in the pointwise case, but also use the large deviation and absorption time results for Markov chains discussed in Section 3.2. Before proceeding, however, we must first define a number of constants that will be used in the proofs and introduce some additional notation.

Let $f : \mathcal{Z} \to \mathbb{R} \cup \{\infty\}$ be defined by:

$$f(i, \bar{i}, x) \equiv \log_2 \left(\frac{\mathbb{P}_i(x)}{\mathbb{P}_{\bar{i}}(x)} \right)$$

For each nontrivial component $\mathcal{C}_k$, let $\epsilon_k \equiv \gamma_k/2$ and take $\beta_k \in (0,1)$ to be the constant $\beta(\epsilon_k)$ given by Lemma 5 for deviations in the empirical average of f evaluated on the irreducible, Markov edge sequence (Z_t) of the component $\mathcal{C}_k$ [14], so that:

$$\limsup_{t\to\infty} \ \mathbb{P}_{(i,\bar{i})}^{foll} \left(\left| \frac{1}{t} \sum_{n=0}^{t-1} f(Z_n) - \mathbb{E}_{\mu^k}(f) \right| \geq \epsilon_k \right)^{1/t} \leq \beta_k$$

for each $(i, \bar{i}) \in \mathcal{C}_k$.

We define the constants $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, b_1 \in (0,1)$ and $b_2 > 0$ as follows:

- $\alpha_1 \equiv 2^{-(\gamma_{min}/2)}$ where $\gamma_{\min} \equiv \min\{\gamma_k : \mathcal{C}_k \text{ is nontrivial}\}$.
- $\alpha_2 \equiv (1 + \alpha_1)/2$.

[14] Note that for each nontrivial $\mathcal{C}_k$ and $(i, \bar{i}, x) \in \mathcal{Z}_k$, $f(i, \bar{i}, x) < \infty$ by point (iii) above for the recurrent decomposition. So, f restricted to $\mathcal{Z}_k$ is always a real-valued map and Lemma 5 may be applied.

- α_3' is the spectral radius of $\widetilde{\mathcal{Q}}$, where $\widetilde{\mathcal{Q}}$ is the restriction of the transition matrix $\mathcal{Q}$ for the follower HMM M_{foll} to the transient states $(\mathcal{S} \times \overline{\mathcal{S}})/\mathcal{R}$. In the case there are no transient states, $\alpha_3' \equiv 0$. [15]
- $\alpha_4' \equiv \max\{\beta_k : \mathcal{C}_k \text{ is nontrivial}\}$.
- α_3 and α_4 are taken as any values such that $\alpha_3 \in (\alpha_3', 1)$ and $\alpha_4 \in (\alpha_4', 1)$.
- $b_1 \equiv \ln(\alpha_2 \cdot \frac{p_{\min}}{p_{\max}})/\ln(\alpha_1 \cdot \frac{p_{\min}}{p_{\max}})$ (where $p_{\min}$, $p_{\max}$ are the values for M).
- $b_2 \equiv (|\mathcal{S}| - 1) \cdot r_{\max} \cdot \frac{p_{\max}}{p_{\min}} \cdot \frac{1}{\alpha_1}$ (where $r_{\max}$, $p_{\min}$, $p_{\max}$ are the values for M).
- $\alpha_5' \equiv \max\{\alpha_3^{1-b_1}, \alpha_4^{b_1}\}$.
- α_5 is taken as any value such that $\alpha_5 \in (\alpha_5', 1)$.

Also, we define the following events for each $t \in \mathbb{N}$ and nontrivial $\mathcal{C}_k$:

- $A_t \equiv \{(R_t, \overline{R}_t) \in \mathcal{R}\}$
- $B_{k,t} \equiv \left\{ \left| \frac{1}{t} \sum_{n=0}^{t-1} f(Z_n) - \mathbb{E}_{\mu^k}(f) \right| < \epsilon_k \right\}$
- $C_{1,t} \equiv \{R_\tau = \overline{R}_\tau\}$
- $C_{2,t} \equiv \left\{ \frac{\mathbb{P}_{\overline{R}_\tau}(Y_\tau^{t-1})}{\mathbb{P}_{R_\tau}(Y_\tau^{t-1})} \leq \alpha_1^{t-\tau} \right\}$
- $C_t \equiv A_\tau \cap (C_{1,t} \cup C_{2,t})$

where $\tau = \tau(t)$ is defined by:

$$\tau \equiv \lceil (1 - b_1)t \rceil$$

We will show below in Proposition 4 that the expected state uncertainty $\mathbb{E}(U_t)$ decays exponentially at rate α or faster, where $0 < \alpha < 1$ is defined by:

$$\alpha \equiv \max\{\alpha_2, (\alpha_3')^{1-b_1}, \left(\alpha_4'\right)^{b_1}\} \tag{27}$$

First, though, we will need to introduce a few lemmas.

[15] Note that α_3' is always strictly less than 1 by Lemma 6, since some recurrent state $(j, \overline{j}) \in \mathcal{R}$ is accessible from each $(i, \overline{i}) \in (\mathcal{S} \times \overline{\mathcal{S}})/\mathcal{R}$.

Lemma 16. *For each* $(i,\bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ *:*

$$\mathbb{P}^{foll}_{(i,\bar{i})}(C_t^c) \leq \alpha_5^t \text{ , } \textit{for all sufficiently large } t$$

Proof. Note that:

1. For each $(i,\bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$:

$$\mathbb{P}^{foll}_{(i,\bar{i})}(A_t^c) \leq \alpha_3^t \text{ , for all sufficiently large } t$$

by Lemma 6, since $\alpha_3 > \alpha_3'$.

2. For each nontrivial $\mathcal{C}_k$ and $(j,\bar{j}) \in \mathcal{C}_k$:

$$\mathbb{P}^{foll}_{(j,\bar{j})}(B_{k,t}^c) \leq \alpha_4^t \text{ , for all sufficiently large } t$$

since $\alpha_4 > \alpha_4' \geq \beta_k$.

3. If $(R_0, \overline{R}_0) = (j, \bar{j})$ for some $(j,\bar{j}) \in \mathcal{C}_k$, $\mathcal{C}_k$ nontrivial, then on the event $B_{k,t}$ we have:

$$\begin{aligned}
&\left| \frac{1}{t} \sum_{n=0}^{t-1} f(Z_n) - \mathbb{E}_{\mu^k}(f) \right| < \epsilon_k = \gamma_k/2 \\
&\overset{(a)}{\Longrightarrow} \frac{1}{t} \sum_{n=0}^{t-1} f(Z_n) > \gamma_k/2 \geq \gamma_{\min}/2 \\
&\Longrightarrow \frac{1}{t} \log_2 \left(\frac{\prod_{n=0}^{t-1} \mathbb{P}_{R_n}(Y_n)}{\prod_{n=0}^{t-1} \mathbb{P}_{\overline{R}_n}(Y_n)} \right) = \frac{1}{t} \sum_{n=0}^{t-1} \log_2 \left(\frac{\mathbb{P}_{R_n}(Y_n)}{\mathbb{P}_{\overline{R}_n}(Y_n)} \right) > \gamma_{\min}/2 \\
&\overset{(b)}{\Longrightarrow} \frac{1}{t} \log_2 \left(\frac{\mathbb{P}_j(Y_0^{t-1})}{\mathbb{P}_{\bar{j}}(Y_0^{t-1})} \right) > \gamma_{\min}/2 \\
&\Longrightarrow \frac{\mathbb{P}_{\bar{j}}(Y_0^{t-1})}{\mathbb{P}_j(Y_0^{t-1})} < \left(2^{-(\gamma_{\min}/2)} \right)^t = \alpha_1^t
\end{aligned}$$

Here (a) follows from the fact that $\mathbb{E}_{\mu^k}(f) = \gamma_k$ and (b) from Lemma 12.

Now, if A_τ occurs C_t can only fail to occur if $(R_\tau, \overline{R}_\tau)$ lies in some nontrivial recurrent component $\mathcal{C}_k$ and $\frac{\mathbb{P}_{\overline{R}_\tau}(Y_\tau^{t-1})}{\mathbb{P}_{R_\tau}(Y_\tau^{t-1})} > \alpha_1^{t-\tau}$. Thus, by Points 2 and 3 above, we know that for each initial state

$(i,\bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ and all sufficiently large t:

$$\begin{aligned}
\mathbb{P}^{foll}_{(i,\bar{i})}\left(C_t^c | A_\tau\right) &\leq \max_{\text{nontrivial } \mathcal{C}_k} \max_{(j,\bar{j})\in \mathcal{C}_k'} \mathbb{P}^{foll}_{(i,\bar{i})} \left(\frac{\mathbb{P}_{\overline{R}_\tau}(Y_\tau^{t-1})}{\mathbb{P}_{R_\tau}(Y_\tau^{t-1})} > \alpha_1^{t-\tau} \middle| (R_\tau, \overline{R}_\tau) = (j,\bar{j}) \right) \\
&\leq \max_{\text{nontrivial } \mathcal{C}_k} \max_{(j,\bar{j})\in \mathcal{C}_k} \mathbb{P}^{foll}_{(j,\bar{j})} \left(\frac{\mathbb{P}_{\bar{j}}(Y_0^{t-\tau-1})}{\mathbb{P}_{j}(Y_0^{t-\tau-1})} > \alpha_1^{t-\tau} \right) \\
&\leq \max_{\text{nontrivial } \mathcal{C}_k} \max_{(j,\bar{j})\in \mathcal{C}_k} \mathbb{P}^{foll}_{(j,\bar{j})} \left(B^c_{k,(t-\tau)} \right) \\
&\leq \alpha_4^{t-\tau} && (28)
\end{aligned}$$

where $\mathcal{C}_k' \equiv \left\{ (j,\bar{j}) \in \mathcal{C}_k : \mathbb{P}^{foll}_{(i,\bar{i})}\left((R_\tau, \overline{R}_\tau) = (j,\bar{j}) \right) > 0 \right\}$. Combining the bound (28) with Point 1 above we find that for each $(i,\bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$ and all sufficiently large t:

$$\begin{aligned}
\mathbb{P}^{foll}_{(i,\bar{i})}(C_t^c) &= \mathbb{P}^{foll}_{(i,\bar{i})}(A_\tau^c) + \mathbb{P}^{foll}_{(i,\bar{i})}(A_\tau) \cdot \mathbb{P}^{foll}_{(i,\bar{i})}\left(C_\tau^c | A_\tau\right) \\
&\leq \mathbb{P}^{foll}_{(i,\bar{i})}(A_\tau^c) + \mathbb{P}^{foll}_{(i,\bar{i})}\left(C_\tau^c | A_\tau\right) \\
&\leq \alpha_3^\tau + \alpha_4^{t-\tau} \\
&\leq \alpha_3^{(1-b_1)t} + \frac{1}{\alpha_4} \cdot \alpha_4^{b_1 t} \\
&\leq \alpha_5^t
\end{aligned}$$

□

Lemma 17. *If the initial follower model state is* $(R_0, \overline{R}_0) = (i, \bar{i})$*, then on the event* $C_{2,t}$:

$$\frac{\mathbb{P}_{\bar{i}}(Y_0^{t-1})}{\mathbb{P}_{i}(Y_0^{t-1})} \leq \frac{p_{\max}}{p_{\min}} \cdot \frac{1}{\alpha_1} \cdot \alpha_2^t$$

Proof. If $(R_0, \overline{R}_0) = (i, \bar{i})$, then on the event $C_{2,t}$ we have by Lemma 12:

$$\begin{aligned}
\frac{\mathbb{P}_{\bar{i}}(Y_0^{t-1})}{\mathbb{P}_{i}(Y_0^{t-1})} &= \frac{\mathbb{P}_{\bar{i}}(Y_0^{\tau-1})}{\mathbb{P}_{i}(Y_0^{\tau-1})} \cdot \frac{\mathbb{P}_{\overline{R}_\tau}(Y_\tau^{t-1})}{\mathbb{P}_{R_\tau}(Y_\tau^{t-1})} \\
&\leq \left(\frac{p_{\max}}{p_{\min}} \right)^\tau \cdot \alpha_1^{t-\tau} \\
&\leq \left(\frac{p_{\max}}{p_{\min}} \right)^{(1-b_1)t+1} \cdot \alpha_1^{b_1 t - 1} \\
&= \frac{p_{\max}}{p_{\min}} \cdot \frac{1}{\alpha_1} \cdot \alpha_2^t
\end{aligned}$$

□

Proposition 4. *The expected state uncertainty decays exponentially with:*

$$\limsup_{t\to\infty} \ \mathbb{E}(U_t)^{1/t} \leq \alpha$$

Proof. Since M_{foll} is unifilar the follower model state sequence $(R_0^t, \overline{R}_0^t)$, and hence the occurrence of all relevant events up to time t, are completely determined by the initial state $(R_0, \overline{R}_0)$ and the symbol sequence Y_0^{t-1}. For fixed $(i, \bar{i}) \in \mathcal{S} \times \overline{\mathcal{S}}$, let us define $C'_{(i,\bar{i}),t}$ as the set of symbol sequences $y_0^{t-1} \in \mathcal{L}_t(i)$ such that for initial state $(R_0, \overline{R}_0) = (i, \bar{i})$ the event C_t occurs when $Y_0^{t-1} \in C'_{(i,\bar{i}),t}$:

$$C'_{(i,\bar{i}),t} \equiv \{y_0^{t-1} \in \mathcal{L}_t(i) : \text{ if } (R_0, \overline{R}_0) = (i, \bar{i}) \text{ and } Y_0^{t-1} = y_0^{t-1}, \text{ then } C_t \text{ occurs}\}$$

Also, for each $i \in \mathcal{S}$ let:

$$C'_{i,t} = \bigcap_{j \in \mathcal{S}} C'_{(i,j),t}$$

Then, since C_t always occurs from the initial state $(R_0, \overline{R}_0) = (i, i)$, we know by Lemmas 12 and 16 that for all sufficiently large t:

$$\mathbb{P}_i\left((C'_{i,t})^c\right) \leq \sum_{j \neq i} \mathbb{P}_i\left((C'_{(i,j),t})^c\right) = \sum_{j \neq i} \mathbb{P}^{foll}_{(i,j)}\left(C_t^c\right) \leq \left(|\mathcal{S}| - 1\right) \alpha_5^t \tag{29}$$

Now, fix $y_0^{t-1} \in C'_{i,t}$. Then, for each $j \in \mathcal{S}_i^c(y_0^{t-1})$, we know by unifilarity that $\overline{\delta}_i(y_0^{\tau-1}) \neq \overline{\delta}_j(y_0^{\tau-1})$. So, $C_{1,t}$ cannot occur when $(R_0, \overline{R}_0) = (i, j)$, $Y_0^{t-1} = y_0^{t-1}$. Instead, $C_{2,t}$ must occur when $(R_0, \overline{R}_0) = (i, j)$, $Y_0^{t-1} = y_0^{t-1}$. Thus, by Lemma 17, we have for any such $j \in \mathcal{S}_i^c(y_0^{t-1})$ that:

$$\frac{\mathbb{P}_j(y_0^{t-1})}{\mathbb{P}_i(y_0^{t-1})} \leq \frac{p_{\max}}{p_{\min}} \cdot \frac{1}{\alpha_1} \cdot \alpha_2^t$$

Applying the bound (24) yields:

$$\|\phi(y_0^{t-1}) - \phi_i(y_0^{t-1})\|_{TV} \leq (|\mathcal{S}| - 1) \cdot r_{\max} \cdot \left(\frac{p_{\max}}{p_{\min}} \cdot \frac{1}{\alpha_1} \cdot \alpha_2^t\right) = b_2 \alpha_2^t$$

Therefore, assuming t is sufficiently large that $b_2\alpha_2^t \leq 1/e$, the entropy $u(y_0^{t-1})$ may be bounded as follows using Lemma 3:

$$u(y_0^{t-1}) = \left|H(\phi(y_0^{t-1})) - H(\phi_i(y_0^{t-1}))\right| \leq |\mathcal{S}| \cdot b_2\alpha_2^t \cdot \log_2\left(\frac{1}{b_2\alpha_2^t}\right) \tag{30}$$

Using the relations (29) and (30) it follows that for all sufficient large t:

$$\begin{aligned}
\mathbb{E}(U_t) &= \sum_{w\in\mathcal{L}_t(M)} \mathbb{P}(w)u(w) \\
&= \sum_i \pi_i \sum_{w\in\mathcal{L}_t(M)} \mathbb{P}_i(w)u(w) \\
&\leq \max_i \left\{ \sum_{w\in\mathcal{L}_t(M)} \mathbb{P}_i(w)u(w) \right\} \\
&= \max_i \left\{ \sum_{w\in C'_{i,t}} \mathbb{P}_i(w)u(w) + \sum_{w\in(C'_{i,t})^c} \mathbb{P}_i(w)u(w) \right\} \\
&\leq \max_i \left\{ \mathbb{P}_i(C'_{i,t}) \cdot \left(|\mathcal{S}| b_2\alpha_2^t \log_2\left(\frac{1}{b_2\alpha_2^t}\right)\right) + \mathbb{P}_i((C'_{i,t})^c) \cdot \log_2|\mathcal{S}| \right\} \\
&\leq 1 \cdot \left(|\mathcal{S}| b_2\alpha_2^t \log_2\left(\frac{1}{b_2\alpha_2^t}\right)\right) + (|\mathcal{S}|-1)\,\alpha_5^t \cdot \log_2|\mathcal{S}|
\end{aligned}$$

and, hence, that:

$$\limsup_{t\to\infty}\ \mathbb{E}(U_t)^{1/t} \leq \alpha' \equiv \max\{\alpha_2, \alpha_5\}$$

The claim follows since α' can be chosen arbitrarily close to α by taking α_3, α_4, and α_5 arbitrarily close to α_3', α_4', and α_5', respectively. □

5.5 Decay of the State Uncertainty U_t under PD Assumption

Throughout Section 5.5 we assume that $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is a nonexact, PD, irreducible, unifilar HMM, and define M^n as the nonexact, PD1, irreducible, unifilar HMM with $n \in \mathbb{N}$ given by (17). As we will show below, the results of the previous sections for the decay of the state uncertainty in the PD1 block model can easily be extended to the PD HMM M.

Lemma 18. *Let $0 < \alpha < 1$ be the constant defined by (26) for the block model M^n. Then for each* *state $i \in \mathcal{S}$:*

$$\lim_{t\to\infty} \|\phi(X_0^{t-1}) - \phi_i(X_0^{t-1})\|_{TV}^{1/t} = \alpha^{1/n} \ , \ \mathbb{P}_i \ a.s.$$

Proof. By Lemma 7 and Lemma 15 (applied to the block model) we have:

$$\lim_{t\to\infty} \|\phi(X_0^{nt-1}) - \phi_i(X_0^{nt-1})\|_{TV}^{1/(nt)} = \Big(\lim_{t\to\infty} \|\phi(X_0^{nt-1}) - \phi_i(X_0^{nt-1})\|_{TV}^{1/t}\Big)^{1/n} = \alpha^{1/n} \ , \ \mathbb{P}_i \text{ a.s.}$$

The claim follows from this fact and the inequalities (25), which ensure that for each $x_0^\infty \in \mathcal{L}_\infty^+(i)$:

$$\left(\frac{p_{\min}}{p_{\max}}\right)^{n-1} \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV} \leq \|\phi(x_0^{t+m}) - \phi_i(x_0^{t+m})\|_{TV} \leq \left(\frac{p_{\max}}{p_{\min}}\right)^{n-1} \cdot \|\phi(x_0^t) - \phi_i(x_0^t)\|_{TV}$$

for all sufficiently large t and $1 \leq m \leq n-1$. □

Proposition 5. *Let $0 < \alpha < 1$ be the constant defined by (26) for the block model M^n. Then the state uncertainty U_t for M a.s. decays at exponential rate $\alpha^{1/n}$:*

$$\lim_{t\to\infty} U_t^{1/t} = \alpha^{1/n} \ , \ \mathbb{P} \ a.s.$$

Proof. The proof of this proposition from Lemma 18 is identical to the proof of Proposition 3 from Lemma 15. □

Proposition 6. *Let $0 < \alpha < 1$ be the constant defined by (27) for the block model M^n. Then the expected state uncertainty for M decays exponentially with:*

$$\limsup_{t\to\infty} \ \mathbb{E}(U_t)^{1/t} \leq \alpha^{1/n}$$

Proof. This follows immediately from Proposition 4 (applied to the block model) and Lemma 8. □

5.6 The Entropy Rate and Convergence of the Block Estimates

In general, there is no simple closed form expression for the entropy rate of a HMM. However, in the case of unifilar HMMs a simple formula does exist, as given by the following proposition.

Proposition 7. *For an irreducible, unifilar HMM M:*

$$h = H(X_0|S_0) = \sum_i \pi_i h_i \tag{31}$$

where $h_i \equiv H(X_0|S_0 = i)$ is the uncertainty in the next symbol to be generated from state i.

Proof. We establish the bounds from above and below separately. Both are direct computations using:

(a) Unifilarity of the HMM M, which implies $H(X_t|S_0, X_0^{t-1}) = H(X_t|S_t)$, $\forall t \in \mathbb{N}$, and

(b) Stationarity of the joint process $(S_t, X_t)_{t\in\mathbb{Z}}$, which implies $H(X_t|S_t) = H(X_0|S_0)$, $\forall t \in \mathbb{N}$.

Upper bound:

$$\begin{aligned}
h &= \lim_{t\to\infty} \frac{H(X_0^{t-1})}{t} \\
&\leq \lim_{t\to\infty} \frac{H(S_0, X_0^{t-1})}{t} \\
&= \lim_{t\to\infty} \frac{H(S_0) + \sum_{\tau=0}^{t-1} H(X_\tau|S_0, X_0^{\tau-1})}{t} \\
&\overset{(a)}{=} \lim_{t\to\infty} \frac{H(S_0) + \sum_{\tau=0}^{t-1} H(X_\tau|S_\tau)}{t} \\
&\overset{(b)}{=} \lim_{t\to\infty} \frac{H(S_0) + H(X_0|S_0)\cdot t}{t} \\
&= H(X_0|S_0)
\end{aligned}$$

Lower bound:

$$\begin{aligned}
h &= \lim_{t\to\infty} \frac{H(X_0^{t-1})}{t} \\
&\geq \lim_{t\to\infty} \frac{H(X_0^{t-1}|S_0)}{t} \\
&= \lim_{t\to\infty} \frac{\sum_{\tau=0}^{t-1} H(X_\tau|S_0, X_0^{\tau-1})}{t} \\
&\overset{(a)}{=} \lim_{t\to\infty} \frac{\sum_{\tau=0}^{t-1} H(X_\tau|S_\tau)}{t} \\
&\overset{(b)}{=} \lim_{t\to\infty} \frac{H(X_0|S_0)\cdot t}{t} \\
&= H(X_0|S_0)
\end{aligned}$$

□

This unifilar entropy rate formula (31) was originally given in the seminal information theory paper [32], though the method of proof was somewhat different. The proof given above, which we find somewhat cleaner, is adapted from a proof of the analogous result for state-emitting HMMs in [33].

Of course, in light of this formula, there is no need to worry about the convergence rate of estimates $h(t)$ if one is only interested in trying to get a good approximation of the true entropy rate h. Nevertheless, we still find the rate of convergence of the estimates an interesting question itself for other reasons, as discussed in the introduction.

The following upper bound on the error in the estimates $h(t)$ is a simple consequence of the entropy rate formula.

Lemma 19. *For an irreducible, unifilar HMM M:*

$$h(t+1) - h \leq \mathbb{E}(U_t) \ , \ \textit{for each } t \in \mathbb{N}$$

Proof. Applying the entropy rate formula (31) and stationarity we have:

$$\begin{aligned}
h(t+1) &= H(X_t | X_0^{t-1}) \\
&\leq H(X_t, S_t | X_0^{t-1}) \\
&= H(S_t | X_0^{t-1}) + H(X_t | S_t, X_0^{t-1}) \\
&= H(S_t | X_0^{t-1}) + H(X_t | S_t) \\
&= H(S_t | X_0^{t-1}) + H(X_0 | S_0) \\
&= \mathbb{E}(U_t) + h
\end{aligned}$$

Thus:

$$h(t+1) - h \leq \mathbb{E}(U_t)$$

□

Now, for unifilar HMMs with probabilistically distinguishable states we know that the expected state uncertainty decays exponentially fast by Proposition 6. So, by this lemma, the block estimates $h(t)$ must also converge exponentially fast in this case.

For unifilar HMMs without probabilistically distinguishable states, the expected state uncertainty may not limit to 0 (at any rate). But, as detailed in the following lemma, such a model M can always be converted to an equivalent unifilar HHM $\widetilde{M}$ with probabilistically distinguishable states, by collapsing together all states with the same probability distribution over future output.

Lemma 20. *Let $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ be an irreducible, unifilar HMM, and let $\sim$ be the equivalence relation on $\mathcal{S}$ of equality of future measures:*

$$i \sim j \Longleftrightarrow \mathbb{P}_i(X_0^\infty) = \mathbb{P}_j(X_0^\infty)$$

Let $\widetilde{\mathcal{S}}$ be the set of equivalence classes under the relation $\sim$ and, for $i \in \mathcal{S}$, let $[i]$ denote the equivalence class of state i. Also, for $[i] \in \widetilde{\mathcal{S}}$, $w \in \mathcal{X}^*$ define $\widetilde{\mathbb{P}}_{[i]}(w)$ and $\widetilde{\delta}_{[i]}(w)$ by:

$$\begin{aligned} \widetilde{\mathbb{P}}_{[i]}(w) &\equiv \mathbb{P}_i(w) \\ \widetilde{\delta}_{[i]}(w) &\equiv \begin{cases} [j], \ \textit{if } \widetilde{\mathbb{P}}_{[i]}(w) > 0 \ \textit{and } \overline{\delta}_i(w) = j \\ [d], \ \textit{if } \widetilde{\mathbb{P}}_{[i]}(w) = 0 \end{cases} \end{aligned}$$

where $[d] \notin \widetilde{\mathcal{S}}$ is a new formal symbol representing the dead state equivalence class.

Then, the HMM $\widetilde{M} = (\widetilde{\mathcal{S}}, \mathcal{X}, \{\widetilde{\mathcal{T}}^{(x)}\})$ with transition matrices $\widetilde{\mathcal{T}}^{(x)}$ defined by:

$$\widetilde{\mathcal{T}}^{(x)}_{[i][j]} \equiv \widetilde{\mathbb{P}}_{[i]}(x) \cdot \mathbb{1}\{\widetilde{\delta}_{[i]}(x) = [j]\}$$

is irreducible, unifilar, and has probabilistically distinguishable states. Moreover, its stationary output process $(\widetilde{X}_t)$ is equal in distribution to the stationary output process (X_t) for the original HMM M.

Before proceeding to the proof of the lemma itself a few comments are in order concerning well definedness of the model $\widetilde{M}$.

1. For each $i' \in [i]$, $w \in \mathcal{X}^*$ we have by definition $\mathbb{P}_i(w) = \mathbb{P}_{i'}(w)$. Thus, $\widetilde{\mathbb{P}}_{[i]}(w)$ is well defined and independent of the representative $i' \in [i]$.

2. For each $i' \in [i]$, $v \in \mathcal{X}^*$, and $w \in \mathcal{X}^*$ with $\widetilde{\mathbb{P}}_{[i]}(w) > 0$ we have by unifilarity of M:

   $$\mathbb{P}_i(wv) = \mathbb{P}_i(w) \cdot \mathbb{P}_j(v) \text{ and } \mathbb{P}_{i'}(wv) = \mathbb{P}_{i'}(w) \cdot \mathbb{P}_{j'}(v)$$

   where $j = \overline{\delta}_i(w)$ and $j' = \overline{\delta}_{i'}(w)$. Since, $\mathbb{P}_i(wv) = \mathbb{P}_{i'}(wv)$ by definition, and $\mathbb{P}_i(w) = \mathbb{P}_{i'}(w) > 0$, it follows that $\mathbb{P}_j(v) = \mathbb{P}_{j'}(v)$. Since this holds for all $v \in \mathcal{X}^*$ it follows that j and j' are in the same equivalence class. Thus, the transition function $\widetilde{\delta}_{[i]}(w)$ is well defined and independent of the representative $i' \in [i]$.

3. Combining points 1 and 2 shows that the transition matrices $\widetilde{\mathcal{T}}^{(x)}$ are well defined. Thus, to verify $\widetilde{M}$ is a well defined HMM we need only show that the overall state transition matrix $\widetilde{\mathcal{T}} = \sum_x \widetilde{\mathcal{T}}^{(x)}$ is stochastic. This a direct computation:

   $$\sum_{[j] \in \widetilde{\mathcal{S}}} \widetilde{\mathcal{T}}_{[i][j]} = \sum_x \sum_{[j] \in \widetilde{\mathcal{S}}} \widetilde{\mathcal{T}}^{(x)}_{[i][j]} = \sum_x \widetilde{\mathbb{P}}_{[i]}(x) = \sum_x \mathbb{P}_i(x) = 1$$

Proof (of lemma). Given that the transition function $\widetilde{\delta}_{[i]}(w)$ is well defined (and single valued), as shown above, unifilarity is immediate from the construction. Also, irreducibility follows directly from the definition:

$$\widetilde{\delta}_{[i]}(w) = [j] \Longleftrightarrow \overline{\delta}_i(w) = j \ , \ w \in \mathcal{X}^*, i \in \mathcal{S}, j \in \mathcal{S} \cup \{d\} \tag{32}$$

and the fact that the original HMM M is irreducible.

To show that $\widetilde{M}$ has probabilistically distinguishable states note that by unifilarity and the relation (32), we have for any $i \in \mathcal{S}, w = x_0^{t-1} \in \mathcal{X}^*$:

$$\mathbb{P}_i(w) = \prod_{\tau=0}^{t-1} \mathbb{P}_{s_\tau}(x_\tau) = \prod_{\tau=0}^{t-1} \widetilde{\mathbb{P}}_{[s_\tau]}(x_\tau) = \widetilde{\mathbb{P}}_{[i]}(w) \tag{33}$$

where the state sequence $s_0^{t-1} \in \mathcal{S}^t$ is defined by the relations:

$$s_0 = i \quad , \quad s_\tau = \overline{\delta}_{s_{\tau-1}}(x_{\tau-1}), \ \tau = 1, ..., t-1$$

and $\widetilde{\mathbb{P}}_{[d]}(x) \equiv \mathbb{P}_d(x) = 0$, for all $x \in \mathcal{X}$. Therefore, no two distinct equivalence classes $[i], [j] \in \widetilde{\mathcal{S}}$ can have the same probability distribution over future output $\widetilde{X}_0^\infty$, because this would imply that the states i and j of M also have the same probability distribution over future output X_0^∞ (contradicting the fact that $[i]$ and $[j]$ are distinct equivalence classes).

Finally, equality of the stationary output processes (X_t) and $(\widetilde{X}_t)$ for the two HMMs M and $\widetilde{M}$ follows from the relation (33) and the fact that the stationary distribution $\widetilde{\pi}$ for $\widetilde{M}$ is:

$$\widetilde{\pi}_{[i]} = \sum_{i' \in [i]} \pi_{i'}$$

which is easily verified by direct computation. □

An example of the $\widetilde{M}$ construction is presented below in Figure 7.

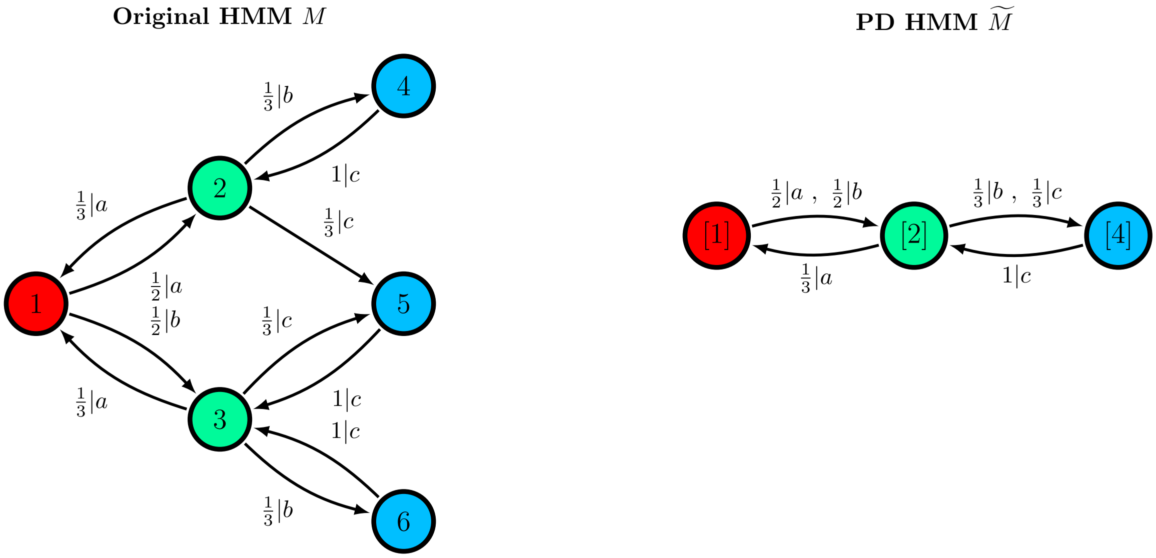


Figure 7: The $\widetilde{M}$ construction for a unifilar HMM M. The HMM M has 6 states, $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$, which reduce to 3 equivalence classes under the relation $\sim$: $[1] = \{1\}$, $[2] = \{2, 3\}$, and $[4] = \{4, 5, 6\}$. In the figure states in [1] are colored red, states in [2] are colored green, and states in [4] are colored blue.

Applying the $\widetilde{M}$ construction, we now state our main theorem for uniflar HMMs.

Theorem 2. *Let M be a nonexact, irreducible, unifilar HMM, and let $\widetilde{M}$ be the corresponding equivalent unifilar HMM with probabilistically distinguishable states given by Lemma 20. Assume that $\widetilde{M}$ is also nonexact. Let $0 < \alpha < 1$ and $n \in \mathbb{N}$ be defined as in Proposition 6 for the PD model*

$\widetilde{M}$. *Then:*

$$\limsup_{t \to \infty} \{h(t) - h\}^{1/t} \leq \alpha^{1/n}$$

Proof. Since the stationary process $(\widetilde{X}_t)$ generated by $\widetilde{M}$ has the same distribution as the stationary process (X_t) generated by M we have:

$$\widetilde{h} = h \quad \text{and} \quad \widetilde{h}(t) = h(t), t \in \mathbb{N}$$

where $\widetilde{h}$ and $\widetilde{h}(t)$ are, respectively, the entropy rate and length-t block estimate for $\widetilde{M}$. The claim follows immediately from this fact, the statement of Proposition 6:

$$\limsup_{t \to \infty} \ \mathbb{E}(\widetilde{U}_t)^{1/t} \leq \alpha^{1/n}$$

and the bound of Lemma 19:

$$\widetilde{h}(t+1) - \widetilde{h} \leq \mathbb{E}(\widetilde{U}_t)$$

□

Remarks

1. In Appendix A we give an algorithm to determine all pairs of probabilistically distinguishable states for a unifilar HMM M. Thus, construction of the reduced HMM $\widetilde{M}$ is always possible in practice, not just in theory.

2. Theorem 2 explicitly assumes that the HMM M, as well as the reduced HMM $\widetilde{M}$, are nonexact. However, for an exact, unifilar HMM (M or $\widetilde{M}$) one may apply Theorem 1 for a bound on the convergence rate of the estimates $h(t)$. This bound, using the exactness property, is at least as strong as what could be obtained using the follower model construction and unifilarity.

6 Results for Flag-State HMMs

In this section we establish exponential bounds on the speed of convergence of the entropy rate block estimates $h(t)$ for flag-state HMMs. The basic method of proof is based on the coupling technique used in [14] to prove similar bounds for state-emitting HMMs with strictly positive transition probabilities in the underlying Markov chain. However, the details are substantially more involved because our assumption is somewhat weaker than the strict positivity assumed in [14][16].

Initially, we will establish our bounds under an FS1 assumption in Section 6.2, and then later translate them to the general FS case in Section 6.3, by passing to a block model presentation. First, however, it will be necessary to introduce two auxiliary probability spaces that will be used in the proofs[17] (in addition to the standard space $(\Omega, \mathcal{F}, \mathbb{P})$ for the HMM M, generated by the random variables $(S_t, X_t)_{t \in \mathbb{Z}}$).

6.1 Auxiliary Spaces

Throughout Section 6.1 we assume that $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is an FS1, irreducible HMM and denote the flag symbol for state k by y_k. We define two important auxiliary probability spaces that will be useful for the proofs in Section 6.2 below: The pair chain coupling space and the reverse time generation space.

6.1.1 The Pair Chain Coupling Space $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mathbb{P}})$

For fixed states $k, \widehat{k} \in \mathcal{S}$ and a symbol sequence $x_0^t \in \mathcal{X}^{t+1}$ $(t \in \mathbb{N})$ such that $\mathbb{P}_k(x_0^t), \mathbb{P}_{\widehat{k}}(x_0^t) > 0$ the *pair chain coupling space* $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{\mathbb{P}})$ is defined as follows:

- $\widehat{\Omega} = \{(r_0^{t+1}, \widehat{r}_0^{t+1}) : r_\tau, \widehat{r}_\tau \in \mathcal{S} \text{ for } 0 \leq \tau \leq t+1\}$ is the set of length-$(t+2)$ state sequence pairs.

- $\widehat{\mathcal{F}}$ is the discrete σ-algebra on $\widehat{\Omega}$ (i.e. all subsets of $\widehat{\Omega}$ are measurable).

[16] Technically, the assumptions in [14] are not directly comparable to ours, as so far stated, since we consider edge-emitting HMMs, whereas [14] dealt with the state-emitting variety. However, in Appendix C we give a translation of the flag-state condition to the case of state-emitting HMMs, which facilitates more direct comparison and shows our assumptions are substantially weaker than those used in [14].

[17] In Sections 4 and 5 a similar approach was taken, implicitly, with the possibility machine and follower model, each of which had their own associated probability space and probability measure. However, it will be helpful now to be more explicit about the underlying probability spaces, especially for the reverse time generation space.

- The measure $\widehat{\mathbb{P}}$ on state sequence pairs $(r_0^{t+1}, \widehat{r}_0^{t+1}) \in \widehat{\Omega}$ is the pull back measure of the time-inhomogeneous Markov chain $(R_\tau, \widehat{R}_\tau)_{\tau=0}^{t+1}$ with initial distribution

  $$\widehat{\mathbb{P}}(R_0 = k, \widehat{R}_0 = \widehat{k}) = 1$$

  and transition probabilities

  $$\begin{aligned}
  &\widehat{\mathbb{P}}(R_{\tau+1} = j, \widehat{R}_{\tau+1} = \widehat{j} | R_\tau = i, \widehat{R}_\tau = \widehat{i}) \\
  &\quad = \begin{cases} \mathbb{P}(S_{\tau+1} = j | S_\tau = i, X_\tau^t = x_\tau^t) \cdot \mathbb{P}(S_{\tau+1} = \widehat{j} | S_\tau = \widehat{i}, X_\tau^t = x_\tau^t), \text{ if } i \neq \widehat{i} \\ \mathbb{P}(S_{\tau+1} = j | S_\tau = i, X_\tau^t = x_\tau^t), \text{ if } i = \widehat{i} \text{ and } j = \widehat{j} \\ 0, \text{ if } i = \widehat{i} \text{ and } j \neq \widehat{j} \end{cases}
  \end{aligned}$$

  for $0 \leq \tau \leq t$.

By marginalizing it follows that the state sequences (R_τ) and $(\widehat{R}_\tau)$ are each individually (time-inhomogeneuos) Markov chains with transition probabilities

$$\begin{aligned}
&\widehat{\mathbb{P}}(R_{\tau+1} = j | R_\tau = i) = \mathbb{P}(S_{\tau+1} = j | S_\tau = i, X_\tau^t = x_\tau^t), \text{ and} \\
&\widehat{\mathbb{P}}(\widehat{R}_{\tau+1} = \widehat{j} | \widehat{R}_\tau = \widehat{i}) = \mathbb{P}(S_{\tau+1} = \widehat{j} | S_\tau = \widehat{i}, X_\tau^t = x_\tau^t).
\end{aligned}$$

Hence:

$$\begin{aligned}
&\widehat{\mathbb{P}}(R_0^{t+1} = r_0^{t+1}) = \mathbb{P}(S_0^{t+1} = r_0^{t+1} | S_0 = k, X_0^t = x_0^t), \text{ and} \\
&\widehat{\mathbb{P}}(\widehat{R}_0^{t+1} = \widehat{r}_0^{t+1}) = \mathbb{P}(S_0^{t+1} = \widehat{r}_0^{t+1} | S_0 = \widehat{k}, X_0^t = x_0^t).
\end{aligned}$$

So, we have the following coupling bound:

$$\begin{aligned}
\|\phi_k(x_0^t) - \phi_{\widehat{k}}(x_0^t)\|_{TV} &\equiv \|\mathbb{P}_k(S_{t+1} | X_0^t = x_0^t) - \mathbb{P}_{\widehat{k}}(S_{t+1} | X_0^t = x_0^t)\|_{TV} \\
&\leq \widehat{\mathbb{P}}(R_{t+1} \neq \widehat{R}_{t+1}) \qquad (34)
\end{aligned}$$

This bound is the primary reason for introducing the pair chain coupling space.

6.1.2 The Reverse Time Generation Space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$

Fix a state $k \in \mathcal{S}$, and assume without loss of generality (up to relabeling) that the output alphabet is $\mathcal{X} = \{1, 2, ..., |\mathcal{X}|\}$ with $y_k = 1$. We construct the *reverse time generation space* $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ and random variables $(\widetilde{X}_t)_{t \in -\mathbb{N}}$ for state k as follows.

- $(\widetilde{U}_n)_{n \in \mathbb{N}}$ and $(\widetilde{V}_n)_{n \in \mathbb{N}}$ are i.i.d. sequences of uniform($[0,1]$) random variables independent of one another.
- $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ is the canonical probability space generated by the sequences $(\widetilde{U}_n)$ and $(\widetilde{V}_n)$.
- On this space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ the random partitions $\widetilde{P}_t, t \in -\mathbb{N}$ of the interval $[0,1]$ and random variables $\widetilde{X}_t, t \in -\mathbb{N}$ are defined inductively as follows:

  1. $\widetilde{P}_{-1} = \{\widetilde{I}^x_{-1} : x \in \mathcal{X}\}$ where each $\widetilde{I}^x_{-1}$ is an interval of length $\mathbb{P}(X_{-1} = x)$, and the intervals are consecutively placed on $[0,1]$ and closed at the right endpoint[18]. $\widetilde{X}_{-1}$ is defined by $\widetilde{X}_{-1} = x \Leftrightarrow \widetilde{U}_1 \in \widetilde{I}^x_{-1}$.
  2. Conditioned on $\widetilde{X}^{-1}_{t+1} = x^{-1}_{t+1}$, for $t \leq -2$, $\widetilde{P}_t = \{\widetilde{I}^x_t : x \in \mathcal{X}\}$ where the $\widetilde{I}^x_t$'s are intervals of length $\mathbb{P}(X_t = x | X^{-1}_{t+1} = x^{-1}_{t+1})$ placed consecutively on $[0,1]$ and closed at the right endpoint and:

     - If $t + 1 = \widetilde{T}^k_n$ for some $n \in \mathbb{N}$, then $\widetilde{X}_t$ is defined by $\widetilde{X}_t = x \Leftrightarrow \widetilde{V}_n \in \widetilde{I}^x_t$.
     - Otherwise, $\widetilde{X}_t$ is defined by $\widetilde{X}_t = x \Leftrightarrow \widetilde{U}_{-t} \in \widetilde{I}^x_t$.

     Here, the random stopping times $\widetilde{T}^k_n$, $n \in \mathbb{N}$ are defined by:

     - $\widetilde{T}^k_1 = \max\{t \leq -1 : \mathbb{P}_k(\widetilde{X}^{-1}_t) \geq \mathbb{P}_j(\widetilde{X}^{-1}_t)\ ,\ \text{for each } j \in \mathcal{S}\}$
     - $\widetilde{T}^k_2 = \max\{t < \widetilde{T}^k_1 : \mathbb{P}_k(\widetilde{X}^{-1}_t) \geq \mathbb{P}_j(\widetilde{X}^{-1}_t)\ ,\ \text{for each } j \in \mathcal{S}\}$
     - $\widetilde{T}^k_3 = \max\{t < \widetilde{T}^k_2 : \mathbb{P}_k(\widetilde{X}^{-1}_t) \geq \mathbb{P}_j(\widetilde{X}^{-1}_t)\ ,\ \text{for each } j \in \mathcal{S}\}$

       $\vdots$

     If there is no such t for some $n \geq 1$, then $\widetilde{T}^k_n \equiv -\infty$ and $\widetilde{T}^k_m \equiv -\infty$ for all $m > n$.

[18] For example, if $\mathbb{P}(X_{-1} = 1) = 1/2$, $\mathbb{P}(X_{-1} = 2) = 1/3$, and $\mathbb{P}(X_{-1} = 3) = 1/6$ then $\widetilde{I}^1_{-1} = [0, 1/2]$, $\widetilde{I}^2_{-1} = (1/2, 5/6]$, and $\widetilde{I}^3_{-1} = (5/6, 1]$.

By induction on the length of w it is easily seen that $\widetilde{\mathbb{P}}\left(\widetilde{X}_{-|w|}^{-1}=w\right)=\mathbb{P}\left(X_{-|w|}^{-1}=w\right)$ for any word $w \in \mathcal{X}^*$. Hence:

$$\widetilde{\mathbb{P}}(\widetilde{X}_{-\infty}^{-1})=\mathbb{P}(X_{-\infty}^{-1}) \tag{35}$$

Of course, this is a somewhat unnecessarily complicated way of generating the output process of the HMM M in reverse time. However, the explicit and relatively simple nature of the underlying space $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ will be useful. In particular, large deviation estimates for the i.i.d. sequences $(\widetilde{U}_n)$ and $(\widetilde{V}_n)$ generating the space can be translated to large deviation estimates for the sequence $(\widetilde{X}_t)$, which would be more difficult to do directly on the sequence (X_t).

6.2 Bounds for FS1 Models

Throughout Section 6.2 we assume $M=(\mathcal{S},\mathcal{X},\{\mathcal{T}^{(x)}\})$ is an irreducible, FS1 HMM and denote the flag symbol for state j by y_j. Also, we define the quantities $p_j, q_j, p_*, q_* \in (0,1]$ by:

$$p_j \equiv \mathbb{P}(X_0=y_j|S_1=j) \quad \text{and} \quad p_* \equiv \min_j p_j$$

$$q_j \equiv \min_{i\in\mathcal{S}(y_j)} \mathbb{P}_i(S_1=j|X_0=y_j) \quad \text{and} \quad q_* \equiv \min_j q_j$$

where:

$$\mathcal{S}(w)\equiv\{i\in\mathcal{S}: w\in\mathcal{L}(i)\}\ ,\ w\in\mathcal{X}^*$$

Our goal is to establish an exponential bound on the speed of convergence of the entropy rate block estimates $h(t)$. The basic steps of the arguments are as follows:

(i) Using large deviation estimates on the $(\widetilde{\Omega},\widetilde{\mathcal{F}},\widetilde{\mathbb{P}})$ space we show that there exists a set of "good" length-$(t+1)$ sequences G_t of combined probability $1-O$(exponentially small), such that for each $x_0^t \in G_t$ there is some state k with $N_k(x_0^t)\geq \eta t$. Here $\eta>0$ is a constant (depending on the HMM, but not on t) and:

$$N_k(x_0^t)\equiv\left|\{0\leq\tau\leq t-1: x_\tau=y_k,\ \text{and}\ \mathbb{P}_k(x_{\tau+1}^t)\geq\mathbb{P}_j(x_{\tau+1}^t)\ \text{for all}\ j\}\right| \tag{36}$$

(ii) Using a coupling argument similar to the one used in [14] we show that $\|\phi_k(x_0^t) - \phi_{\widehat{k}}(x_0^t)\|_{TV}$ is exponentially small for any sequence $x_0^t \in G_t$ and states $k, \widehat{k} \in \mathcal{S}(x_0^t)$.

(iii) Using (ii) we show that the difference $H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|X_0^t = x_0^t, S_0)$ is exponentially small for any $x_0^t \in G_t$.

(iv) Using (iii) and the fact that $\mathbb{P}(G_t^c)$ is exponentially small we show that the difference $H(X_{t+1}|X_0^t) - H(X_{t+1}|X_0^t, S_0)$ is exponentially small.

(v) Finally, using (iv) and the fact that $H(X_{t+1}|X_0^t, S_0) \leq h$, for all t, we conclude that the differences $h(t) - h$ must be exponentially small.

6.2.1 Large Deviation Estimates for the Space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$

Let the random variables $\widetilde{Z}_n, n \in \mathbb{N}$ and $\widetilde{Z}_n^{avg}, n \in \mathbb{N}$ on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ be defined by:

$$\widetilde{Z}_n \equiv \begin{cases} 1, \text{ if } \widetilde{V}_n \leq p_* r_{\min}/|\mathcal{S}| \\ 0, \text{ otherwise} \end{cases}$$

and:

$$\widetilde{Z}_n^{avg} \equiv \frac{1}{n} \sum_{m=1}^{n} \widetilde{Z}_m$$

Also, define the random fractions $\widetilde{F}_n^k$, $n \in \mathbb{N}$ by:

$$\widetilde{F}_n^k \equiv \begin{cases} -1, \text{ if } \widetilde{T}_n^k = -\infty \\ c/n, \text{ if } \widetilde{T}_n^k > -\infty \text{ where } c = |\{t \leq -1 : t = \widetilde{T}_m^k \text{ for some } 1 \leq m \leq n \text{ and } \widetilde{X}_{t-1} = y_k\}| \end{cases}$$

Lemma 21. $\widetilde{\mathbb{P}}\left(\widetilde{Z}_n^{avg} \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right) \leq \alpha_1^n$, *where* $\alpha_1 \equiv \exp\left(-\frac{p_*^2 r_{\min}^2}{2|\mathcal{S}|^2}\right) < 1$.

Proof. The $\widetilde{Z}_n$ are i.i.d. with $0 \leq \widetilde{Z}_n \leq 1$ and $\mathbb{E}(\widetilde{Z}_n) = \frac{p_* r_{\min}}{|\mathcal{S}|}$. Thus, applying Hoeffding's inequality to the i.i.d. sequence $\widetilde{Z}_n$ yields:

$$\begin{aligned}\widetilde{\mathbb{P}}\left(\widetilde{Z}_n^{avg} \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right) &= \widetilde{\mathbb{P}}\left(\widetilde{Z}_n^{avg} - \frac{p_* r_{\min}}{|\mathcal{S}|} \leq -\frac{p_* r_{\min}}{2|\mathcal{S}|}\right) \\ &\leq \exp\left(\frac{-2\left(\frac{p_* r_{\min}}{2|\mathcal{S}|}\right)^2}{(1-0)^2} \cdot n\right) \\ &= \alpha_1^n\end{aligned}$$

□

Lemma 22. *If $w \in \mathcal{L}_t(M)$ is a word such that $\mathbb{P}_k(w) \geq \mathbb{P}_j(w)$, for all j, then:*

$$\begin{aligned}&(i)\ \ \mathbb{P}(S_{-t} = k | X_{-t}^{-1} = w) \geq r_{\min} \cdot \mathbb{P}(S_{-t} = j | X_{-t}^{-1} = w)\ ,\ \textit{for all } j \\ &(ii)\ \ \mathbb{P}(S_{-t} = k | X_{-t}^{-1} = w) \geq r_{\min}/|\mathcal{S}|\end{aligned}$$

Proof. By Bayes' theorem:

$$\begin{aligned}&\mathbb{P}_k(w) \geq \mathbb{P}_j(w) \\ &\Longrightarrow \mathbb{P}(S_{-t} = k | X_{-t}^{-1} = w) \cdot \frac{\mathbb{P}(X_{-t}^{-1} = w)}{\mathbb{P}(S_{-t} = k)} \geq \mathbb{P}(S_{-t} = j | X_{-t}^{-1} = w) \cdot \frac{\mathbb{P}(X_{-t}^{-1} = w)}{\mathbb{P}(S_{-t} = j)} \\ &\Longrightarrow \mathbb{P}(S_{-t} = k | X_{-t}^{-1} = w) \geq \frac{\mathbb{P}(S_{-t} = k)}{\mathbb{P}(S_{-t} = j)} \cdot \mathbb{P}(S_{-t} = j | X_{-t}^{-1} = w) \geq r_{\min} \cdot \mathbb{P}(S_{-t} = j | X_{-t}^{-1} = w)\end{aligned}$$

This proves (i), and (ii) follows. □

Lemma 23. *If $\widetilde{T}_m^k > -\infty$ and $\widetilde{Z}_m = 1$, then $\widetilde{X}_{\widetilde{T}_m^k - 1} = y_k$.*

Proof. Note that for any $t \leq -1$ the event $\{\widetilde{T}_m^k = t\}$ depends only on $\widetilde{X}_t^{-1}$. Define:

$$W_m^k \equiv \{x_t^{-1} \in \mathcal{L}(M) : t \leq -1, \text{ and } \widetilde{X}_t^{-1} = x_t^{-1} \Longrightarrow \widetilde{T}_m^k = t\}$$

Then for any $x_t^{-1} \in W_m^k$ we have by Lemma 22:

$$\begin{aligned}\mathbb{P}(X_{t-1} = y_k | X_t^{-1} = x_t^{-1}) &\geq \mathbb{P}(S_t = k | X_t^{-1} = x_t^{-1}) \cdot \mathbb{P}(X_{t-1} = y_k | S_t = k) \\ &\geq \frac{r_{\min}}{|\mathcal{S}|} \cdot p_*\end{aligned}$$

But:

$$\widetilde{Z}_m = 1 \Longrightarrow \widetilde{V}_m \leq \frac{p_* r_{\min}}{|\mathcal{S}|}$$

Thus:

$$\begin{aligned}\widetilde{X}_t^{-1} = x_t^{-1} \text{ and } \widetilde{Z}_m = 1 &\Longrightarrow \widetilde{V}_m \in [0, \mathbb{P}(X_{t-1} = y_k | X_t^{-1} = x_t^{-1})] = \widetilde{I}_{t-1}^{y_k} \\ &\Longrightarrow \widetilde{X}_{t-1} = y_k\end{aligned}$$

Since this holds for any $x_t^{-1} \in W_m^k$ the claim follows. □

Lemma 24. *If $\widetilde{T}_n^k > -\infty$ and $\widetilde{Z}_n^{avg} > \frac{p_* r_{\min}}{2|\mathcal{S}|}$, then $\widetilde{F}_n^k > \frac{p_* r_{\min}}{2|\mathcal{S}|}$.*

Proof. Assume $\widetilde{T}_n^k > -\infty$. Then, by Lemma 23, $\widetilde{X}_{\widetilde{T}_m^k - 1} = y_k$ for each $m \in \{1, ..., n\}$ with $\widetilde{Z}_m = 1$. So:

$$\begin{aligned}\widetilde{F}_n^k &= \frac{1}{n} \left| \{1 \leq m \leq n : \widetilde{X}_{\widetilde{T}_m^k - 1} = y_k\} \right| \\ &\geq \frac{1}{n} \left| \{1 \leq m \leq n : \widetilde{Z}_m = 1\} \right| \\ &= \widetilde{Z}_n^{avg}\end{aligned}$$

Hence:

$$\widetilde{T}_n^k > -\infty \text{ and } \widetilde{Z}_n^{avg} > \frac{p_* r_{\min}}{2|\mathcal{S}|} \Longrightarrow \widetilde{F}_n^k > \frac{p_* r_{\min}}{2|\mathcal{S}|}$$

□

Lemma 25. $\widetilde{\mathbb{P}}\left(\left\{\widetilde{T}_n^k > -\infty \textit{ and } \widetilde{F}_n^k \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\}\right) \leq \alpha_1^n$.

Proof. By Lemmas 21 and 24 we have:

$$\begin{aligned}
\widetilde{\mathbb{P}}\left(\left\{\widetilde{T}_n^k > -\infty \text{ and } \widetilde{F}_n^k \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\}\right) &\leq \widetilde{\mathbb{P}}\left(\left\{\widetilde{T}_n^k > -\infty \text{ and } \widetilde{Z}_n^{avg} \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\}\right) \\
&\leq \widetilde{\mathbb{P}}\left(\left\{\widetilde{Z}_n^{avg} \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\}\right) \\
&\leq \alpha_1^n
\end{aligned}$$

□

6.2.2 Bound from Below on $\mathbb{P}(G_t)$

The random variables $\widetilde{T}_n^k$ and $\widetilde{F}_n^k$ on $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{\mathbb{P}})$ depend only on the symbol sequence $(\widetilde{X}_t)_{t \in -\mathbb{N}}$: $\widetilde{T}_n^k = \widetilde{T}_n^k(\widetilde{X}_{-\infty}^{-1})$ and $\widetilde{F}_n^k = \widetilde{F}_n^k(\widetilde{X}_{-\infty}^{-1})$. As such, we may define corresponding random variables T_n^k and F_n^k for the standard HMM probability space $(\Omega, \mathcal{F}, \mathbb{P})$, depending on $(X_t)_{t \in -\mathbb{N}}$. Since the distribution of $(\widetilde{X}_t)_{t \in -\mathbb{N}}$ is the same as that of $(X_t)_{t \in -\mathbb{N}}$ we have, necessarily, equality in distribution between these derived random variables: $(\widetilde{T}_n^k)_{n \in \mathbb{N}} \overset{d.}{=} (T_n^k)_{n \in \mathbb{N}}$ and $(\widetilde{F}_n^k)_{n \in \mathbb{N}} \overset{d.}{=} (F_n^k)_{n \in \mathbb{N}}$. This equivalence will be quite useful in the following development.

Before proceeding, however, we first need to introduce some more notation and terminology. We say $x_{-\infty}^{-1}$ is a *tail extension* of the word w if $x_{-|w|}^{-1} = w$. We define the constant $\eta = \frac{p_* r_{\min}}{2|\mathcal{S}|^2}$ and, for $t \in \mathbb{N}$, we define $n_t = \lceil t/|\mathcal{S}| \rceil$ and $N_k(x_0^t)$ as in (36). Finally, we define the following sets:

$$\begin{aligned}
A_n^k &\equiv \left\{x_{-\infty}^{-1} \in \mathcal{L}_\infty^-(\mathcal{P}) : T_n^k(x_{-\infty}^{-1}) > -\infty \;\text{ and }\; F_n^k(x_{-\infty}^{-1}) \leq \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\} \;,\; k \in \mathcal{S} \text{ and } n \in \mathbb{N} \\
B_n &\equiv \left\{x_{-\infty}^{-1} \in \mathcal{L}_\infty^-(\mathcal{P}) : F_n^k(x_{-\infty}^{-1}) > \frac{p_* r_{\min}}{2|\mathcal{S}|} \;\text{ for each } k \text{ with }\; T_n^k(x_{-\infty}^{-1}) > -\infty\right\} \;,\; n \in \mathbb{N} \\
C_t &\equiv \left\{x_{-\infty}^{-1} \in \mathcal{L}_\infty^-(\mathcal{P}) : \text{there exists } k \text{ with } T_{n_t}^k(x_{-\infty}^{-1}) \geq -t \;\text{ and }\; F_{n_t}^k(x_{-\infty}^{-1}) > \frac{p_* r_{\min}}{2|\mathcal{S}|}\right\} \;,\; t \in \mathbb{N} \\
C'_t &\equiv \left\{x_{-t-1}^{-1} \in \mathcal{L}_{t+1}(\mathcal{P}) : \text{ tail extensions of } x_{-t-1}^{-1} \text{ are in } C_t\right\} \;,\; t \in \mathbb{N} \\
G_t &\equiv \left\{x_0^t \in \mathcal{L}_{t+1}(\mathcal{P}) : \text{ there exists } k \text{ with } N_k(x_0^t) \geq \eta t\right\} \;,\; t \in \mathbb{N}
\end{aligned}$$

where $\mathcal{P} = (X_t)_{t \in \mathbb{Z}}$ is the stationary output process for M.

Note that C'_t and G_t may both be considered simply as sets of length-$(t+1)$ words w. For C'_t there is the added interpretation of the words as length-$(t+1)$ past sequences, and for G_t there is the added interpretation of the words as length-$(t+1)$ future sequences. However, C'_t and G_t are both well defined simply as sets of words. As such, we may reasonably ask questions about when one set is contained in another and the probability of these sets as the combined (stationary) probability of all words in the sets.

Lemma 26. *For any $n \in \mathbb{N}$:*

(i) $\mathbb{P}(A_n^k) \leq \alpha_1^n$*, for each* k

(ii) $\mathbb{P}(B_n) \geq 1 - |\mathcal{S}|\alpha_1^n$

Proof. (i) Follows from (35) and Lemma 25. (ii) follows from (i) and the fact that the complement of the set B_n is $B_n^c = \bigcup_k A_n^k$. (Note: Complement here means complement with respect to the set of sequences $\mathcal{L}_\infty^-(\mathcal{P})$, i.e. $B_n \cup B_n^c = \mathcal{L}_\infty^-(\mathcal{P})$. The combined probability of all other sequences is 0, so they can be ignored.) □

Lemma 27. $C_t \supseteq B_{n_t}$.

Proof. For any $x_{-\infty}^{-1} \in \mathcal{L}_\infty^-(\mathcal{P})$ there is some k such that $T_{n_t}^k(x_{-\infty}^{-1}) \geq -t$. And, for any $x_{-\infty}^{-1} \in B_{n_t}$ with $T_{n_t}^k(x_{-\infty}^{-1}) \geq -t > -\infty$, we have $F_{n_t}^k(x_{-\infty}^{-1}) > \frac{p_* r_{\min}}{2|\mathcal{S}|}$, which implies $x_{-\infty}^{-1} \in C_t$. □

Lemma 28. $G_t \supseteq C'_t$.

Proof. Let $w \in C'_t$. Then there exists k such that:

$$T_{n_t}^k(x_{-\infty}^{-1}) \geq -t \;\text{ and }\; F_{n_t}^k(x_{-\infty}^{-1}) > \frac{p_* r_{\min}}{2|\mathcal{S}|}$$

for any tail extension $x_{-\infty}^{-1}$ of w in $\mathcal{L}_\infty^-(\mathcal{P})$. It follows that:

$$N_k(w) > n_t \cdot \frac{p_* r_{\min}}{2|\mathcal{S}|} \geq \left(\frac{p_* r_{\min}}{2|\mathcal{S}|^2}\right) t = \eta t$$

which implies $w \in G_t$. □

Lemma 29. *For any $t \in \mathbb{N}$, $\mathbb{P}(G_t) \geq 1 - |\mathcal{S}|\alpha_2^t$, where $\alpha_2 \equiv \alpha_1^{1/|\mathcal{S}|} < 1$.*

Proof. By Lemmas 26, 27, and 28 we have:

$$\mathbb{P}(G_t) \geq \mathbb{P}(C_t') = \mathbb{P}(C_t) \geq \mathbb{P}(B_{n_t}) \geq 1 - |\mathcal{S}|\alpha_1^{n_t} \geq 1 - |\mathcal{S}|\alpha_2^t$$

□

6.2.3 Pair Chain Coupling Bound

Assume now, without loss of generality, that the state set is $\mathcal{S} = \{1, 2, ..., |\mathcal{S}|\}$, and for $x_0^t \in G_t$ let the index $l(x_0^t)$ and time set $\Gamma(x_0^t)$ be defined by:

$$l \equiv \min\{k : N_k(x_0^t) \geq \eta t\}$$
$$\Gamma \equiv \{0 \leq \tau \leq t-1 : x_\tau = y_l, \text{ and } \mathbb{P}_l(x_{\tau+1}^t) \geq \mathbb{P}_j(x_{\tau+1}^t) \text{ for all } j\}$$

Lemma 30. *For any $x_0^t \in G_t$, $\tau \in \Gamma(x_0^t)$, and $i \in \mathcal{S}(x_\tau^t)$:*

$$\mathbb{P}(S_{\tau+1} = l | S_\tau = i, X_\tau^t = x_\tau^t) \geq q_*$$

Proof. Let $x_0^t \in G_t$ and $\tau \in \Gamma(x_0^t)$. Then, for any $i \in \mathcal{S}(x_\tau^t)$ we must have have:

$$\mathbb{P}(S_\tau = i, X_\tau = y_l) = \mathbb{P}(S_\tau = i, X_\tau = x_\tau) > 0$$

and:

$$\mathbb{P}(X_{\tau+1}^t = x_{\tau+1}^t | S_\tau = i, X_\tau = y_l) \leq \mathbb{P}(X_{\tau+1}^t = x_{\tau+1}^t | S_{\tau+1} = l)$$

Thus:

$$\begin{aligned}
&\mathbb{P}(S_{\tau+1} = l | S_\tau = i, X_\tau^t = x_\tau^t) \\
&\quad = \frac{\mathbb{P}(S_{\tau+1} = l, S_\tau = i, X_\tau = y_l, X_{\tau+1}^t = x_{\tau+1}^t)}{\mathbb{P}(S_\tau = i, X_\tau = y_l, X_{\tau+1}^t = x_{\tau+1}^t)} \\
&\quad = \frac{\mathbb{P}(S_\tau = i, X_\tau = y_l) \cdot \mathbb{P}(S_{\tau+1} = l | S_\tau = i, X_\tau = y_l) \cdot \mathbb{P}(X_{\tau+1}^t = x_{\tau+1}^t | S_{\tau+1} = l)}{\mathbb{P}(S_\tau = i, X_\tau = y_l) \cdot \mathbb{P}(X_{\tau+1}^t = x_{\tau+1}^t | S_\tau = i, X_\tau = y_l)} \\
&\quad \geq \mathbb{P}(S_{\tau+1} = l | S_\tau = i, X_\tau = y_l) \\
&\quad \geq q_*
\end{aligned}$$

□

Lemma 31. *For any $x_0^t \in G_t$ and states $k, \widehat{k} \in \mathcal{S}(x_0^t)$:*

$$\|\phi_k(x_0^t) - \phi_{\widehat{k}}(x_0^t)\|_{TV} \leq \alpha_3^t$$

where $\alpha_3 \equiv (1 - q_*^2)^\eta < 1$.

Proof. Applying the pair chain coupling bound (34) we have:

$$\begin{aligned}
\|\phi_k(x_0^t) - \phi_{\widehat{k}}(x_0^t)\|_{TV} &\leq \widehat{\mathbb{P}}(R_{t+1} \neq \widehat{R}_{t+1}) \\
&= \prod_{\tau=0}^{t} \widehat{\mathbb{P}}\left(R_{\tau+1} \neq \widehat{R}_{\tau+1} | R_\tau \neq \widehat{R}_\tau\right) \\
&\leq \prod_{\tau \in \Gamma(x_0^t)} \widehat{\mathbb{P}}\left(R_{\tau+1} \neq \widehat{R}_{\tau+1} | R_\tau \neq \widehat{R}_\tau\right) \\
&\leq \prod_{\tau \in \Gamma(x_0^t)} \max_{i, \widehat{i} \in \mathcal{S}(x_\tau^t), i \neq \widehat{i}} \widehat{\mathbb{P}}\left(R_{\tau+1} \neq \widehat{R}_{\tau+1} | R_\tau = i, \widehat{R}_\tau = \widehat{i}\right) \\
&\leq \prod_{\tau \in \Gamma(x_0^t)} \max_{i, \widehat{i} \in \mathcal{S}(x_\tau^t), i \neq \widehat{i}} \left(1 - \widehat{\mathbb{P}}\left(R_{\tau+1} = \widehat{R}_{\tau+1} = l | R_\tau = i, \widehat{R}_\tau = \widehat{i}\right)\right) \\
&\overset{(a)}{\leq} \prod_{\tau \in \Gamma(x_0^t)} (1 - q_*^2) \\
&\overset{(b)}{\leq} (1 - q_*^2)^{\eta t} \\
&= \alpha_3^t
\end{aligned}$$

Here (a) follows from Lemma 30 and (b) from the fact that $|\Gamma(x_0^t)| \geq \eta t$. (Note: We have assumed in the proof that $\widehat{\mathbb{P}}(R_\tau \neq \widehat{R}_\tau) > 0$, for all $0 \leq \tau \leq t$. If this is not the case, then the conclusion follows trivially since $\widehat{\mathbb{P}}(R_{t+1} \neq \widehat{R}_{t+1}) = 0$.) □

6.2.4 Convergence of the Entropy Rate Estimates

Lemma 32. *For any two probability distributions μ, ν on $\mathcal{S}$:*

$$\|\mathbb{P}_\mu(X_0) - \mathbb{P}_\nu(X_0)\|_{TV} \leq \|\mu - \nu\|_{TV}$$

Proof. It is equivalent to prove the statement for 1-norms. In this case we have:

$$\begin{aligned}\|\mathbb{P}_\mu(X_0) - \mathbb{P}_\nu(X_0)\|_1 &= \left\|\sum_k \mu_k \cdot \mathbb{P}_k(X_0) - \sum_k \nu_k \cdot \mathbb{P}_k(X_0)\right\|_1 \\ &\leq \sum_k |\mu_k - \nu_k| \cdot \|\mathbb{P}_k(X_0)\|_1 \\ &= \|\mu - \nu\|_1\end{aligned}$$

□

Lemma 33. *Let $t_0 = t_0(\alpha_3) \equiv \lceil \log_{\alpha_3}(1/e) \rceil$. Then for each $t \geq t_0$ and $x_0^t \in G_t$:*

$$H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|S_0, X_0^t = x_0^t) \leq |\mathcal{X}| \cdot \alpha_3^t \cdot \log_2(1/\alpha_3^t)$$

Proof. Fix $x_0^t \in G_t$ with $t \geq t_0$. Then for each $k \in \mathcal{S}(x_0^t)$ we have by Lemmas 31 and 32:

$$\begin{aligned}&\|\mathbb{P}_k(X_{t+1}|X_0^t = x_0^t) - \mathbb{P}(X_{t+1}|X_0^t = x_0^t)\|_{TV} \\ &\qquad \leq \|\mathbb{P}_k(S_{t+1}|X_0^t = x_0^t) - \mathbb{P}(S_{t+1}|X_0^t = x_0^t)\|_{TV} \\ &\qquad \leq \max_{\widehat{k} \in \mathcal{S}(x_0^t)} \|\mathbb{P}_k(S_{t+1}|X_0^t = x_0^t) - \mathbb{P}_{\widehat{k}}(S_{t+1}|X_0^t = x_0^t)\|_{TV} \\ &\qquad = \max_{\widehat{k} \in \mathcal{S}(x_0^t)} \|\phi_k(x_0^t) - \phi_{\widehat{k}}(x_0^t)\|_{TV} \\ &\qquad \leq \alpha_3^t\end{aligned}$$

Thus, by Lemma 3:

$$H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|X_0^t = x_0^t, S_0 = k) \leq |\mathcal{X}| \cdot \alpha_3^t \cdot \log_2(1/\alpha_3^t)$$

for all $k \in \mathcal{S}(x_0^t)$, since $t \geq t_0$. The claim follows since:

$$\begin{aligned}&H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|S_0, X_0^t = x_0^t) \\ &\qquad = \sum_{k \in \mathcal{S}(x_0^t)} \mathbb{P}(S_0 = k|X_0^t = x_0^t) \cdot \big(H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|X_0^t = x_0^t, S_0 = k)\big)\end{aligned}$$

□

Theorem 3. *The entropy rate block estimates $h(t)$ converge exponentially with:*

$$\limsup_{t\to\infty} \{h(t) - h\}^{1/t} \leq \alpha$$

where $\alpha \equiv \max\{\alpha_2, \alpha_3\}$.

Proof. For any $t \in \mathbb{N}$:

$$\begin{aligned} h &= \lim_{\tau\to\infty} H(X_0|X_{-\tau}^{-1}) \\ &\geq \lim_{\tau\to\infty} H(X_0|X_{-\tau}^{-1}, S_{-(t+1)}) \\ &= H(X_0|X_{-(t+1)}^{-1}, S_{-(t+1)}) \\ &= H(X_{t+1}|X_0^t, S_0) \end{aligned}$$

Thus, using Lemmas 29 and 33, the difference $h(t+2) - h$ may be bounded as follows for all $t \geq t_0 = t_0(\alpha_3)$:

$$\begin{aligned} h(t+2) - h &= H(X_{t+1}|X_0^t) - h \\ &\leq H(X_{t+1}|X_0^t) - H(X_{t+1}|X_0^t, S_0) \\ &= \sum_{x_0^t \in \mathcal{L}_{t+1}(M)} \mathbb{P}(x_0^t) \cdot \left[H(X_{t+1}|X_0^t = x_0^t) - H(X_{t+1}|X_0^t = x_0^t, S_0)\right] \\ &\leq \mathbb{P}(G_t) \cdot \left(|\mathcal{X}| \cdot \alpha_3^t \cdot \log_2(1/\alpha_3^t)\right) + \mathbb{P}(G_t^c) \cdot \log_2|\mathcal{X}| \\ &\leq 1 \cdot \left(|\mathcal{X}| \cdot \alpha_3^t \cdot \log_2(1/\alpha_3^t)\right) + |\mathcal{S}|\alpha_2^t \cdot \log_2|\mathcal{X}| \end{aligned}$$

The claim follows directly from this inequality. □

6.3 Bounds for General FS Models

The following theorem gives an upper bound on the rate of convergence of the entropy rate estimates $h(t)$ for a FS HMM M. It is an immediate consequence of Theorem 3 and Lemma 8.

Theorem 4. *Let M be an irreducible, FS HMM, and let M^n be the irreducible, FS1 block model with $n \in \mathbb{N}$ given by (18). Also, let the constant $\alpha < 1$ be defined as in Theorem 3 for the FS1 block model M^n. Then the entropy rate estimates $h(t)$ for the HMM M converge exponentially with:*

$$\limsup_{t \to \infty} \{h(t) - h\}^{1/t} \leq \alpha^{1/n}$$

7 Conclusion

There is, in general, no closed form expression for the entropy rate of a hidden Makov model. However, the entropy rate h can often be well approximated by the finite length block estimates $h(t)$, and the speed of convergence of these estimates is of independent interest as well, even in cases where the entropy rate can be calculated analytically, such as for unifilar HMMs.

In particular, the rate of convergence of the estimates $h(t)$ is important for numerical approximations of the excess entropy [13] and for an observer making predictions from finite data. It is also closely related to the rate at which the HMM "forgets its initial distribution", a problem that has been investigated extensively [17, 34–39], though primarily in the case of an $\mathbb{R}$ or $\mathbb{R}^n$ valued output sequence (X_t) and often with a more general state space $\mathcal{S}$ as well.

In the finite case, it is well known that the estimates $h(t)$ converge exponentially for state-emitting HMMs if the underlying Markov chain has strictly positive transition probabilities [14]. Similarly, in [16] exponential convergence is demonstrated for finite HMMs (both state-emitting and edge-emitting) with the assumption of strictly positive symbol emission probabilities and an aperiodic underlying Markov chain. However, no exponential bounds have been previously established for the rate of convergence of the block estimates in all finite HMMs. The weakest previously used assumptions come from [18], but these still require the output process (X_t) to have full support.

We have demonstrated, here, exponential convergence not for all finite HMMs, but under a number of alternative assumptions that do not require positivity of either the state transition matrix or symbol emission probabilities and also do not require the output process to have full support. The first two conditions we studied, exactness and unifilarity, are fairly strong, but we believe, as discussed in the introduction, that they are natural assumptions to make and interesting cases to consider. The flag-state condition is not as natural, but it is the most widely applicable, and, when translated to state emitting HMMs, is strictly weaker than either of the positivity assumptions used in [14] and [16]. These ideas will be discussed further and made more precise in Appendices B and C, respectively.

Remark. *The aperiodicity assumption in [16] is not essential. Exponential convergence for aperiodic HMMs with positive symbol emission probabilities could be used to show exponential convergence for all HMMs with positive symbol emission probabilities. Similarly, exponential convergence will still hold for a HMM that is "flag-state up to periodicity".*

Appendices

A Property Testing

In this section we provide tests for several important HMM properties, including the flag-state property, and whether a given pair of states (i, j) is (a) path-mergeable, (b) incompatible, or (c) probabilistically distinguishable, in the case that the HMM is unifilar. Throughout, $\overline{\delta}_i(w)$ denotes the unifilar HMM transition function defined by (23), and for a HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ and states $i, j \in \mathcal{S}$ we denote the minimum path-merge length, minimum incompatibility length, and minimum probabilistically distinguishing length, respectively, by ℓ^1_{ij}, ℓ^2_{ij}, and ℓ^3_{ij}:

$$\ell^1_{ij} \equiv \inf\{n : \exists w \in \mathcal{L}_n(M) \text{ and } k \in \mathcal{S} \text{ with } k \in \delta_i(w) \cap \delta_j(w)\} \qquad (\ell^1_{ij} \equiv \infty, \text{ if no such } n)$$

$$\ell^2_{ij} \equiv \inf\{n : \mathcal{L}_n(i) \cap \mathcal{L}_n(j) = \emptyset\} \qquad (\ell^2_{ij} \equiv \infty, \text{ if no such } n)$$

$$\ell^3_{ij} \equiv \inf\{n : \mathbb{P}_i(X_0^{n-1}) \neq \mathbb{P}_j(X_0^{n-1})\} \qquad (\ell^3_{ij} \equiv \infty, \text{ if no such } n)$$

Note that, by definition, a state pair (i, j) is path-mergeable if and only if $\ell^1_{ij} < \infty$. Similarly, the pair (i, j) is incompatible if and only if $\ell^2_{i,j} < \infty$ and probabilistically distinguishable if and only if $\ell^3_{i,j} < \infty$.

Testing if $\ell^*_{i,j} = 1$ ($* = 1$, 2, or 3) for any state pair (i, j) can be done directly by examining the 1 symbol transition probabilities $\mathcal{T}^{(x)}_{ik}$, $\mathcal{T}^{(x)}_{jk}$ for each $x \in \mathcal{X}, k \in \mathcal{S}$. If $\ell^*_{i,j} \neq 1$ then determining if $\ell^*_{ij} = n$ for some $1 < n < \infty$ is more difficult, but it can be accomplished using modified versions of the standard table filling algorithm for deterministic finite automata (DFAs) [40], as presented below.

Algorithm 1. *Determine Path-Mergeable State Pairs for HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$*

Initialization Step

- *Create a table with boxes for each pair of distinct states (i, j).*
  Initially all boxes are unmarked.

- *Then, for each pair (i, j) check if $\ell^1_{ij} = 1$. If so, mark box for pair (i, j).*

Inductive Step

- *If for some i, j, x there exist states $i' \in \delta_i(x), j' \in \delta_j(x)$ such that the box for pair (i', j') is already marked, then mark the box for pair (i, j).*

- *Repeat until no more boxes can be marked this way.*

Algorithm 2. *Determine Incompatible State Pairs for HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$*

<u>*Initialization Step*</u>

- *Create a table with boxes for each pair of distinct states (i, j).*

  Initially all boxes are unmarked.

- *Then, for each pair (i, j) check if $\ell_{ij}^2 = 1$. If so, mark box for pair (i, j).*

<u>*Inductive Step*</u>

- *If there exist states i, j such that for each $x \in \mathcal{X}$ and $i' \in \delta_i(x)$, $j' \in \delta_j(x)$ the box for pair (i', j') is already marked, then mark the box for pair (i, j).*

- *Repeat until no more boxes can be marked this way.*

Algorithm 3. *Determine Probabilistically Distinguishable State Pairs for Unifilar HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$*

<u>*Initialization Step*</u>

- *Create a table with boxes for each pair of distinct states (i, j).*

  Initially all boxes are unmarked.

- *Then, for each pair (i, j) check if $\ell_{ij}^3 = 1$. If so, mark box for pair (i, j).*

<u>*Inductive Step*</u>

- *If for some i, j, x the box for pair $(\overline{\delta}_i(x), \overline{\delta}_j(x))$ is already marked, then mark the box for pair (i, j).*

- *Repeat until no more boxes can be marked this way.*

By induction on n, it can be shown that for each $n \in \mathbb{N}$ all state pairs (i, j) with $\ell_{ij}^1 = n$ end up with marked boxes under Algorithm 1. Thus, all path-mergeable state pairs (i, j), i.e. pairs with $\ell_{ij}^1 < \infty$, end up with marked boxes under this algorithm. On the other hand, a box never becomes marked under Algorithm 1 unless the pair (i, j) is path-mergeable. Thus, under Algorithm 1 the box for pair (i, j) ends up marked if and only if the pair (i, j) is path-mergeble.

Similarly, one can show, by induction, that under Algorithm 2 a pair (i, j) ends up with a marked box if and only if the pair (i, j) is incompatible, and under Algorithm 3 (operating on a unifilar HMM) that a pair (i, j) ends up with a marked box if and only if the pair (i, j) is probabilistically distinguishable. Thus, these three algorithms provide ways to determine all the state pair properties of interest.

In light of Proposition 8 in Appendix B we also now have the following one way test to see if a given (irreducible) HMM is a flag-state HMM:

1. Determine all path-mergeable state pairs with Algorithm 1.

2. Determine all incompatible state pairs with Algorithm 2.

3. If each state pair (i, j) satisfies at least one of these two properties, then the HMM is a flag-state HMM. If not, the test is inconclusive.

The three algorithms presented above, and thus the flag-state one way test, are all “reasonably fast”, at least if the inductions in the second step are carried out in an efficient manner. For example, as described in [40] for the DFA table filling algorithm, with small modifications as necessary. In particular, the running times are all polynomial in the number of states and symbols.

However, the one way flag-state test is not definitive in the negative direction, and we are unaware of a similar modification of the DFA table filling algorithm that gives a definitive answer for whether a given HMM is flag-state or not. Instead, we present below an auxiliary construction, which we call simply the *flag-state test machine*, that allows one to test definitively if a given HMM is a flag-state HMM. Unfortunately, though, the machine M_{test} itself can be very large, even if the original HMM is reasonably sized. In particular, its state set $\mathcal{R}$ has $2^{(|\mathcal{S}|^2)}$ states, where $\mathcal{S}$ is the state set of the original HMM M. In general, not all states of M_{test} will be accessible, and it is only necessary to construct the restricted graph consisting of accessible states. Nevertheless, there still may be very many accessible states, so the construction and analysis may take quite some time.

Definition 10. *For an irreducible HMM $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ the* flag-state test machine M_{test} *is the unifilar HMM $(\mathcal{R}, \mathcal{X}, \{\mathcal{Q}^{(x)}\})$ with state set:*

$$\mathcal{R} \equiv \{r = (A_1, ..., A_{|\mathcal{S}|}) : A_n \textit{ is a subset of } \mathcal{S}, \textit{ for each } n = 1, ..., |\mathcal{S}|\}$$

and unifilar transition function:

$$\overline{\delta}_r^{test}(w) \equiv (\delta_{A_1}(w), ..., \delta_{A_{|\mathcal{S}|}}(w))$$

where $\delta_A(w)$ is the standard transition function for the original HMM M given by (7). Transition probabilities are, by convention[19]*, taken to be uniform over all outgoing edges from each state:*

$$\mathcal{Q}_{rr'}^{(x)} \equiv \frac{1}{|\mathcal{X}|} \cdot \mathbb{1}\left\{\overline{\delta}_r^{test}(x) = r'\right\}$$

For notational convenience, let us assume that the state set of M is $\mathcal{S} = \{1, ..., |\mathcal{S}|\}$. Also, let:

$$r_0 \equiv (\{1\}, ..., \{|\mathcal{S}|\})$$

and, for $w \in \mathcal{X}^*$, let $r = r(w)$ be defined by:

$$r = (A_1, ..., A_{|\mathcal{S}|}) \equiv \overline{\delta}_{r_0}^{test}(w) = (\delta_i(w), ..., \delta_{|\mathcal{S}|}(w))$$

Then w is a flag word for state k if and only if:

(a) For each $i \in \mathcal{S}$, either $k \in A_i$ or $A_i = \emptyset$, and

(b) There exists some $i \in \mathcal{S}$ with $A_i \neq \emptyset$.

Thus, state k has a flag word if and only if there is some $r = (A_1, ..., A_{|\mathcal{S}|}) \in \mathcal{R}$ that is accessible from r_0 and satisfies (a) and (b). To test if M is a flag-state HMM one can, therefore, restrict the graph of M_{test} to r_0 and those states accessible from r_0, and then for each $k \in \mathcal{S}$ examine all states r in the restricted graph to make sure there is some r that satisfies (a) and (b). If for each $k \in \mathcal{S}$ there is some such r then the HMM is a flag-state HMM, otherwise it is not. In fact, as noted

[19] Transition probabilities of the allowed transitions are unimportant. It is only the topology of the test machine M_{test} that is relevant for deciding if M is a flag-state HMM.

previously in Section 2.5.4, if any state k has a flag word w then all of them do (by irreducibility). Thus, if suffices to check that there is some r satisfying (a) and (b) for a single state $k \in \mathcal{S}$.

Of course, once it is determined that the HMM is a flag-state HMM, it is also easy to determine flag words for it in order to apply the bounds given in Section 6. For a state r satisfying (a) and (b) for a given k, if we take any path in the graph from r_0 to r and let w be the word defined by this path then w is a flag word for k. Generally, better bounds will tend to result from shorter paths, but finding an ideal path may also require some searching.

An example of the M_{test} construction is presented below in Figure 8.

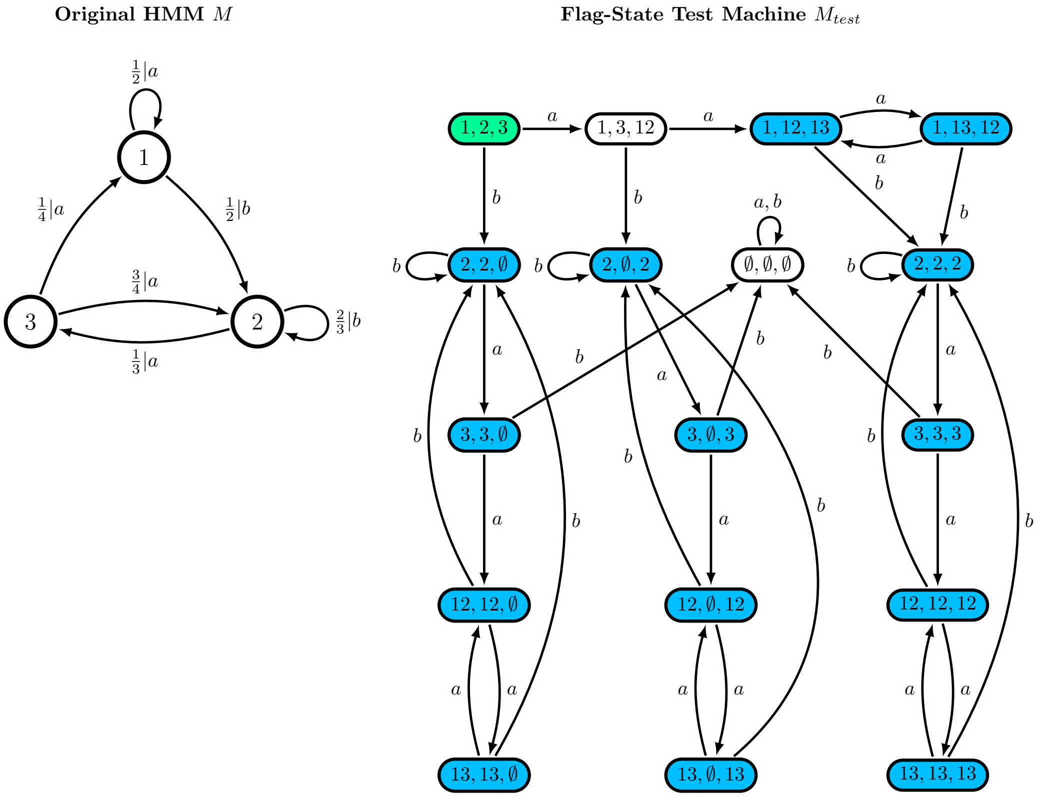


Figure 8: The flag-state test machine construction for a 3 state, 2 symbol HMM. In the test machine M_{test} state names have been abbreviated by dropping the outer parentheses, set brackets, and commas between elements of each set. Thus, for example, the state $(\{1\},\{1,2\},\{1,3\})$ is denoted simply by $1,12,13$, and the state $(\{1,3\},\{1,3\},\emptyset)$ is denoted $13,13,\emptyset$. Also, transition probabilities have been omitted from the edge labels of the test machine, because they are irrelevant, and the graph of the test machine has been restricted to the state r_0 and other states accessible from it, since these are all that matter for determining if M is a flag-state HMM. For ease of recognition, r_0 is colored in green, and each state r satisfying conditions (a) and (b) for some $k \in \mathcal{S}$ is colored in blue, while all other states are colored white. Since there are indeed blue states in this restricted graph of M_{test} we know that M is a flag-state HMM.

B Weakness of the Flag-State Condition

Proposition 8 below states than an irreducible HMM M is a flag-state HMM if each pair of distinct states i, j is either path-mergeable or incompatible. This is not a necessary condition for a HMM to be flag-state, but it is a fairly weak sufficient condition.

In particular, if the number of symbols m and states n are both large (or even if the number of states n is large and the alphabet is binary, $m = 2$) and a HMM topology is selected uniformly at random from all topologies with n states and m symbols, then with high probability the HMM will have a significant degree of branching or non-unifilarity. That is, most states will have at least one symbol upon which they can transition to many (i.e. $\approx n/2$) other states. If this is the case, then most state pairs should be path-mergeable. The most likely way for a given state pair i, j to not be path-mergeable (at least if m is small) is that i and j have no common output symbol, or possibly a common output symbol but no common words of some short length. If this is the case, though, the state pair i, j is incompatible. Thus, when n is large (and especially when n and m are both large) with high probability each pair i, j should be either path-mergeable or incompatible.

Proposition 8. *If $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ is an irreducible HMM and each pair of distinct states $i, j \in \mathcal{S}$ is either path-mergeable or incompatible, then M is a flag-state HMM.*

Proof. We will explicitly construct a word v^*, which is a flag word for some state i^*. As noted previously, irreducibility then implies that the HMM M is indeed a flag-state HMM.

To construct the word v^* we need to introduce a little notation. If a pair of states i, j is path-mergeable we denote by w_{ij} and k_{ij} the special word w and state k in the definition of path-mergeability, so that $k_{ij} \in \delta_i(w_{ij}) \cap \delta_j(w_{ij})$. Also, for a state i we denote by $\mathcal{M}(i)$ the set of all states that are path-mergeable with i. Finally, for notational convenience, we assume, without loss of generality, that the state set is $\mathcal{S} = \{1, 2, ..., |\mathcal{S}|\}$ and the alphabet is $\mathcal{X} = \{1, 2, ..., |\mathcal{X}|\}$.

v^* is then defined inductively by the following algorithm:

```
(i) $t := 0$, $v_0 := \lambda$ (the empty word), $i_0 := 1$, $\mathcal{R} := \mathcal{S}/\{1\}$
(ii) While $\mathcal{R} \neq \emptyset$ do:
        $k_t := \min\{k : k \in \mathcal{R}\}$
        If $\delta_{k_t}(v_t) \cap \mathcal{M}(i_t) \neq \emptyset$:
```

```
        $j_t := \min\{j : j \in \delta_{k_t}(v_t) \cap \mathcal{M}(i_t)\}$
        $w_t := w_{i_t j_t}$
        $v_{t+1} := v_t w_t$
        $i_{t+1} := k_{i_t j_t}$
    Else:
        $w_t := \min\{x : \mathbb{P}_{i_t}(x) > 0\}$
        $v_{t+1} := v_t w_t$
        $i_{t+1} := \min\{i : i \in \delta_{i_t}(w_t)\}$
    $\mathcal{R} := \mathcal{R} / \left(\{k : i_{t+1} \in \delta_k(v_{t+1})\} \cup \{k : \mathbb{P}_k(v_{t+1}) = 0\}\right)$
    $t := t + 1$
(iii) $v^* := v_t, i^* := i_t$
```

If $\delta_{k_t}(v_t) \cap \mathcal{M}(i_t) = \emptyset$ then, by assumption, each $j \in \delta_{k_t}(v_t)$ is incompatible with i_t. So, in this case, we will always have $\mathbb{P}_{k_t}(v_{t+1}) = 0$ with our choice of w_t. On the other hand, if $\delta_{k_t}(v_t) \cap \mathcal{M}(i_t) \neq \emptyset$, then by our choice of w_t we will always have $i_{t+1} \in \delta_{k_t}(v_{t+1})$. Thus, the state k_t is always removed from the set $\mathcal{R}$ in each iteration of the loop, so the algorithm must terminate after a finite number of steps.

Now, let us assume that the loop terminates at time t with word $v^* = w_0 w_1 ... w_{t-1}$ and state $i^* = i_t$. Then, by the construction, we have $i_0 = 1$ and $i_{\tau+1} \in \delta_{i_\tau}(w_\tau)$ for each $\tau = 0, 1, ..., t-1$, which implies $i^* \in \delta_1(v^*)$, and, hence, that $v^* \in \mathcal{L}(M)$.

Further, since each state $k \neq 1$ must be removed from the list $\mathcal{R}$ before the algorithm terminates, we know that for each $k \neq 1$ one of the following must hold:

1. $i_{\tau+1} \in \delta_k(v_{\tau+1})$, for some $0 \leq \tau \leq t-1$, which implies $i^* \in \delta_k(v^*)$.
2. $\mathbb{P}_k(v_{\tau+1}) = 0$, for some $0 \leq \tau \leq t-1$, which implies $\mathbb{P}_k(v^*) = 0$.

It follows that v^* is a flag word for state i^*. □

C State-Emitting vs. Edge-Emitting HMMs

In this section we show how the results derived in this dissertation for edge-emitting HMMs apply to the more commonly used state-emitting HMMs. We begin in Sections C.1 and C.2 with the state-emitting definition and standard model type equivalence. Then in Section C.3 we show how our edge-emitting results translate to the state-emitting case.

C.1 State-Emitting Definition

Definition 11. *A (finite)* state-emitting hidden Markov model *is a 4-tuple* $(\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ *where:*

- $\mathcal{S}$ *is a finite set of states.*
- $\mathcal{X}$ *is a finite alphabet of output symbols.*
- $\mathcal{T}$ *is an* $|\mathcal{S}| \times |\mathcal{S}|$ *stochastic state transition matrix:* $\mathcal{T}_{ij} = \mathbb{P}(S_{t+1} = j | S_t = i)$.
- $\mathcal{O}$ *is an* $|\mathcal{S}| \times |\mathcal{X}|$ *stochastic observation matrix*[20]*:* $\mathcal{O}_{ix} = \mathbb{P}(X_t = x | S_t = i)$.

The state sequence (S_t) for the state-emitting HMM $(\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ is a Markov chain with transition matrix $\mathcal{T}$, and the output sequence (X_t) has conditional distribution defined by the observation matrix $\mathcal{O}$. More precisely, for each initial state $i \in \mathcal{S}$ we have measure $\mathbb{P}_i$ on joint state-symbol sequences $(S_t, X_t)_{t \in \mathbb{Z}^+}$ defined by:

$$\mathbb{P}_i(S_0^t = s_0^t) = \mathbb{1}\{s_0 = i\} \cdot \prod_{n=0}^{t-1} \mathcal{T}_{s_n s_{n+1}}$$

$$\mathbb{P}_i(X_0^t = x_0^t | S_0^t = s_0^t) = \prod_{n=0}^{t} \mathcal{O}_{s_n x_n}$$

If the HMM $(\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ is irreducible, then we denote by $\mathbb{P}$ the (unique) stationary measure on joint sequences $(S_t, X_t)_{t \in \mathbb{Z}^+}$ given by choosing the initial state S_0 according to the stationary distribution π for the underlying Markov model $(\mathcal{S}, \mathcal{T})$: $\mathbb{P}(\cdot) \equiv \mathbb{P}_\pi(\cdot) \equiv \sum_i \pi_i \mathbb{P}_i(\cdot)$.

[20] In much of the early literature [1–3, 14], as well as some of the more recent [15], it is assumed that the observation matrix $\mathcal{O}$ is deterministic: $\mathcal{O}_{ix} = \mathbb{1}\{f(i) = x\}$, for some observation function $f : \mathcal{S} \to \mathcal{X}$. Such HMMs are sometimes referred to as *functional HMMs*. However, we will not distinguish them as such explicitly here. It can be shown that functional HMMs are equivalent to general state-emitting HMMs in the same way that edge-emitting HMMs are, as discussed below in Section C.2.

C.2 Equivalence of Model Types

Though they are different objects, state-emitting and edge-emitting HMMs are equivalent in the following sense: Given an irreducible HMM M of either of these types there exists an irreducible HMM M' of the other type such that the stationary output processes (X_t) for the two HMMs M and M' are equal in distribution. Thus, if one is only concerned with questions about the observable output process, such as the convergence of the entropy rate block estimates $h(t)$, then either the state-emitting or edge-emitting model may be used.

We recall below the standard model type conversions.

1. *State-Emitting to Edge-Emitting* - If $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ then $M' = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{'(x)}\})$, where $\mathcal{T}^{'(x)}_{ij} = \mathcal{T}_{ij}\mathcal{O}_{jx}$.

2. *Edge-Emitting to State-Emitting* - If $M = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}^{(x)}\})$ then $M' = (\mathcal{S}', \mathcal{X}, \mathcal{T}', \mathcal{O}')$, where $\mathcal{S}' = \{(i,x) : \sum_j \mathcal{T}^{(x)}_{ij} > 0\}$, $\mathcal{T}'_{(i,x)(j,y)} = \left(\mathcal{T}^{(x)}_{ij} / \sum_k \mathcal{T}^{(x)}_{ik}\right) \cdot \left(\sum_k \mathcal{T}^{(y)}_{jk}\right)$, and $\mathcal{O}'_{(i,x)y} = \mathbb{1}\{x = y\}$.

C.3 Translation of Results

For a state-emitting HMM $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ let $\delta_i(w)$ be defined, as for edge-emitting HMMs, as the set of states j, which the HMM M can transition to upon emitting w if it starts in state i [21]:

$$\delta_i(w) \equiv \{j \in \mathcal{S} : \mathbb{P}_i(X_1^{|w|} = w, \mathcal{S}_{|w|} = j) > 0\} \tag{37}$$

Also, for $A \subseteq \mathcal{S}$ let:

$$\delta_A(w) \equiv \bigcup_{i \in \mathcal{S}} \delta_i(w)$$

As for edge-emitting HMMs, we say that a state-emitting HMM $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ is:

- *unifilar* if $|\delta_i(x)| \leq 1$, for each $i \in \mathcal{S}$, $x \in \mathcal{X}$.

- *exact* if there exists some *synchronizing word* w and state $j \in \mathcal{S}$, such that $\delta_{\mathcal{S}}(w) = \{j\}$.

[21] In the edge-emitting case we had instead $\delta_i(w) = \{j \in \mathcal{S} : \mathbb{P}_i(X_0^{|w|-1} = w, \mathcal{S}_{|w|} = j) > 0\}$, but this is because of the offset that comes from using edge-emitting HMMs where the symbol X_t is generated on the transition from state S_t to state S_{t+1}, rather than being emitted directly from state S_t, as for state-emitting HMMs.

- a *flag-state HMM* if for each state k there is some *flag word* $w \in \mathcal{L}(M)$ such that $k \in \delta_i(w)$, for all i with $\delta_i(w) \neq \emptyset$.

The following lemma shows that these three properties are preserved under the state-emitting to edge-emitting conversion.

Lemma 34. *Let $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ be an irreducible, state-emitting HMM, and let $M' = (\mathcal{S}, \mathcal{X}, \{\mathcal{T}'^{(x)}\})$ be the corresponding, irreducible, edge-emitting HMM defined as in the previous section with $\mathcal{T}'^{(x)}_{ij} = \mathcal{T}_{ij}\mathcal{O}_{jx}$. Then:*

1. *M is unifilar if and only if M' is unifilar.*

2. *M is exact if and only if M' is exact.*

3. *M is a flag-state HMM if and only if M' is a flag-state HMM.*

Proof. By induction on t it is easily seen that for each $t \in \mathbb{N}$, $x_1^t \in \mathcal{X}^t$, and $s_0^t \in \mathcal{S}^{t+1}$:

$$\mathbb{P}_i(S_0^t = s_0^t, X_1^t = x_1^t) = \mathbb{P}_i'(S_0^t = s_0^t, X_0^{t-1} = x_1^t)$$

where $\mathbb{P}_i$ and $\mathbb{P}_i'$ are, respectively, the measures on state-symbol sequences $(S_t, X_t)_{t \in \mathbb{Z}^+}$ for M and M' with initial state $S_0 = i$. This implies that the transition function $\delta_i(w)$ for M defined by (37) is the same as the transition function $\delta_i'(w)$ for M' defined by (6):

$$\delta_i(w) = \delta_i'(w) \ , \text{ for each } i \in \mathcal{S}, w \in \mathcal{X}^*$$

All three claims follow immediately from the equivalence of transition functions. □

Thus, although some of these properties may be slightly less natural in the state-emitting case (particularly unifilarity), they all do have well defined analogs, and a state-emitting HMM has one of these properties if and only if the corresponding edge-emitting HMM has the same property. Of course, in practice, if one wants bound on convergence of the entropy rate estimates $h(t)$ for the output process of an irreducible, state-emitting HMM M it is possible to:

1. Convert it to the corresponding edge-emitting HMM M'.

2. Test M' for the various properties (exactness, unifilarity, flag-state property) as described previously.

3. Apply the bounds established for edge-emitting HMMs with whatever property(s) (if any) M' satisfies.

And, none of this requires any direct knowledge of the properties of the original state-emitting HMM M.

However, to compare conditions on state-emitting HMMs to the conditions studied here for edge-emitting HMMs the model type independent nature of the properties is important. In particular, observe that, on the one hand:

1. If the underlying transition matrix $\mathcal{T}$ for a state-emitting HMM $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ has strictly positive entries, as is often assumed, then for each state j and symbol x with $\mathcal{O}_{jx} > 0$:

$$j \in \delta_i(x) \ , \ \text{for each } i \in \mathcal{S}$$

Thus, such a HMM is a flag-state HMM.

2. If a state-emitting HMM $M = (\mathcal{S}, \mathcal{X}, \mathcal{T}, \mathcal{O})$ is aperiodic and the observation matrix $\mathcal{O}$ has all positive entries, then for any word $w \in \mathcal{X}^n$ and state $j \in \mathcal{S}$:

$$j \in \delta_i(w) \ , \ \text{for each } i \in \mathcal{S}$$

where:

$$n \equiv \min\{m : \mathcal{T}^m_{ij} > 0, \ \text{for all } i, j \in \mathcal{S}\} < \infty \quad \text{(by aperiodicity)}$$

Hence, such a HMM is also a flag-state HMM.

On the other hand, it is easy to construct state-emitting, flag-state HMMs (or state-emitting unifilar or state-emitting exact HMMs) for which neither the underlying transition matrix $\mathcal{T}$ or observation matrix $\mathcal{O}$ has all positive entries. So, our exponential decay bounds on the error in the estimates $h(t)$ are more widely applicable than many of those established previously, if sometimes not as strong.

References

[1] E. J. Gilbert. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 30(3):688–697, 1959.

[2] D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 28(4):1011–1015, 1957.

[3] D. Blackwell. The entropy of functions of finite-state Markov chains. In *Transactions of the first Prague conference on information theory, statistical decision functions, random processes*, pages 13–20. Publishing House of the Czechoslovak Academy of Sciences, 1957.

[4] B. H. Juang and L. R. Rabiner. Hidden Markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[5] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Proc.*, 77:257–286, 1989.

[6] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE Trans. Pattern Analysis and Machine Intelligence*, PAMI-5:179–190, 1983.

[7] F. Jelinek. Continuous speech recognition by statistical methods. *Proc. IEEE*, 64:532–536, 1976.

[8] A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comp. Bio.*, 11(2-3):413–428, 2004.

[9] S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[10] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.

[11] S. R. Eddy, G. Mitchison, and R. Durbin. Maximum discrimination hidden Markov models of sequence consensus. *J. Comp. Bio.*, 2(1):9–23, 1995.

[12] P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov models of biological primary sequence information. *PNAS*, 91:1059–1063, 1994.

[13] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: Levels of entropy convergence. *CHAOS*, 13(1):25–54, 2003.

[14] J. Birch. Approximations for the entropy for functions of Markov chains. *Ann. Math. Statist*, 33(3):930–938, 1962.

[15] B.M. Hochwald and P.R. Jelenkovic. State learning and mixing in entropy of hidden Markov processes and the Gilbert-Elliott channel. *IEEE Trans. Info. Theory*, 45(1):128–138, 1999.

[16] H. Pfister. *On the capacity of finite state channels and analysis of convolutional accumulate-m codes*. PhD thesis, University of California, San Diego, 2003.

[17] F. Le Gland and L. Mevel. Exponential forgetting and geometric ergodicity in hidden Markov models. *Math. Control Signals Systems*, 13:63–93, 2000.

[18] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Trans. Info. Theory*, 52(12):5251–5266, 2006.

[19] N. Jonoska. Sofic shifts with synchronizing presentations. *Theor. Comput. Sci.*, 158:81–115, 1996.

[20] K. Marchand, J. Mulherkar, and B. Nachtergaele. Entropy rate calculations of algebraic measures. In *ISIT*, pages 1072–1076. 2012.

[21] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105–108, 1989.

[22] C. R. Shalizi and J. P. Crutchfield. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.*, 104:817–879, 2001.

[23] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, Hoboken, NJ, second edition, 2006.

[24] D. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times.* American Mathematical Society, Providence, RI, 2006.

[25] B. Weiss. Subshifts of finite type and sofic systems. *Monatsh. Math.*, 77:462–474, 1973.

[26] N. F. Travers and J. P. Crutchfield. Exact synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1181–1201, 2011.

[27] N. F. Travers and J. P. Crutchfield. Asymptotic synchronization for finite-state sources. *J. Stat. Phys.*, 145(5):1202–1223, 2011.

[28] P.W. Glynn and D. Ormoneit. Hoeffding's inequality for uniformly ergodic Markov chains. *Stat. Prob. Lett.*, 56:143–146, 2002.

[29] A. Dembo and O. Zeitouni. *Large Deviations - Techniques and Applications.* Springer Verlag, New York, NY, 1998.

[30] S. Still, J.P. Crutchfield, and C.J. Ellison. Optimal causal inference: Estimating stored information and approximating causal architecture. *CHAOS*, 20, 2010.

[31] Ben Morris. Personal communication.

[32] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[33] J. L. Massey. Markov information sources. In J. K. Skwirzynski, editor, *New Directions in Signal Processing in Communication and Control*, volume E25 of *NATO Advanced Study Institutes Series*, pages 15–26. Noordhoff, Leyden, The Netherlands, 1975.

[34] R. B. Sowers and A. M. Makowski. Discrete-time filtering for linear systems in correlated noise with non-gaussian initial conditions: Formulas and asymptotics. *IEEE Trans. Automat. Control*, 37:114–121, 1992.

[35] R. Atar and O. Zeitouni. Lyapunov exponents for finite state nonlinear filtering. *SIAM J. Control Optim.*, 35:36–55, 1995.

[36] R. Atar and O. Zeitouni. Exponential stability for nonlinear filtering. *Ann. Inst. H. Poincare Prob. Statist.*, 33(6):697–725, 1997.

[37] P. Chigansky and R. Lipster. Stability of nonlinear filters in nonmixing case. *Ann. App. Prob.*, 14(4):2038–2056, 2004.

[38] R. Douc, G. Fort, E. Moulines, and P. Priouret. Forgetting the initial distribution for hidden Markov models. *Stoch. Proc. App.*, 119(4):1235–1256, 2009.

[39] P. Collet and F. Leonardi. Loss of memory of hidden Markov models and Lyapunov exponents. *arXiv/0908.0077*, 2009.

[40] J. E. Hopcroft, R. Motwani, and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation.* Addison-Wesley, second edition, 2001.