Adjusted Viterbi training for hidden Markov models

Technical Report 07-01

School of Mathematical Sciences, Nottingham University, UK

Jüri Lember

Tartu University, Liivi 2-507, Tartu 50409, Estonia; jyril@ut.ee

Alexey Koloydenko*

School of Mathematical Sciences. Division of Statistics.
University of Nottingham. University Park. Nottingham NG7 2RD, UK.
Tel: +44(0)115.951.4937, alexey.koloydenko@nottingham.ac.uk

April, 24, 2006

Abstract

To estimate the emission parameters in hidden Markov models one commonly uses the EM algorithm or its variation. Our primary motivation, however, is the Philips speech recognition system wherein the EM algorithm is replaced by the Viterbi training algorithm. Viterbi training is faster and computationally less involved than EM, but it is also biased and need not even be consistent. We propose an alternative to the Viterbi training – adjusted Viterbi training – that has the same order of computational complexity as Viterbi training but gives more accurate estimators. Elsewhere, we studied the adjusted Viterbi training for a special case of mixtures, supporting the theory by simulations. This paper proves the adjusted Viterbi training to be also possible for more general hidden Markov models.

Keywords: Consistency; EM algorithm; hidden Markov models; parameter estimation; Viterbi Training

1 Introduction

We consider a set of procedures to estimate the emission parameters of a finite state hidden Markov model given observations $x_1, \dots, x_n$. Thus, Y is a Markov chain with (finite) state space S, transition matrix $\mathbb{P} = (p_{ij})$, and initial distribution π. To every state $l \in S$ there corresponds an emission distribution P_l with density f_l that is known up to the parametrization $f_l(x; \theta_l)$. When Y reaches state l, an observation according to P_l and independent of everything else, is emitted.

The standard method for finding the maximum likelihood estimator of the emission parameters θ_l is the EM-algorithm that in the present context is also known as the *Baum-Welch* or *forward-backward algorithm* [1, 2, 8, 9, 18, 19]. Since the EM-algorithm can in practice be slow and computationally expensive, one seeks reasonable alternatives. One such alternative is *Viterbi training* (VT). VT is used in speech recognition [8, 15, 19, 20, 21, 22], natural language modeling [16], image analysis [14], bioinformatics [5, 17]. We are also motivated by connections with constrained vector quantization [4, 6]. The basic idea behind VT is to replace the computationally costly expectation (E) step of the EM-algorithm by an appropriate maximization step with fewer and simpler computations. In speech recognition, essentially the same training procedure was already described by L. Rabiner *et al.* in [10, 20] (see also [18, 19]). Rabiner considered this procedure as a variation of the *Lloyd algorithm* used in vector quantization, referring to Viterbi training as the *segmential K-means training*. The analogy with the vector quantization is especially pronounced when the underlying chain is simply a sequence of *i.i.d.* variables, observations on which are consequently an *i.i.d.* sample from a mixture distribution. For such mixture models, VT was also described by R. Gray *et al.* in [4], where the training algorithm was considered in the vector quantization context under the name of *entropy constrained vector quantization (ECVQ)*.

The VT algorithm for estimation of the emission parameters of the hidden Markov model can be described as follows. Using some initial values for the parameters, find a realization of Y that maximizes the likelihood of the given observations. Such an n-tuple of states is called a *Viterbi alignment*. Every Viterbi alignment partitions the sample into subsamples corresponding to the states appearing in the alignment. A subsample corresponding to state l is regarded as an *i.i.d.* sample from P_l and is used to find $\hat{\mu}_l$, the maximum likelihood estimate of θ_l. These estimates are then used to find an alignment in the next step of the training, and so on. It can be shown that in general this procedure converges in finitely many steps; also, it is usually much faster than the EM-algorithm.

Although VT is computationally feasible and converges fast, it has a significant disadvantage: The obtained estimators need not be (local) maximum likelihood estimators; moreover, they are generally biased and inconsistent. (VT does not necessarily increase the likelihood, it is, however, an ascent algorithm maximizing a certain other objective function.) Despite this deficiency, speech recognition experiments do not show any significant degradation of the recognition performance when the EM algorithm is replaced by VT. There appears no other explanation of this phenomena but the "curse of complexity" of the very speech recognition system based on HMM.

This paper considers VT largely outside the speech recognition context. We regard the VT procedure merely as a parameter estimation method, and we address the following question: Is it possible to adjust VT in such a way that the adjusted training still has the attractive properties of VT (fast convergence and computational feasibility) and that the estimators are, at the same time, "more accurate" than those of the baseline proce-

dure? In particular, we focus on a special property of the EM algorithm that VT lacks. This property ensures that the true parameters are asymptotically a fixed point of the algorithm. In other words, for a sufficiently large sample, the EM algorithm "recognizes" the true parameters and does not change them much. VT does not have this property; even when the initial parameters are correct (and n is arbitrarily large), an iteration of the training procedure would in general disturb them. *We thus attempt to modify VT in order to make the true parameters an asymptotic fixed point of the resulting algorithm.* In accomplishing this task it is crucial to understand the asymptotic behavior of $\hat{P}_l^n$, the empirical measures corresponding to the subsamples obtained from the alignment. These measures depend on the set of parameters used by the alignment, and in order for the true parameters to be asymptotically fixed by (adjusted) VT, the following must hold: If $\hat{P}_l^n$ is obtained by the alignment with the true parameters, and n is sufficiently large, then $\hat{\mu}_l$, the estimator obtained from P_l^n, must be close to the true parameters. The latter would hold if

$$\hat{P}_l^n \Rightarrow P_l, \quad \text{a.s.} \tag{1}$$

and if the estimators $\hat{\mu}_l$ were continuous[1] at P_l with respect to the convergence in (1). The reason why VT does not enjoy the desired fixed point property is, however, different and is that (1) need not in generally hold. Hence, in order to improve VT in the aforementioned sense, one needs to study the asymptotics of the measures $\hat{P}_l^n$. First of all, one needs to know if there exist any limiting probability measures Q_l such that for every $l \in S$

$$\hat{P}_l^n \Rightarrow Q_l, \quad l \in S \quad \text{a.s..} \tag{2}$$

If such limiting measures exist, then under the above continuity assumption, the estimators $\hat{\mu}_l$ will converge to μ_l, where

$$\mu_l = \arg\max_{\theta_l} \int \ln f_l(x;\theta_l) Q_l(dx).$$

Taking now into account the difference between μ_l and the true parameter, the appropriate adjustment of VT, so called adjusted Viterbi training (VA) can be defined (§2.2).

Let us briefly introduce the main ideas of the paper. Let X stand for the observable subprocess of our HMM. The core of the problem is that the alignment is not defined for infinite sequences of observations, hence the asymptotic behavior of $\hat{P}_l^n$ is not straightforward. To handle this, we introduce the notion of *barrier* (§3). Roughly, a barrier is a block of observations from a predefined cylinder set that has the following property: Alignments for contiguous subsequences of observations enclosed by barriers can be performed independently of the observations outside these enclosing barriers. A simple example of a barrier is an observation z that determines, or indicates, the underlying state: $x_u = z \Rightarrow y_u = l$, $u \leq n$. This happens if z can only be emitted from l. This also implies that any Viterbi alignment has to go through l at time u, and in particular, the alignment up to u does not depend on the observations after time u. If a realization had many such special z's, then the alignment could be obtained piecewise, gluing together

[1] Loosely speaking, the requirement is that $\hat{\mu}_l$ is *consistent*.

subalignments each for each segment enclosed by two consecutive z's.

Barriers are a generalization of this concept. A barrier is characterized by containing a special observation termed a *node* (of order $r \geq 0$). Suppose a barrier is observed with x_u being its node. The node guarantees the existence of state l such that any alignment goes through l at time u independently of the observations outside the barrier.

Lemma 3.1 states (under certain assumptions) the existence of a special path, or a block, of Y states such that, first, the path itself occurs with a positive probability, and second, the (conditional) probability of it emitting a barrier is positive. Hence, by ergodicity of the full HMM process, *almost every* sequence of observations has infinitely many barriers emitted from this special block. Next, we introduce random times τ_i's at which such nodes are emitted. Note that τ_i's are unobservable: We do observe the barriers but without knowing whether or not the underlying MC is going through that special block at the same time. It is, however, not difficult to see that the times $T_i = \tau_i - \tau_{i-1}$ are *renewal times*, and furthermore, the process X is *regenerative* with respect to the times τ_i (Proposition 4.2).

Recall that almost every sequence of observations has infinitely many barriers and that every barrier contains a node. For a generic such sequence, let u_i be the times of its nodes. Note that u_i-s are observable and that also every for all $j = 1, 2, \ldots$, $\tau_j = u_i$ for some $i \geq j$ (there may be more nodes than those emitted from the special block). Using these u_i's as dividers, we define infinite alignment piecewise (Definition 4.1). Formally we have defined a mapping $v : \mathcal{X}^\infty \to S^\infty$, where $\mathcal{X}^\infty$ is the set of all possible observation sequences, and S^∞ is the set of all possible state-sequences. Hence, $V = v(X)$ is a well defined *alignment process*. We consider the two-dimensional process $Z := (X, V)$, and we note that this process is also regenerative with respect to τ_i's. We now define empirical measures $\hat{Q}_l^n$ that are based on the first n elements of Z (Definition 4.2). Using the regenerativity, it is not hard to show that there exists a limit measure Q_l such that $\hat{Q}_l^n \Rightarrow Q_l$, a.s. and $\hat{P}_l^n \Rightarrow Q_l$ (Theorem 4.4). *This is the main result of the paper.*

To implement VA in practice, a closed form of Q_l (or $\hat{\mu}_l$) as a function of the true parameters is necessary. The measures Q_l depend on both the transition and the emission parameters, and computing Q_l can be very difficult. However, in the special case of mixture models, the measures Q_l are easier to find. In [12], VA is described for the mixture case. The simulations in [12, 11] verify that VA indeed recovers the asymptotic fixed point property. Also, since the appropriate adjustment function does not depend on the data, each iteration of VA enjoys the same order of computational complexity (in terms of the sample size) as the baseline VT. Moreover, for commonly used mixtures, such as, for example mixtures of multivariate normal distributions with unknown means and known covariances, the adjustment function is available in a closed form (requiring integration with the mixture densities). Depending on the dimension of the emission, the number of components, and on the available computational resources, one can vary the accuracy of the adjustment. We reiterate that, unlike the computations of the EM algorithm, com-

putations of our adjustment do not involve evaluation and subsequent summation of the mixture density at every data point. Also, instead of calculating the measures Q_l exactly, one can easily simulate them producing in effect a stochastic version of VA. Although simulations do require extra computations, the overall complexity of the stochastically adjusted VT can still be considerably lower than that of EM, but this, of course, requires further investigation.

2 Adjusted Viterbi training

In this section, we define the adjusted Viterbi training and we state the main question of the paper. We begin with the formal definition of the model.

2.1 The model

Let Y be a Markov chain with finite state space $S = \{1, \ldots, K\}$. We assume that Y is irreducible and aperiodic with transition matrix $\mathbb{P} = (p_{ij})$ and initial distribution π that is also the stationary distribution of Y. We consider a hidden Markov model (HMM), in which to every state $l \in S$ there corresponds an *emission distribution* P_l on $(\mathcal{X}, \mathcal{B})$. We assume $\mathcal{X}$ and $\mathcal{B}$ are a separable metric space and the corresponding Borel σ-algebra, respectively. Let f_l be a density function of P_l with respect to a certain dominating measure λ on $(\mathcal{X}, \mathcal{B})$. Two most important concrete examples are $(\mathbb{R}^d, \mathcal{B})$ with Lebesgue measure and discrete spaces with the counting measure. We define support of P_l as the interesection of all closed sets of probability 1 under P_l, and denote such supports by G_l.

In our model, to any realization $y_1, y_2, \ldots$ of Y there corresponds a sequence of independent random variables, $X_1, X_2, \ldots$, where X_n has the distribution P_{y_n}. We do not know the realizations y_n (the Markov chain Y is hidden), as we only observe the process $X = X_1, X_2, \ldots$, or, more formally:

Definition 2.1 *We say that the stochastic process X is a hidden Markov model if there is a (measurable) function h such that for each n,*

$$X_n = h(Y_n, e_n), \quad \textit{where} \quad e_1, e_2, \ldots \quad \textit{are i.i.d. and independent of } Y. \tag{3}$$

Hence, the emission distribution P_l is the distribution of $h(l, e_n)$. The distribution of X is completely determined by the chain parameters (P, π) and the emission distributions P_l, $l \in S$. Moreover, the processes Y and X have the following properties:

- given Y_n, the observation X_n is independent of Y_m, $m \neq n$. Thus, the conditional distribution of X_n given $Y_1, Y_2 \ldots$ depends on Y_n only;
- the conditional distribution of X_n given Y_n depends only on the state of Y_n and not on n;
- given $Y_1, \ldots, Y_n$, the random variables $X_1, \ldots, X_n$ are independent.

The process X is also mixing and, therefore, ergodic.

2.2 Viterbi alignment and training

Suppose we observe $x_1, \ldots, x_n$, the first n elements of X. Throughout the paper, we will also use the shorter notation $x_{1\ldots n}$. A central concept of the paper is the *Viterbi alignment*, which is any sequence of states $q_{1\ldots n} \in S^n$ that maximizes the likelihood of observing $x_{1\ldots n}$. In other words, the Viterbi alignment is a maximum-likelihood estimate of the realization of $Y_1, \ldots, Y_n$ given $x_1, \ldots, x_n$. In the following, the Viterbi alignment will be referred to as the *alignment*. We start with the formal definition of the alignment. First note that for any sequence $q_{1\ldots n} \in S^n$ of states and sets $B_i \in \mathcal{B}$ $i = 1, \ldots, n$,

$$\mathbf{P}(X_1 \in B_1, \ldots, X_n \in B_n, Y_1 = q_1, \ldots, Y_n = q_n) = \mathbf{P}(Y_1 = q_1, \ldots, Y_n = q_n) \prod_{i=1}^{n} \int_{B_i} f_{q_i} d\lambda,$$

and define $\Lambda(q_1, \ldots, q_n; x_1, \ldots, x_n)$ to be the likelihood function:

$$\Lambda(q_{1\ldots n}; x_{1\ldots n}) \stackrel{\text{def}}{=} \mathbf{P}(Y_i = q_i,\ i = 1, \ldots, n) \prod_{i=1}^{n} f_{q_i}(x_i).$$

Definition 2.2 *For each $n \geq 1$, let the set of all the alignments be defined as follows:*

$$\mathcal{V}(x_{1\ldots n}) = \{v \in S^n :\ \forall w \in S^n\ \Lambda(v; x_{1\ldots n}) \geq \Lambda(w; x_{1\ldots n})\}. \tag{4}$$

Any map $v : \mathcal{X}^n \mapsto \mathcal{V}(x_{1\ldots n})$ as well as any element $v \in \mathcal{V}(x_1, \ldots, x_n)$ will also be called an alignment.

Note that alignments require the knowledge of all the parameters of X: (π, P) and P_l $\forall l \in S$.

Throughout the paper we assume that the sample $x_{1\ldots n}$ is generated by an HMM with transition parameters $(\pi, \mathbb{P})$ and with the emission distributions $f_i(x; \theta_l^*)$, where $\theta^* = (\theta_1^*, \ldots, \theta_K^*)$ are the unknown true parameters. We assume that the transition parameters $\mathbb{P}$ and π are known, but the emission densities are known only up to the parametrization $f_l(\cdot; \theta_l)$, $\theta_l \in \Theta_l$. A straightforward generalization to the case when $\psi = (\mathbb{P}, \theta^*)$, all of the free parameters, are unknown, can be found in [13]. In the present case, the likelihood function Λ as well as the set of alignments $\mathcal{V}$ can be viewed as a function of θ. In the following, we shall write $\mathcal{V}_\theta$ for the set of alignments using the parameters θ. Also, unless explicitly specified, $v_\theta \in \mathcal{V}_\theta$ will denote an arbitrary element of $\mathcal{V}_\theta$.

The classical method for computing MLE of θ^* is the EM algorithm. However, if the dimension of X is high, n is big and f_i's are complex, then EM can be (and often is) computationally involved. For this reason, a shortcut, the so-called *Viterbi training* is used. The Viterbi training replaces the computationally expensive expectation (E-)step by an appropriate maximization step that is based on the alignment, and is generally computationally cheaper in practice than the expectation. We now describe the Viterbi training in the HMM case.

Viterbi training

1. Choose an initial value $\theta^o = (\theta_1^o, \ldots, \theta_K^o)$.

2. Given θ^j, obtain alignment
$$v_{\theta^j}(x_{1\ldots n}) = v_{1\ldots n}$$
and partition the sample $x_1, \ldots, x_n$ into K sub-samples, where the observation x_k belongs to the l^{th} subsample if and only if $v_k = l$. Equivalently, we define (at most) K empirical measures
$$\hat{P}_l^n(A; \theta^j, x_{1\ldots n}) \stackrel{\text{def}}{=} \frac{\sum_{i=1}^n I_{A\times l}(x_i, v_i)}{\sum_{i=1}^n I_l(v_i)}, \quad A \in \mathcal{B}, \quad l \in S. \tag{5}$$

3. For every sub-sample find MLE given by:
$$\hat{\mu_l}^n(\theta^j, x_{1\ldots n}) = \arg\max_{\theta_l \in \Theta_l} \int \ln f_l(\theta_l, x) \hat{P}_l^n(dx; \theta^j, x_{1\ldots n}), \tag{6}$$
and take
$$\theta_l^{j+1} = \hat{\mu}_l(\theta^j, x_{1\ldots n}), \quad l \in S.$$
If for some $l \in S$ $v_i \neq l$ for any $i = 1, \ldots, n$ (l^{th} subsample is empty), then the empirical measure $\hat{P}_l^n$ is formally undefined, in which case we take $\theta_l^{j+1} = \theta_l^j$. We will be omitting this exceptional case from now on.

The Viterbi training can be interpreted as follows. Suppose that at some step j, $\theta^j = \theta^*$ and hence v_{θ^j} is obtained using the true parameters. The training is then based on the assumption that the alignment $v_{1\ldots n} = v(x_{1\ldots n})$ is correct, i.e., $v_i = Y_i$, $i = 1, \ldots, n$. In this case, the empirical measures $\hat{P}_l^n$, $l \in S$ would be obtained from the i.i.d. sample generated from $P_l(\theta^*)$, and the MLE $\hat{\mu_l}^n(\theta^*, X_{1\ldots n})$ would be a natural estimator to use. Clearly, under these assumptions $\hat{P}_l^n(\theta^*, X_{1\ldots n}) \Rightarrow P_l(\theta^*)$ a.s. ("$\Rightarrow$" denotes the weak convergence of probability measures) and, provided that $\{f_l(\cdot;\theta) : \theta \in \Theta_l\}$ is a P_l-Glivenko-Cantelli class and Θ_l is equipped with some suitable metric, $\lim_{n\to\infty} \hat{\mu}_l^n(\theta^*, X_{1\ldots n}) = \theta_l^*$ a.s. Hence, if n is sufficiently large, then $\hat{P}_l^n \approx P_l$ and
$$\theta_l^{j+1} = \hat{\mu}_l^n(\theta^*, x_{1\ldots n}) \approx \theta_l^* = \theta_l^j, \quad \forall l$$
i.e. $\theta^j = \theta^*$ would be (approximately) a fixed point of the training algorithm.

A weak point of the foregoing argument is that the alignment in general is not correct even when the parameters used to find it, are. So, generally $v_i \neq Y_i$. In particular, this implies that the empirical measures $\hat{P}_l^n(\theta^*, x_{1\ldots n})$ are not obtained from an i.i.d. sample from $P_l(\theta^*)$. Hence, we have no reason to believe that $\hat{P}_l^n(\theta^*, X_{1\ldots n}) \Rightarrow P_l(\theta^*)$ a.s. and $\lim_{n\to\infty} \hat{\mu}_l^n(\theta^*, X_{1\ldots n}) = \theta_l^*$ a.s. Moreover, we do not even know whether the sequences of empirical measures $\{\hat{P}_l^n(\theta^*, X_{1\ldots n})\}$ and MLE estimators $\{\hat{\mu}_l^n(\theta^*, X_{1\ldots n})\}$ converge (a.s.) at all.

In this paper, we prove the existence of probability measures $Q_l(\theta, \theta^*)$ (that depend on both θ, the parameters used to obtain the alignments, as well as θ^*, the true parameters used to generate the training samples), such that for every $l \in S$,

$$\hat{P}_l^n(\theta^*, X_{1\ldots n}) \Rightarrow Q_l(\theta^*, \theta^*), \quad \text{a.s.} \tag{7}$$

for a special choice of the alignment $v_{\theta^*} \in \mathcal{V}_{\theta^*}$ used to define $\hat{P}_l^n(\theta^*, x_{1\ldots n})$. (In fact, adding certain mild restrictions on P_l, one can eliminate the dependence of the above result on the particular choice of the alignment $v_{\theta^*} \in \mathcal{V}_{\theta^*}$.) We will also be writing $Q_l(\theta)$ for $Q_l(\theta, \theta)$ whenever appropriate.

Suppose also that the parameter space Θ_l is equipped with some metric. Then, under certain consistency assumptions on classes $\mathcal{F}_l = \{f_l(\cdot; \theta_l) : \theta_l \in \Theta_l\}$, the convergence

$$\lim_{n\to\infty} \hat{\mu}_l(\theta^*, X_{1\ldots n}) = \mu_l(\theta^*) \quad \text{a.s.} \tag{8}$$

can be deduced from (7), where

$$\mu_l(\theta) \stackrel{\text{def}}{=} \arg\max_{\theta_l' \in \Theta_l} \int \ln f_l(x; \theta_l') Q_l(dx; \theta). \tag{9}$$

We also show that in general, for the baseline Viterbi training $Q_l(\theta^*) \neq P_l(\theta^*)$, implying $\mu_l(\theta^*) \neq \theta_l^*$. In an attempt to reduce the bias $\theta_l^* - \mu_l(\theta^*)$, we next propose the *adjusted Viterbi training*.

Suppose (7) and (8) hold. Based on (9), we now consider the mapping

$$\theta \mapsto \mu_l(\theta), \quad l = 1, \ldots, K, \tag{10}$$

The calculation of $\mu_l(\theta)$ can be rather involved and it may have no closed form. Nonetheless, since this function is independent of the sample, we can define the following correction for the bias:

$$\Delta_l(\theta) = \theta_l - \mu_l(\theta), \quad l = 1, \ldots, K. \tag{11}$$

Thus, the adjusted Viterbi training emerges as follows:

Adjusted Viterbi training

1. Choose an initial value $\theta^0 = (\theta_1^0, \ldots, \theta_K^0)$.
2. Given θ^j, perform the alignment and define K empirical measures $\hat{P}_l^n(\theta^j, \theta^*)$ as in (5).
3. For every $\hat{P}_l^n(\theta^j, x_{1\ldots n})$, find $\hat{\mu}_l^n(\theta^j, x_{1\ldots n})$ as in (6).
4. For each l, define
$$\theta_l^{j+1} = \hat{\mu}_l^n(\theta^j, x_{1\ldots n}) + \Delta_l(\theta^j),$$
where Δ_l as in (11).

Note that, as desired, for a sufficiently large n, the adjusted training algorithm has θ^* as its (approximately) fixed point: Indeed, suppose $\theta^j = \theta^*$, then $\hat{\mu}_l^n(\theta^j, x_{1\ldots n}) = \hat{\mu}_l^n(\theta^*, x_{1\ldots n})$. Recalling (8), it then follows that $\hat{\mu}_l^n(\theta^*, x_{1\ldots n}) \approx \mu_l(\theta^*) = \mu_l(\theta^j)$, for all $l \in S$. Hence,

$$\theta_l^{j+1} = \hat{\mu}_l(\theta^*, x_{1\ldots n}) + \Delta_l(\theta^*) \approx \mu_l(\theta^*) + \Delta_l(\theta^*) = \theta_l^* = \theta^j, \quad l \in S. \tag{12}$$

In [12], we considered i.i.d. sequence $X_1, X_2, \ldots$, where X_1 has a mixture distribution, i.e. the density of X_1 is $\sum_{i=1}^K p_i f_i$. Here $p_i > 0$ are the mixture weights. Such a sequence is an HMM with the transition matrix satisfying $p_{ij} = p_j \ \forall i, j$. In this particular case, the alignment and the measures Q_l are easy to find. Indeed, for any set of parameters $\theta = (\theta_1, \ldots \theta_K)$, the alignment v_θ can be obtained via a *Voronoi partition* $\mathcal{S}(\theta) = \{S_1(\theta), \ldots, S_K(\theta)\}$, where

$$S_1(\theta) = \{x : p_1 f_1(x; \theta_1) \geq p_j f_j(x; \theta_j), \quad \forall j \in S\} \tag{13}$$

$$S_l(\theta) = \{x : p_l f_l(x; \theta_l) \geq p_j f_j(x; \theta_j), \quad \forall j \in S\} \backslash (S_1 \cup \ldots \cup S_{l-1}), \quad l = 2, \ldots, K. \tag{14}$$

Now, the alignment can be defined point-wise as follows: $v_\theta(x_1, \ldots, x_n) = v_\theta(x_1) \cdots v_\theta(x_n)$, where $v_\theta(x) = l$ if and only if $x \in S_l(\theta)$.
The convergence (7) now follows immediately from the strong law of large numbers as $\hat{P}_l^n(\theta^*, X_{1\ldots n}) \Rightarrow Q_l(\theta^*)$ a.s., where

$$q_l(x; \theta^*) \propto f(x; \theta^*) I_{S_l(\theta^*)} = (\sum_i p_i f_i(x; \theta^*)) I_{S_l(\theta^*)}, \quad l = 1, \ldots, K$$

are the densities of respective $Q_l(\theta^*)$.

Thus, in the special case of mixtures, the adjustments Δ_l are easy to calculate and the adjusted Viterbi training is easy to implement. Simulations in [12] have largely supported the expected gain in estimation accuracy due to the adjustment Δ with a small extra cost for computing Δ. Indeed, this extra computation does not affect the algorithm's overall computational complexity as a function of the sample size, since Δ depends on the training sample only through θ^j, the current value of the parameter.

Due to the time-dependence in the general HMM, the convergence (7) does not follow immediately from the law of large numbers. However, the very concept of the adjusted Viterbi training is based on the existence of the Q_l-measures. Thus, in order to generalize this concept to an arbitrary HMM, one has to begin with the existence of the Q_l-measures, which is *the objective of this paper*.

3 Nodes and barriers

In this section, we present some preliminaries that will allow us to prove the convergences (7) and (8). We choose to introduce the necessary concepts gradually, building up the general notions on special cases that we find more intuitive and insightful. For a comprehensive introduction to HMM's and related topics we refer to [8, 18, 19], and an overview

of the basic concepts related to HMM's follows below in §3.1. We then proceed to the notion of *infinite (Viterbi) alignment* (§4.2), developing on the way several auxiliary notions such as *nodes* and *barriers*.

Throughout the rest of this section, we will be writing f_l and $\mathcal{V}$ for $f_l(\cdot;\theta^*_l)$, the true emission distributions, and $\mathcal{V}_{\theta^*}$, the set of alignments with the true parameters, respectively.

3.1 Nodes

3.1.1 Preliminaries

Let $1 \le u_1 < u_2 < \ldots < u_k \le n$. Given any sequence $a = (a_1, \ldots, a_n)$, write $a_{u_1 \ldots u_k}$ for $(a_{u_1}, \ldots, a_{u_k})$ and define also the following objects:

$$S^{l_1 \ldots l_k}_{u_1 \ldots u_k}(n) \stackrel{\text{def}}{=} \{v \in S^n : v_{u_1 \ldots u_k} = (l_1, \ldots, l_k)\}.$$

Next, given observations $x_{1\ldots n}$, let us introduce the set of constrained likelihood maximizers defined below:

$$\mathcal{W}^l_u(x_{1\ldots n}) = \{v \in S^l_u(n) : \forall w \in S^l_u(n)\ \Lambda(v; x_{1\ldots n}) \ge \Lambda(w; x_{1\ldots n})\}.$$

Next, define the *scores*

$$\delta_u(l) \stackrel{\text{def}}{=} \max_{q \in S^l_u(u)} \Lambda(q; x_{1\ldots u}), \tag{15}$$

and notice the trivial case: $\delta_l(1) = \pi_l f_l(x_1)$. Then, we have the following recursion (see, for example, [19]):

$$\delta_{u+1}(j) = \max_{l \in S}(\delta_u(l)p_{lj})f_j(x_{u+1}). \tag{16}$$

The Viterbi training as well as the Viterbi alignment inherit their names from the *Viterbi algorithm*, which is a dynamic programming algorithm for finding $v \in \mathcal{V}(x_{1\ldots n})$. In fact, due to potential non-uniqueness of such v, the Viterbi algorithm requires a selection rule as part of its specification. However, for our purposes we will often be manipulating by $\mathcal{V}(x_{1\ldots n})$ as opposed to by individual v's, in which case we will also be identifying the entire $\mathcal{V}(x_{1\ldots n})$ with the output of the algorithm. This algorithm is based on recursion (16) and on the following relations:

$$t(u, j) = \{l \in S : \forall i \in S\ \delta_u(l)p_{lj} \ge \delta_u(i)p_{ij}\}, \quad u = 1, \ldots, n-1, \tag{17}$$

$$\mathcal{V}(x_{1\ldots n}) = \{v \in S^n : \delta_n(v_n) \ge \delta_n(i)\ \forall i \in S, v_u \in t(u, v_{u+1})\ 1 \le u < n\}. \tag{18}$$

It can also be shown that

$$\mathcal{W}^l_n(x_{1\ldots n}) = \{v \in S^l_n(n) :\ v_u \in t(u, v_{u+1})\ u = 1, \ldots, n-1\}. \tag{19}$$

We shall also need the following notation:

$$\mathcal{V}^{l_1 \ldots l_k}_{u_1 \ldots u_k}(x_{1\ldots n}) = \{v \in \mathcal{V}(x_{1\ldots n}) : v_{u_i u_{i+1} \ldots u_k} = (l_1, \ldots, l_k)\}.$$

and will use subscript (l) to refer to alignments obtained using $(p_{li})_{i\in S}$ (instead of π) as the initial distribution. Thus $\mathcal{V}_{(l)}(x_{1\dots n})$ stands for the set of all such alignments, and

$$\mathcal{V}_{(l)u_1\dots u_k}^{l_1\dots l_k}(x_{1\dots n}) = \{v \in \mathcal{V}_{(l)}(x_{1\dots n}) : v_{u_i u_{i+1}\dots u_k} = (l_1, \dots, l_k)\}.$$

Similarly, $\mathcal{W}_{(l)u_1\dots u_k}^{l_1\dots l_k}(x_{1\dots n})$ will be referring to the constrained alignments obtained using $(p_{li})_{i\in S}$ as the initial distribution. The following Proposition and Corollary reveal more structure of the alignments.

Proposition 3.1 *Let $1 \le u \le n$, then*

$$\mathcal{W}_u^l(x_{1\dots n}) = \mathcal{W}_u^l(x_{1\dots u}) \times \mathcal{V}_{(l)}(x_{u+1\dots n}), \tag{20}$$

$$\mathcal{V}_u^l(x_{1\dots n}) \neq \emptyset \Rightarrow \mathcal{V}_u^l(x_{1\dots n}) = \mathcal{W}_u^l(x_{1\dots n}). \tag{21}$$

Proof. The Markov property implies: for any $q = (q_1, \dots, q_n)$.

$$\Lambda(q; x_{1\dots n}) = \Lambda(q_{1\dots u}; x_{1\dots u}) \cdot \Lambda(q_{u+1\dots n}; x_{u+1\dots n}|q_u),$$

where

$$\Lambda(q_{u+1\dots n}; x_{u+1\dots n}|l) = \mathbf{P}(Y_{u+1\dots n} = q_{u+1\dots n}|Y_u = l) \prod_{i=u+1}^{n} f_{q_i}(x_i).$$

Thus, (20) follows from the equivalence between maximizing $\Lambda(q; x_{1\dots n})$ over $S_u^l(n)$ on one hand, and maximizing $\Lambda(q_{1\dots u}; x_{1\dots u})$ and $\Lambda(q_{u+1\dots n}; x_{u\dots n}|l)$ over S^{n-u} and $S_u^l(n)$, respectively and independently, on the other. (21) follows immediately from the definitions of the involved sets. ■

Corollary 3.1

$$\mathcal{V}_u^l(x_{1\dots n}) \neq \emptyset \text{ and } \mathcal{V}_u^l(x_{1\dots u}) \neq \emptyset \Rightarrow \mathcal{V}_u^l(x_{1\dots n}) = \mathcal{V}_u^l(x_{1\dots u}) \times \mathcal{V}_{(l)}(x_{u+1\dots n}). \tag{22}$$

Proof. The hypotheses of (22) together with (21) imply $\mathcal{V}_u^l(x_{1\dots n}) = \mathcal{W}_u^l(x_{1\dots n})$ and $\mathcal{V}_u^l(x_{1\dots u}) = \mathcal{W}_u^l(x_{1\dots u})$. The latter statements and (20) yield the claim. ■

3.1.2 Nodes and alignment

We aim at extending the notion of alignment for infinite HMM's. In order to fulfil this objective, we investigate properties of finite alignments (e.g. Propositions 3.1, and 3.2) and identify necessary ingredients (e.g. "node", and "barrier") for the development of the extended theory. We start with the notion of nodes:

Definition 3.1 *For $1 \le u < n$, we call x_u an l-node if*

$$\delta_u(l)p_{lj} \ge \delta_u(i)p_{ij}, \quad \forall i, j \in S. \tag{23}$$

We also say that x_u is a node if it is an l-node for some $l \in S$.

Figure 1 illustrates the newly introduced notion.

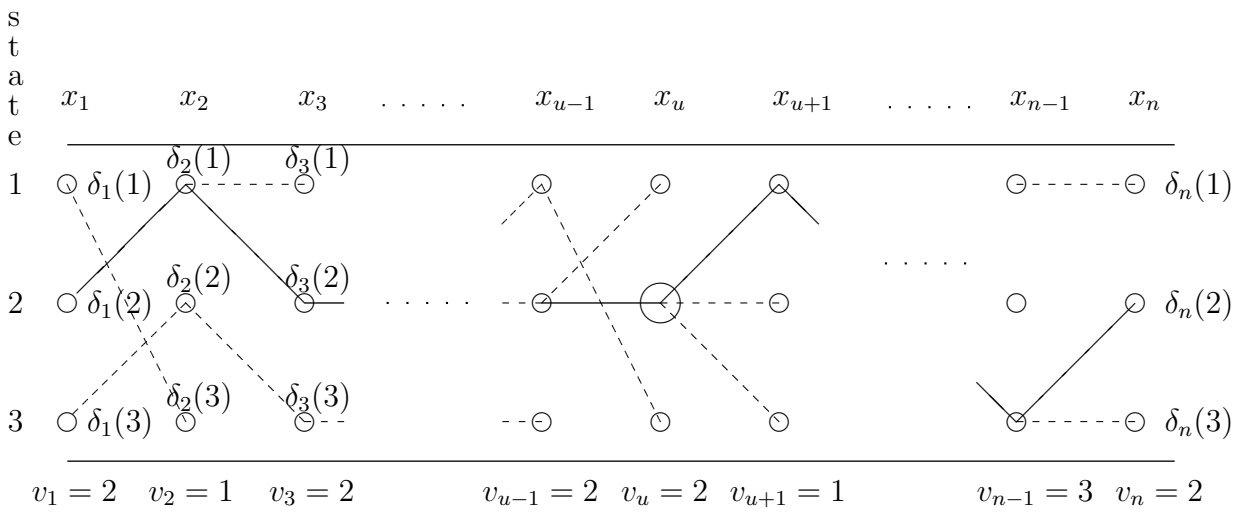

Figure 1: An example of the Viterbi algorithm in action. The solid line corresponds to the final alignment $v_{1\ldots n}$. The dashed links are of the form $(k,l)-(k+1,j)$ with $l \in t(k,j)$ and are not part of the final alignment. E.g., $(1,3)-(2,2)-(3,3)$ is because $3 \in t(1,2)$, $2 \in t(2,3)$. The observation x_u is a 2-node, since we have $2 \in t(u,j)\ \forall j \in S$. We also see that $v_{1\ldots u}$ is *fixed*.

Proposition 3.2

$$
\begin{aligned}
x_u \text{ is an } l\text{-node} \quad &\iff \quad l \in t(u,j)\ \forall j \in S, && (24)\\
&\Rightarrow \quad \mathcal{V}_u^l(x_{1\ldots u}) \neq \emptyset, && (25)\\
&\Rightarrow \quad \forall v \in \mathcal{V}(x_{1\ldots n}), \forall v^* \in \mathcal{V}_u^l(x_{1\ldots u})\ (v^*, v_{u+1\ldots n}) \in \mathcal{V}_u^l(x_{1\ldots n}), && (26)\\
&\Rightarrow \quad \mathcal{V}_u^l(x_{1\ldots n}) \neq \emptyset, && (27)\\
&\Rightarrow \quad \textit{Right hand side of } (22). && (28)
\end{aligned}
$$

Whether x_u is a node does not depend on x_i, $i > u$.

Proof. The final statement follows immediately from Definition 3.1 and (15), and (24) also follows immediately from Definition 3.1 and (17). Summing both sides of (23) over $j \in S$, we obtain

$$\delta_u(l) \geq \delta_u(i), \quad \forall i \in S, \tag{29}$$

hence, (25) holds by (18). Note that (26) means that any alignment $v \in \mathcal{V}(x_{1\ldots n})$ can be modified by setting $v_u = l$ and taking $v_i^* \in t(i, v_{i+1})$ for $i = u-1, u-2, \ldots, 1$, and the modified string remains an alignment, i.e. belongs to $\mathcal{V}(x_{1\ldots n})$. Such a modification is evidently always possible, i.e., $(v^*, v_{u+1\ldots n})$ is well-defined since $\mathcal{V}_u^l(x_{1\ldots u}) \neq \emptyset$. For $u = n$ this holds trivially, for $u < n$ this follows from (24) (as the latter implies $l \in t(u, v_{u+1})$ for any value of v_{u+1}), and (18). Also, (26) implies (27). Finally, given (25) and (27), Corollary 3.1 yields (28). ■

Remark 3.2 *Note that a modification of $v \in \mathcal{V}(x_{1\ldots x_n})$ possibly required to enforce $v_u = l$ when x_u is an l-node (see proof of (26) above) depends only on $x_1, \ldots, x_{u-1}$. Thus, if x_u*

is an l-node and if $v^* \in \mathcal{V}_u^l(x_{1\ldots x_u})$*, then for any* $n > u$ *and any* $x_{u+1}, \ldots, x_n$ (26) *always guarantees an alignment* $v \in \mathcal{V}(x_{1\ldots n})$ *with* $v_{1\ldots u} = v^*$*, in which case we can call* v^* fixed*, meaning that* v^* *can be kept as the substring of the first* u *components for any alignment based on the extended observations.*

The fact that $v \in \mathcal{V}(x_{1\ldots n})$ in general does not imply $v_{1\ldots u} \in \mathcal{V}(x_{1\ldots u})$ complicates the structure of the alignments and furthermore emphasizes the significance of nodes in view of (28) and Remark 3.2.

Corollary 3.2 *Suppose the observations* $x_1, \ldots, x_n$ *are such that for some* $1 \le u_1 < u_2 < \cdots < u_k \le n$*, the observations* x_{u_i} *are* l_i*-nodes,* $i = 1, \ldots, k-1$*. Then*

$$\emptyset \neq \mathcal{V}_{u_1 u_2 \cdots u_k}^{l_1 l_2 \cdots l_k}(x_{1\ldots n}) =$$
$$= \mathcal{V}_{u_1}^{l_1}(x_{1\ldots u_1}) \times \mathcal{V}_{(l_1)u_2}^{\ \ l_2}(x_{u_1+1\ldots u_2}) \times \cdots \times \mathcal{V}_{(l_{k-1})u_k}^{\ \ l_k}(x_{u_{k-1}+1\ldots u_k}) \times \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n}). \quad (30)$$

Proof. By (25),

$$\mathcal{V}_{u_i}^{l_i}(x_{1\ldots u_i}) \neq \emptyset, \quad i = 1, \ldots, k.$$

By (27)

$$\mathcal{V}_{u_k}^{l_k}(x_{1\ldots n}) \neq \emptyset, \quad \mathcal{V}_{u_i}^{l_i}(x_{1\ldots u_{i+1}}) \neq \emptyset \quad i = 1, \ldots, k-1.$$

From (26), it now follows

$$\mathcal{V}_{u_i u_{i+1}}^{l_i l_{i+1}}(x_{1\ldots u_{i+1}}) \neq \emptyset, \quad i = 2, \ldots k-1.$$

Now use (22) to decompose

$$\mathcal{V}_{u_k}^{l_k}(x_{1\ldots n}) = \mathcal{V}_{u_k}^{l_k}(x_{1\ldots u_k}) \times \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n}).$$

Use (22) again to decompose

$$\mathcal{V}_{u_{k-1}u_k}^{l_{k-1}l_k}(x_{1\ldots u_k}) = \mathcal{V}_{u_{k-1}}^{l_{k-1}}(x_{1\ldots u_{k-1}}) \times \mathcal{V}_{(l_{k-1})u_k}^{\ \ l_k}(x_{u_{k-1}+1\ldots u_k}).$$

Proceeding this way, we obtain (30). ■

Corollary 3.2 guarantees the existence of an alignment $v(x_{1\ldots n})$ that can be constructed *piecewise*, i.e.

$$(v_1, \ldots, v_{k+1}) \in \mathcal{V}(x_{1\ldots n}), \quad (31)$$

where

$$v_1 \in \mathcal{V}_{u_1}^{l_1}(x_{1\ldots u_1}),\ v_2 \in \mathcal{V}_{(l_1)u_2}^{\ \ l_2}(x_{u_1+1\ldots u_2}), \ldots, v_k \in \mathcal{V}_{(l_{k-1})u_k}^{\ \ l_k}(x_{u_{k-1}+1\ldots u_k}), v_{k+1} \in \mathcal{V}_{(l_k)}(x_{u_k+1\ldots u_n}).$$

3.1.3 Proper alignment

If the sets $\mathcal{V}_{(l_{i-1})u_i}^{\ \ l_i}(x_{u_{i-1}+1\ldots u_i})$, $i = 2, \ldots, k$ as well as $\mathcal{V}_{(l_k)}(x_{u_k+1\ldots n})$ have a single element each, then the concatenation (31) is unique. Otherwise, a single v_i will need to be selected from $\mathcal{V}_{(l_{i-1})u_i}^{\ \ l_i}(x_{u_{i-1}+1\ldots u_i})$. Thus, suppose that $(x_{u_{i-1}+1\ldots u_i}) = (x_{u_{j-1}+1\ldots u_j})$, and $l_i = l_j$ for some $j \neq i$. Ignoring the fact that the actual probability of such realizations may

well be zero in most cases, for technical reasons we are nonetheless going to be general and require that the selection from any $\mathcal{V}_{(q)u+\Delta}^{\ l}(x_{u+1\ldots u+\Delta})$ for which x_u and $x_{u+\Delta}$ are q and l nodes, respectively, be made independently of u. To achieve this, we impose the following (formally even more restrictive) condition on admissible selection schemes $\{w^{ql}(x_{1\ldots m}) : \mathcal{X}^m \to \mathcal{W}_{(q)m}^{\ l}(x_{1\ldots m}),\ m = 1, \ldots, n,\ q, l \in S\}$:

$$\forall q,\ \forall l \in S,\ \forall m \leq n,\ \forall x_{1\ldots n} \in \mathcal{X}^n :\ w_{1\ldots n} = w^{ql}(x_{1\ldots n}) \Rightarrow w_{1\ldots m} = w^{qw_m}(x_{1\ldots m}). \quad (32)$$

The condition (32) above simply states that the ties are broken consistently.

Definition 3.3 *The alignment* (31) *based on* $l_1, \ldots, l_k$ *nodes* $x_{u_1}, \ldots, x_{u_k}$ *is called proper if for every* $i = 2, \ldots, k-1$

$$v_i = w^{l_i l_{i+1}}(x_{u_i+1\ldots u_{i+1}}),$$

where $\{w^{ql}(x_{1\ldots m}) : \mathcal{X}^m \to \mathcal{W}_{(q)m}^{\ l}(x_{1\ldots m}),\ m = 1, \ldots, n,\ q, l \in S\}$ *is some selection scheme satisfying* (32).

Clearly, there may be many such selection schemes and the following discussion is valid for all of them (provided the choice is fixed throughout). One such selection scheme is based on taking maxima under the reverse lexicographic order on S^m (for any positive integer m). According to this order $\prec$, for $a, b \in S^m$, $a \prec b$ if and only if for some i, $1 \leq i < m$, $a_i < b_i$ and $a_j = b_j$ for $j = i+1, \ldots, m$. (Clearly, if neither $a \prec b$ nor $b \prec a$, then $a_j = b_j$ for $j = 1, \ldots, m$, in which case a and b are defined equal for this order.) It is immediate to verify that (32) holds for

$$w^{ql}(x_{1\ldots m}) \stackrel{def}{=} \max\nolimits_{\prec} \mathcal{W}_{(q)m}^{\ l}(x_{1\ldots m}),\ 1 \leq m \leq n,\ q, l \in S. \quad (33)$$

For the sake of concreteness, we are going to refer to this particular selection scheme as *the selection* and base all proper alignments on it. Also, since Definition 3.3 does not concern the initial or terminal components of the concatenated alignment (31), we extend the selection (again, purely for the sake of concreteness of the presentation) to the initial and terminal components of the concatenated alignment (31). Thus, to specify the initial component we have $w^{\pi l}(x_{1\ldots m}) \stackrel{def}{=} \max_{\prec} \mathcal{W}_m^l(x_{1\ldots m}),\ 1 \leq m \leq n$, for all $l \in S$ and for all π, probability mass functions on S. To be concise, we will write $\vee W$ for the selected element of W for any $W \subset S^m$ (where W generally depends on $x_{1\ldots m}$). In particular, the final component is then specified via $\vee \mathcal{V}_{(l)}(x_{1\ldots m})$.

Example 3.4 *Consider an i.i.d. sequence* $X_1, X_2, \ldots$, *where* X_1 *has a mixture distribution, i.e. the density of* X_1 *is* $\sum_{i=1}^K p_i f_i$. *Here* $p_i > 0$ *are the mixture weights. Such a sequence is an HMM with the transition matrix satisfying* $p_{ij} = p_j\ \forall i, j$. *In this case, an observation* x_u *is an* l*-node if*

$$\delta_u(l) \geq \delta_u(i), \quad \forall i.$$

In particular, this means that every observation is an l*-node for some* $l \in S$. *Then* (16) *becomes*

$$\delta_{u+1}(i) = \max_j(\delta_u(j))p_i f_i(x_{u+1}) \propto p_i f_i(x_{u+1}), \quad \forall i$$

and

$$\delta_u(l) \geq \delta_u(i), \quad \forall i \quad \Longleftrightarrow \quad p_l f_l(x_u) \geq p_i f_i(x_u), \quad \forall i. \tag{34}$$

Thus, in a mixture-model, any observation x_u is a node, more precisely it is an l-node for any $l = \arg\max_j (p_j f_j(x_u))$. For this model, the alignment can naturally be concatenated point-wise: $v(x_{1\ldots n}) = (v(x_1), \ldots, v(x_n))$, where

$$v(x) = \arg\max_i p_i f_i(x). \tag{35}$$

The alignment will be proper if ties in (35) *are broken consistently, which is, for example, the case when using the selection* (33)*.*

3.2 r^{th}-order nodes

The concept of nodes is both important and rich, but the existence of a node can also be restrictive in the following sense: Suppose $x_{1\ldots u}$ is such that $\delta_u(i) > 0$ for every i. In this case, (23) is equivalent to

$$\delta_u(l) \geq \max_i \Big(\max_j \Big(\frac{p_{ij}}{p_{lj}} \Big) \delta_u(i) \Big)$$

and actually implies $p_{lj} > 0$ for every $j \in S$. Hence, one cannot guarantee the existence of an l-node for an arbitrary emission distribution since an ergodic $\mathbb{P}$ in general can have a zero in every row, violating the above positivity constraint on the l^{th} row of $\mathbb{P}$. We now generalize the notion of nodes in order to eliminate the aforementioned positivity constraint and to still enjoy the desirable properties of nodes. We need some additional definitions: For each $u \geq 1$ and $r \geq 1$, let

$$p_{ij}^{(r)}(u) = \max_{q_{1\ldots r} \in S^r} p_{iq_1} f_{q_1}(x_{u+1}) p_{q_1 q_2} f_{q_2}(x_{u+2}) p_{q_2 q_3} \cdots p_{q_{r-1} q_r} f_{q_r}(x_{u+r}) p_{q_r j}. \tag{36}$$

Also, for each $u \geq 1$ define $p_{ij}^{(0)}(u) = p_{ij}$, and notice

$$p_{ij}^{(r)}(u) = \max_{q \in S} p_{iq}^{(r'-1)}(u) f_q(x_{u+1}) p_{qj}^{(r-r')}(u+1), \text{ for all } r' = 1, 2, \ldots, r. \tag{37}$$

The recursion (16) then generalizes to

$$\delta_{u+1}(j) = \max_{i \in S} \big(\delta_{u-r}(i) p_{ij}^{(r)}(u-r) \big) f_j(x_{u+1}), \ r < u. \tag{38}$$

For $r \geq 1$ and $u + r \leq n$ define

$$\begin{aligned} t^{(r)}(u,j) &= \{ l \in S : \forall i \in S \ \delta_u(l) p_{lj}^{(r-1)} \geq \delta_u(i) p_{ij}^{(r-1)} \}, \\ t^{(r)}(u,J) &= \{ t^{(r)}(u,j) : \ j \in J \}, \quad J \subset S. \end{aligned} \tag{39}$$

It can be verified that for $1 \leq q, r$, $q + r \leq n - u$

$$t^{(r+q)}(u,j) = t^{(q)}(u, t^{(r)}(u+q, j)), \tag{40}$$

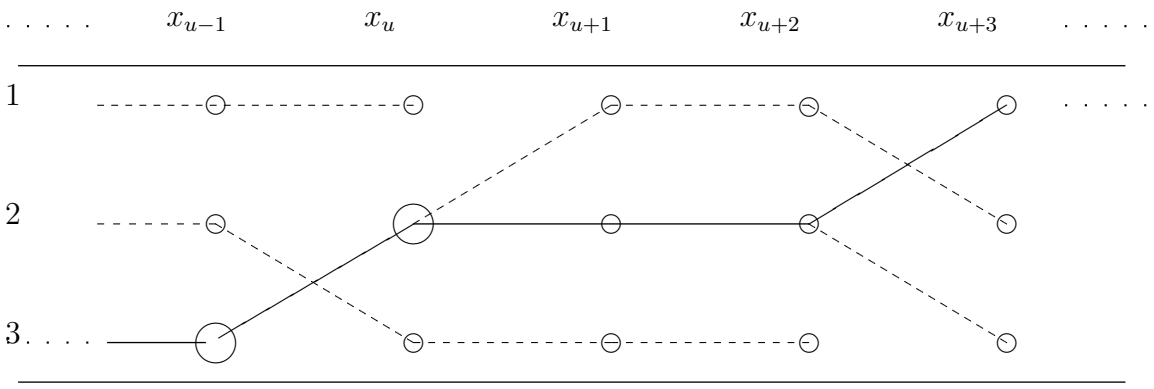

Figure 2: In this example, x_u is a 2^d order 2-node, x_{u-1} is a 3^d-order 3-node. Thus, for given $x_{1...n}$, the alignment includes $v_u = 2$. However, unlike in the case of ordinary nodes (of order 0), x_{u+1} can now destroy the property of x_u being the (second order) node.

where $t^{(1)}(u,j)$ coincides with $t(u,j)$ (18). Thus, $l_1 \in t^{(q)}(u, t^{(r)}(u+q,j))$ in (40) implies the existence of $l_2 \in t^{(r)}(u+q,j)$ such that $l_1 \in t^{(q)}(u,l_2)$. In short,

$$t^{(q)}(u, t^{(r)}(u+q,j)) = \cup_{l \in t^{(r)}(u+q,j)} t^{(q)}(u,l).$$

Note that with this new notation, (18) and (19) can be rewritten respectively as follows:

$$\mathcal{V}(x_1,\ldots,x_n) = \{v \in S^n : \delta_n(v_n) \geq \delta_n(i)\ \forall i \in S, v_u \in t^{(n-u)}(u,v_n)\ 1 \leq u < n\} \quad (41)$$

$$\mathcal{W}^l_u(x_1,\ldots,x_n) = \{v \in S^l_n(n) :\ v_u \in t^{(n-u)}(u,l)\ 1 \leq u < n\} \quad (42)$$

We now generalize the concept of the node:

Definition 3.5 *Let $1 \leq r < n$, $u \leq n-r$ and let $l \in S$. We call x_u an l-node of order r if*

$$\delta_u(l) p^{(r)}_{lj}(u) \geq \delta_u(i) p^{(r)}_{ij}(u), \quad \forall i,j \in S. \quad (43)$$

We also say that x_u is a node of order r if it is an l-node of order r for some $l \in S$.

Note that a 0^{th}-order node is just a node. One immediately obtains the following properties of the (generalized) nodes:

Proposition 3.3 *Let $0 \leq r$, $1 \leq q$ such that $r+q \leq n-u$, then*

1. *If x_u is an r^{th}-order l-node, then it is also an l-node of order $r+q$.*
2. *If x_{u+q} is an r^{th}-order l-node, then x_u is an $(r+q)^{th}$-order l'-node for any $l' \in t^{(q)}(u,l)$.*

Next, we generalize Proposition 3.2:

Proposition 3.4

$$x_u \text{ is an } l\text{-node of order } r \iff l \in t^{(r+1)}(u,j)\ \forall j \in S, \quad (44)$$

$$u+r<n, x_u \text{ is an } l\text{-node of order } r \Rightarrow \forall v \in \mathcal{V}(x_{1\ldots n}), \forall v^* \in \mathcal{W}_u^l(x_{1\ldots u})$$
$$\exists v' \in \mathcal{W}_u^{l\ v_{u+r+1}}{}_{u+r+1}(x_{1\ldots u+r+1}) : v^* = v'_{1\ldots u}, (v', v_{u+r+1\ldots n}) \in \mathcal{V}_u^l(x_{1\ldots n}), \tag{45}$$
$$\Rightarrow \mathcal{V}_u^l(x_{1\ldots n}) \neq \emptyset, \tag{46}$$
$$\Rightarrow \mathcal{V}_u^l(x_{1\ldots n}) = \mathcal{W}_u^l(x_{1\ldots u}) \times \mathcal{V}_{(l)}(x_{u+1\ldots n}). \tag{47}$$

Finding $v'_{u+1\ldots u+r}$ and $v^* \in \mathcal{W}_u^l(x_{1\ldots u})$ in (45) for given $v \in \mathcal{V}(x_{1\ldots n})$ does not require knowledge of any of $x_{u+r+1\ldots n}$. Finally, whether x_u is an l-node of order r depends on $x_1, \ldots, x_{u+r}$ only, i.e. it does not depend on any x_i for $i > u+r$.

Proof. The final statement follows immediately from Definition 3.5 and relations (15) and (36). (44) also follows immediately from Definition 3.5 and (39). In order to see (45), note that applying (40) with $q=1$ to $l \in t^{(r+1)}(u, v_{u+r+1})$ once gives us $\tilde{v}_1 \in t^{(r)}(u+1, v_{u+r+1})$. Applying then (40) with $q=1$ to $\tilde{v}_i \in t^{(r-i+1)}(u+i, v_{u+r+1})$ successively for $i = 2, \ldots, r$ proves the existence of the entire $\tilde{v}_{1\ldots r} \in S^r$ such that $l \in \tilde{t}(u, v'_1)$, $\tilde{v}'_1 \in t(u+1, \tilde{v}_2)$, …, $\tilde{v}_{r-1} \in t(u+r-1, \tilde{v}_r)$, $\tilde{v}_r \in t(u, v_{u+r+1})$. Thus, recalling (42), $\tilde{v} = v'_{u+1\ldots u+r}$ for some $v' \in \mathcal{W}_u^{l\ v_{u+r+1}}{}_{u+r+1}(x_{1\ldots u+r+1})$. Since $v_i^* \in t(i, v_{i+1}^*)$ for $i = 1, \ldots, u-1$ ($v^* \in \mathcal{W}_u^l(x_{1\ldots u})$ and (19)), and $v_i \in t(i, v_{i+1})$ for $i = u+r+1, \ldots, n-1$ and $\delta_n(v_n) \geq \delta_n(j)\ \forall j \in S$ ($v \in \mathcal{V}(x_{1\ldots n})$ and (18)), one gets $(v^*, v', v_{u+r+1\ldots n}) \in \mathcal{V}_u^l(x_{1\ldots n})$. Evidently, v' above involves no x_i for $i > u+r$. Thus, unlike in (26), in addition to setting $v_u = l$ and taking $v_i^* \in t(i, v_{i+1})$ for $i = u-1, u-2, \ldots, 1$ we may have to "realign" $u+1^{st}$, …, $u+r^{th}$ components in order for the modified string to remain in $\mathcal{V}(x_{1\ldots n})$. Moreover, v^* need not belong to $\mathcal{V}(x_{1\ldots u})$. Clearly, (45) implies (46). Finally, given (46), Proposition 3.1 yields (47). ■

Corollary 3.3 *For any fixed $s \in S$, Proposition 3.4 remains valid after replacing π by $(p_{si})_{i \in S}$, wherever appropriate. In particular,*

$$u+r<n, x_u \text{ is an } l\text{-node of order } r \ \Rightarrow\ \emptyset \neq \mathcal{V}_{(s)u}^{\ \ l}(x_{1\ldots n}) = \mathcal{W}_{(s)u}^{\ \ l}(x_{1\ldots u}) \times \mathcal{V}_{(l)}(x_{u+1\ldots n}).$$

Corollary 3.4 *Let $u_i + r_i < u_{i+1}$ $i = 1, \ldots, k-1$, and $u_k + r_k < n$, and suppose $x_{1\ldots n}$ is such that the observations x_{u_i} are l_i-nodes of order r_i, for $i = 1, \ldots, k$. Then*

$$\emptyset \neq \mathcal{V}_{u_1 u_2 \cdots u_k}^{l_1 l_2 \cdots l_k}(x_{1\ldots n}) =$$
$$= \mathcal{W}_{u_1}^{l_1}(x_{1\ldots u_1}) \times \mathcal{W}_{(l_1)u_2}^{\ \ l_2}(x_{u_1+1\ldots u_2}) \times \cdots \times \mathcal{W}_{(l_{k-1})u_k}^{\ \ l_k}(x_{u_{k-1}+1\ldots u_k}) \times \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n}). \tag{48}$$

Proof. By (46), we have

$$\mathcal{V}_{u_i}^{l_i}(x_{1\ldots n}) \neq \emptyset \quad i = 1, \ldots, k.$$

Hence,

$$\emptyset \neq \mathcal{V}_{u_1 u_2 \cdots u_k}^{l_1 l_2 \cdots l_k}(x_{1\ldots n}).$$

By (47),

$$\mathcal{V}_{u_1 u_2 \cdots u_k}^{l_1 l_2 \cdots l_k}(x_{1\ldots n}) = \mathcal{W}_{u_1}^{l_1}(x_{1\ldots u_1}) \times \mathcal{V}_{(l_1)u_2 \cdots u_k}^{\ \ l_2 \cdots l_k}(x_{u_1+1\ldots n}).$$

Apply Corollary 3.3 to get

$$\mathcal{V}_{(l_1)u_2 \cdots u_k}^{\ \ l_2 \cdots l_k}(x_{u_1+1\ldots n}) = \mathcal{W}_{(l_1)u_2}^{\ \ l_2}(x_{u_1+1\ldots u_2}) \times \mathcal{V}_{(l_2)u_3 \cdots u_k}^{\ \ l_3 \cdots l_k}(x_{u_2+1\ldots n}),$$

and repeat similarly to get

$$\mathcal{V}_{(l_i)u_{i+1}\cdots u_k}^{\;\;l_{i+1}\cdots l_k}(x_{u_i+1\ldots n}) = \mathcal{W}_{(l_i)u_{i+1}}^{\;\;l_{i+1}}(x_{u_i+1\ldots u_{i+1}}) \times \mathcal{V}_{(l_{i+1})u_{i+2}\cdots u_k}^{\;\;l_{i+2}\cdots l_k}(x_{u_{i+1}+1\ldots n})$$

for $i = 2, \ldots, k-1$, yielding the desired result. ■

Thus, the assumptions of Proposition 3.4 and Corollary 3.4 establish the existence of piecewise alignments

$$v = (v_1, \ldots, v_{k+1}) \in \mathcal{V}(x_{1\ldots n}), \tag{49}$$

where $v_1 \in \mathcal{W}_{u_1}^{l_1}(x_{1\ldots u_1})$, $v_2 \in \mathcal{W}_{(l_1)u_2}^{\;\;l_2}(x_{u_1+1\ldots u_2}), \ldots,\ v_k \in \mathcal{W}_{(l_{k-1})u_k}^{\;\;l_k}(x_{u_{k-1}+1\ldots u_k}),\ v_{k+1} \in \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n})$. Moreover, for every $i = 1, \ldots, k$, the vectors $w(i) \stackrel{def}{=} (v_1, \ldots, v_i)$ satisfy $w(i) \in \mathcal{W}_{u_i}^{l_i}(x_{1\ldots u_i})$ and $w(i)_{1\ldots u_{i-1}} = w(i-1), \quad i = 2, \ldots, k$. Since $w(i)$ does not depend on $x_{u_i+r_i+1}, \ldots, x_n$ and as long as $x_1, \ldots, x_{u_i+r_i}$ are such that x_{u_i} is a node of order-r_i, an alignment $v(x_{1\ldots n})$ can always be found such that $v_{1\ldots u_i} = w(i)$.

Definition 3.6 *Any alignment of the form in* (49) *is called a piecewise alignment based on nodes* $x_{u_1}, \ldots, x_{u_k}$ *of orders* $r_1, \ldots, r_k$*, respectively.*

Recall that we have previously fixed the selection scheme $\vee$ (33). Based on this selection scheme, we will concern ourselves in §4.2 with proper (Definition 3.3) piecewise (Definition 3.6) alignments (that are based on nodes of possibly non-zero orders) formally defined as follows:

Definition 3.7

$$\begin{aligned} v(x_{1\ldots n}) \;\stackrel{\text{def}}{=}\; & (\vee\mathcal{W}_{u_1}^{l_1}(x_{1\ldots u_1}), \vee\mathcal{W}_{(l_1)u_2}^{\;\;l_2}(x_{u_1+1\ldots u_2}), \ldots, \\ & \vee\mathcal{W}_{(l_{k-1})u_k}^{\;\;l_k}(x_{u_{k-1}+1\ldots u_k}), \vee\mathcal{V}_{(l_k)}(x_{u_k+1\ldots n})) \in \mathcal{V}_{u_1\ldots u_k}^{l_1\ldots l_k}(x_{1\ldots n}), \end{aligned}$$

for $k > 0$*, and* $v(x_{1\ldots n}) \stackrel{\text{def}}{=} \vee\mathcal{V}(x_{1\ldots n})$ *for* $k = 0$*.*

To summarize the above, recall that by defining nodes (of various orders) we aim at extending alignments at infinitum, and we would like to do this for as wide a class of HMM's with irreducible and aperiodic hidden layers as possible. Requiring l-nodes of order 0 immediately restricts the transition probabilities by requiring $p_{lj} > 0$ for $\forall j \in S$. However, this restriction disappears with the introduction of nodes of order r for sufficiently large r. Indeed, suppose that $\forall u\ 0 < u \leq n$, we have $\delta_u(j) > 0\ \forall j \in S$ (which in particular implies $f_j(x_u) > 0\ \forall j \in S\ \forall u\ 0 < u \leq n$). Then, x_u being an l-node of order r, and irreducibility of the underlying chain, imply $p_{lj}^{(r)}(u) > 0\ \forall j \in S$. The latter in turn implies that $r_{lj} > 0$ for every $j \in S$, where r_{lj} is the lj^{th} entry of $\mathbb{P}^r$. Thus, having an l-node of order r for **some** r does not impose any restriction on $\mathbb{P}$: by virtue of irreducibility and aperiodicity of $\mathbb{P}$, there always exists r_0 such that P has all of its entries positive for every $r \geq r_0$.

3.3 Barriers

By Corollary 3.4, x_u being a node of order r fixes the alignment up to u for any possible continuation of $x_{1\dots u+r}$. However, changing the value of an observation before x_{u+r+1}, say x_1 or x_{u+r}, can prevent x_u from being the node. Moreover, in general nothing guarantees that for an arbitrary prefix $x'_{1\dots w} \in \mathcal{X}^w$, $w+u$-th element of $(x'_{1\dots w}, x_{1\dots u+r})$ would be a node of order r. On the other hand, a block of observations $x^b_{1\dots k} \in \mathcal{X}^k$ $(k \geq r)$ can be such that for any $w > 0$ and for any $x'_{1\dots w} \in \mathcal{X}^w$, $w+k-r$-th element of $(x'_{1\dots w}, x^b_{1\dots k})$ is a node of order r. $x^b_{1\dots k}$ in that case will be called *a barrier*.

Definition 3.8 *A block of observations $x^b_{1\dots k} \in \mathcal{X}^k$ $(k \geq r)$ is called an l-barrier of order r and length k if for any $w > 0$ and for any $x'_{1\dots w} \in \mathcal{X}^w$, $w+k-r$-th element of $(x'_{1\dots w}, x^b_{1\dots k})$ is an l-node of order r.*

3.4 Existence of barriers

In this section, we state the main technical result of the paper. For each $i \in S$, we denote by $G_i = \cap_{G\text{-closed},\ P_i(G)=1} G$, the support of P_i.

Definition 3.9 *We define a subset $C \subset S$ to be a cluster, if, simultaneously,*

$$\min_{j\in C} P_j(\cap_{i\in C} G_i \cap \{x \in \mathcal{X} :\ f_i(x) > 0\}) > 0, \quad \text{and} \quad P_j(\cap_{i\in C} G_i) = 0 \quad \forall j \notin C.$$

(Note that C is well-defined that is, if the first condition is satisfied with one choice of density functions f_i, it will certainly be satisfied with any other choice of densities g_i since $\lambda(\{x \in \mathcal{X} :\ f_i \neq g_i\}) = 0$ for all $i \in S$.) Hence, a cluster is a maximal subset of states such that the corresponding emission distributions have a "detectable" intersection of their supports $G_C = \cap_{i\in C} G_i$. Clusters need not necessarily be disjoint and a cluster can consist of a single state. In this latter case such a state is not hidden: Any emission from this state reveals it. If $K = 2$, then, for an HMM, there is only one cluster (otherwise the underlying Markov chain would not be hidden as all observations reveal their states). In many cases in practise there is only one cluster, that is S.
A proof of Lemma 3.1 below is given in Appendix 5.1.

Lemma 3.1 *Assume that for each state $l \in S$,*

$$P_l\left(x : f_l(x) \max_j\{p_{jl}\} > \max_{i,i\neq l}\{f_i(x) \max_j\{p_{ji}\}\}\right) > 0. \tag{50}$$

Moreover, assume that there exist a cluster $C \subset S$ and a finite integer $m < \infty$ such that the m-th power of the sub-stochastic matrix $\mathbb{Q} = (p_{ij})_{i,j\in C}$ has all of its entries non-zero. Then, for some integers M and r, $M > r \geq 0$, there exist a set $B = B_1 \times \cdots \times B_M \subset \mathcal{X}^M$, an M-tuple of states $q_{1\dots M} \in S^M$, and a state $l \in S$, such that every vector $y \in B$ is an l-barrier of order r , $q_{M-r} = l$ and

$$\mathbf{P}\Big((X_1, \dots, X_M) \in \mathcal{Y}\Big| Y_1 = q_1, \dots, Y_M = q_M\Big) > 0, \quad \mathbf{P}(Y_1 = q_1, \dots, Y_M = q_M) > 0.$$

Lemma 3.1 implies that $\mathbf{P}\big((X_1,\ldots,X_M)\in B\big)>0$. Also, since every element of B is a barrier of order r, the ergodicity of X therefore guarantees a.e. realization of X to contain infinitely many l-barriers of order r. Hence, a.e. realization of X also has infinitely many l-nodes of order r.

3.4.1 Separated barriers

If we were to apply Corollary 3.4 to a realization with infinitely many l-nodes of order r, we would first need to ensure that $u_{i+1}>u_i+r$ for $i=1,2,\ldots$, where u_i's are the locations of the nodes. Obviously, one can easily select a subsequence of those nodes to enforce this condition. For certain technical reasons related to the construction of the infinite alignment process (§4), we, however, choose first to define special barriers for which the above "separation" condition is always satisfied. Then, we give a formal statement (Lemma 3.2 below) guaranteeing that these separated barriers occur also infinitely often. Let $B\subset\mathcal{X}^M$ and M and r be as in Lemma 3.1. Assume that for some $l\in S$ and some $j>0$ $x_{j\ldots j+M-1}\in B$, i.e. $x_{j\ldots j+M-1}$ is an l-barrier of order r, and $x_{j+M-r-1}$ is an l-node of order r. However, it might happen that for some i, $j\le i\le j+r$, $x_{i\ldots i+M-1}$ is also in B. Then $x_{i+M-r-1}$ is another node of order r. In this case, $i+M-r-1-(j+M-r-1)\le r$ and Corollary 3.4 can not be used (in the presence of ties) with these two nodes simultaneously.

Definition 3.10 *Let $B^*\subset\mathcal{X}^N$ such that all its elements are l-barriers of order r for some $l\in S$ and $r\le N$. We say that $x^b_{1\ldots N}\in B^*$ is separated (relative to B^*) if for any w, $1\le w\le r$, and for any $x'_{1\ldots w}\in\mathcal{X}^w$ the concatenated block $(x'_{1\ldots w},x^b_{1\ldots N-w})\notin B^*$.*

Thus, roughly, a barrier is separated, if it is at least $r+1$ steps apart from any preceding B^* barrier.

Suppose $B\subset\mathcal{X}^M$ is such that every $x^b_{1\ldots M}\in B$ is a barrier. The barriers from B need not in general be separated. However, it can be possible to extend these barriers to make them separated relative to their own set B^*. For example, suppose further that there exists $x\in\mathcal{X}$ such that no $y\in B$ contains x, i.e. $x^b_i\ne x$ $i=1,\ldots,M$. All the elements of $B^*\stackrel{\rm def}{=}\{x\}\times B$ are evidently barriers, and moreover, they are now also separated (relative to B^*).

This will be used in Appendix §5.2 to prove Lemma 3.2 given below, and which states that under the assumptions of Lemma 3.1, separated barriers are also guaranteed to occur. In other words, a.e. realization of X has infinitely many separated barriers.

Lemma 3.2 *Suppose the assumptions of Lemma 3.1 are satisfied. Then, for some integers M and r, $M>r\ge 0$, there exist a set $B=B_1\times\cdots\times B_M\subset\mathcal{X}^M$, an M-tuple of states $q_{1\ldots M}\in S^M$, and a state $l\in S$, such that every vector $y\in B$ is a separated (relative to B) l-barrier of order r, $q_{M-r}=l$ and*

$$\mathbf{P}\Big((X_1,\ldots,X_M)\in B\Big|Y_1=q_1,\ldots,Y_M=q_M\Big)>0,\quad \mathbf{P}(Y_1=q_1,\ldots,Y_M=q_M)>0.$$

3.4.2 Counterexamples

The condition on C in Lemma 3.1 might seem technical and even unnecessary. We next give an example of an HMM where the cluster condition is not fulfilled and no barriers can occur. Then, we will modify the example (Examples 3.12 3.13) to enforce the cluster condition and consequently gain barriers.

Example 3.11 *Let $K = 4$ and consider an ergodic Markov chain with transition matrix*

$$\mathbb{P} = \begin{pmatrix} \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Let the emission distributions be such that (50) *is satisfied and $G_1 = G_2$ and $G_3 = G_4$ and $G_1 \cap G_3 = \emptyset$. Hence, in this case there are two disjoint clusters $C_1 = \{1, 2\}$, $C_2 = \{3, 4\}$. The matrices $\mathbb{Q}_i$ corresponding to C_i, $i = 1, 2$ are*

$$\mathbb{Q}_1 = \mathbb{Q}_2 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Evidently, the cluster assumption of Lemma 3.1 is not satisfied. Note also that the alignment cannot change (in one step) its state to the other one of the same cluster. Due to the disjoint supports, any observation indicates the corresponding cluster. Hence any sequence of observations can be regarded as a sequence of blocks emitted from alternating clusters. However, the alignment inside each block stays constant.

In order to see that no x_u can be a node (of any order) for $1 \leq u < n$, recall $t(u, j)$ (17) *and $t(u, j)^{(r)}$* (40)*, and Proposition 3.4. Specifically, note that in this setting for any $j \in S$ $t(u, j)$ contains exactly one element, hence for any $r \geq 1$, $t(u, j)^{(r)}$ defines a function from S to S. Now, it is easy to see that depending on x_u, $t(u, j)$ belongs to a single cluster $C(x_u)$ for all $j \in S$. In particular, there are $i, j \in C' \subset S$ for some cluster C' such that $i \neq j$. Given this particular transition matrix, evidently $t(u, i) \neq t(u, j)$. Hence, x_u cannot be a (zero order) node (by* (44)*). Now, starting with $u+1$ (instead of u), the same argument establishes that for some $i, j \in S$, $t(u + 1, i) \neq t(u + 1, j)$ but are in one cluster. Applying the same argument again but now to $t(u + 1, i)$ and $t(u + 1, j)$, we get that $t(u, t(u + 1, i)) \neq t(u, t(u + 1, j))$, i.e. $t^{(2)}(u, i) \neq t^{(2)}(u, j)$. Consequently x_u cannot be a first order node* (44)*; and so forth and so on recursively for any r such that $0 \leq r < n-u$.*

Example 3.12 *Let us modify the HMM in Example 3.11 to ensure the assumptions of Lemma 3.1 hold. At first, let us change the transition matrix. Let $0 < \epsilon < \frac{1}{2}$ and consider the Markov chain Y with transition matrix*

$$\begin{pmatrix} \frac{1}{2} - \epsilon & \epsilon & 0 & \frac{1}{2} \\ \epsilon & \frac{1}{2} - \epsilon & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Let the emission distributions be as in the previous example. In this case, the cluster C_1 satisfies the assumption of Lemma 3.1. As previously, every observation indicates its cluster. Unlike in the previous example, nodes are now possible. To be concrete, let $\epsilon = 1/4$, $f_1(x) = \exp(-x)_{x\geq 0}$, $f_2(x) = 2\exp(-2x)_{x\geq 0}$, and $f_3(x) = \exp(x)_{x\leq 0}$, $f_4(x) = 2\exp(2x)_{x\leq 0}$. It can then be verified that, for example, if $x_1 = 1$, $x_2 = 1$ then x_1 is a 1-node of order 2. Indeed, in that case any element of $B = (0, +\infty) \times (\log(2), +\infty) \times (0, +\infty)$ is a 1-barrier of order 2.

Example 3.13 *Another way to modify the HMM in Example 3.11 to enforce the assumptions of Lemma 3.1 is to change the emission probabilities. Assume that the supports G_i, $i = 1, \ldots, 4$ are such that $P_j(\cap_{i=1}^4 G_i) > 0$ for all $j \in S$, and (50) holds. Now, the model has only one cluster that is $S = \{1, \ldots, 4\}$. Since the matrix $\mathbb{P}^2$ has all its entries positive, the conditions of Lemma 3.1 are now satisfied. A barrier can now be constructed. For example, the following block of observations,*

$$z_1, z_2, z_3, x_1, \ldots, x_k, z_1', z_2', z_3', \tag{51}$$

where $z_i, z_i' \in \cap_{j=1}^4 G_j$, $i = 1, 2, 3$, $x_i \in \mathcal{X}$, $i = 1, \ldots k$ and k is sufficiently large, is a barrier (see proof of Lemma 3.1). The construction of barriers in this case is possible because of the observations z_i and z_i'. These observations can be emitted from any state (i.e. from any distribution P_i, $i = 1, \ldots, 4$) and therefore do not indicate any proper subsets of S. They play a role of a buffer allowing a change in the alignment from a given state to any other state (in 3 steps). The HMM in Example 3.11 does not have r-order nodes, because such buffers do not arise. The cluster assumption in Lemma 3.1 makes these buffers possible.

4 Alignment process

Let $x_{1\infty} = x_1, x_2, \ldots$ be a realization of X. If for some $r < \infty$ $x_{1\infty}$ contains infinitely many r-order nodes, then Corollary 3.4 paves the way for defining an infinite alignment for $x_{1\infty}$.

4.1 Preliminaries

Throughout this Section, we work under the assumptions of Lemma 3.1. Let $M \geq 0$, $B \subset \mathcal{X}^M$, $r \geq 0$, and $l \in S$, $q = q_{1\ldots M} \in S^M$ as promised by Lemma 3.2. Then, for every $n \geq 1$,

$$\mathbf{P}\Big((Y_n, \ldots, Y_{n+M-1}) = q\Big) > 0, \quad \mathbf{P}\Big((X_n, \ldots, X_{n+M-1}) \in B \Big| (Y_n, \ldots, Y_{n+M-1}) = q\Big) > 0$$

hence every $x_{n\ldots n+M-1} \in B$ is a separated (relative to B) l-barrier of order r.
Denote $\mathbf{P}\Big((X_n, \ldots, X_{n+M-1}) \in \mathcal{Y} | (Y_n, \ldots, Y_{n+M-1}) = q\Big)$ by γ^*. Thus, $\gamma^* > 0$, and define

$$U_n \stackrel{\text{def}}{=} (X_n, \ldots, X_{n+M-1}), \quad D_n \stackrel{\text{def}}{=} (Y_n, \ldots, Y_{n+M-1}). \tag{52}$$

Let $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(Y_1, \ldots, Y_n, X_1, \ldots, X_n)$. Define stopping times $\nu_0, \nu_1, \nu_2, \ldots$, $R_0, R_1, R_2, \ldots$, and $\vartheta_0, \vartheta_1, \vartheta_2, \ldots$, of the filtration $\{\mathcal{F}_{n+M-1}\}_{n=1}^{\infty}$ as follows:

$$\nu_0 \stackrel{\text{def}}{=} \min\{n \geq 1 : U_n \in B, D_n = q\}, \quad \nu_i \stackrel{\text{def}}{=} \min\{n > \nu_{i-1} : U_n \in B, D_n = q\}; \tag{53}$$

$$\vartheta_0 \stackrel{\text{def}}{=} \min\{n \geq 1 : U_n \in B\}, \quad \vartheta_i \stackrel{\text{def}}{=} \min\{n > \vartheta_{i-1} : U_n \in B\}; \tag{54}$$

$$R_0 \stackrel{\text{def}}{=} \min\{n \geq 1 : \ D_n = q\}, \quad R_i \stackrel{\text{def}}{=} \min\{n > R_{i-1} : \ D_n = q\}. \tag{55}$$

We use the convention $\min \emptyset = 0$ and $\max \emptyset = -1$. Note the difference between ν and R and ϑ: The stopping times ϑ are observable via X process alone; the stopping times R are observable via the Y process alone; the sopping times ν already require knowledge of the full two-dimensional process (X, Y). Clearly $\vartheta_i \leq \nu_i$, and $R_i \leq \nu_i$.

From (55), it follows that the random variables $R_0, (R_1 - R_0), (R_2 - R_1), \ldots$ are independent and $(R_1 - R_0), (R_2 - R_1), \ldots$ are identically distributed. The same evidently holds for the random variables $\nu_0, (\nu_1 - \nu_0), (\nu_2 - \nu_1), \ldots$.

Proposition 4.1 *For any initial distribution π' of Y, we have $E_{\pi'}\nu_0 < \infty$ and $E_{\pi'}(\nu_1 - \nu_0) < \infty$.*

Proposition 4.1 above is an intuitive result; a proof is provided in Appendix §5.3. To every ν_i, $i = 0, 1, \ldots$ there corresponds an l-barrier of order r. This barrier extends over the interval $[\nu_i, \nu_i + M - 1]$. By Definition 3.8, X_{τ_i} is an l-node of order r, where

$$\tau_i \stackrel{\text{def}}{=} \nu_i + (M - 1) - r, \quad i = 0, 1, \ldots \tag{56}$$

Define

$$T_0 \stackrel{\text{def}}{=} \tau_0, \quad T_i \stackrel{\text{def}}{=} \tau_i - \tau_{i-1} = \nu_i - \nu_{i-1}, \ i = 1, 2, \ldots. \tag{57}$$

Proposition 4.1 says that $E_{\pi'}T_1 < \infty$, $E_{\pi'}T_0 < \infty$, where π' is any initial distribution of Y. Thus, T_i, $i = 0, 1, \ldots$ correspond to a delayed renewal process (for a general reference see, for example, [7]).

Let $u_0, u_1, u_2, \ldots$ be the locations of the order r l-nodes corresponding to the stopping times ϑ, i.e.

$$u_i = \vartheta_i + (M - 1) - r, \quad i = 0, 1, 2, \ldots. \tag{58}$$

Clearly, every τ_i is also u_j for some $j \geq i$. Also, since the barriers are separated, $u_i > u_{i-1} + r$.

4.2 Alignments

We next specify the alignments $v(x_{1\ldots n}) \in \mathcal{V}(x_{1\ldots n})$ and define $v(x_{1\ldots\infty})$ as well as the measures $\hat{P}_l^n$ corresponding to $v(x_{1\ldots n})$.

Let $k(x_{1\ldots n})$ be the number of $x_{u_0},\ x_{u_1},\ldots,\ x_{u_{k(x_{1\ldots n})-1}}$, *all* l nodes of order r such that $u_i > u_{i-1}+r$ for $i=1,\ldots,k(x_{1\ldots n})-1$, and $u_{k(x_{1\ldots n})-1}+r<n$. Recall (Definition 3.7) that based on the selection $\vee$ (33), we single out the following proper piecewise alignment:

$$\begin{aligned} v(x_{1\ldots n}) &= (\vee\mathcal{W}^l_{u_0}(x_{1\ldots u_0}), \vee\mathcal{W}_{(l)u_1}^{\ l}(x_{u_0+1\ldots u_1}),\ldots,\\ &\quad \vee\mathcal{W}_{(l)u_{k-1}}^{\ l}(x_{u_{k-2}+1\ldots u_{k-1}}), \vee\mathcal{V}_{(l)}(x_{u_{k-1}+1\ldots n}))\in\mathcal{V}^{l\ldots l}_{u_0\ldots u_{k-1}}(x_{1\ldots n}), \end{aligned}$$

for $k=k(x_{1\ldots n})>0$, and $v(x_{1\ldots n})=\vee\mathcal{V}(x_{1\ldots n})$ for $k=0$. Corollary 3.4 makes it possible to define the *infinite proper piecewise alignment* that will be consistent with Definition 3.7 (in the sense of (59) below). Namely, we state

Definition 4.1

$$v(x_{1\ldots\infty}) \stackrel{\text{def}}{=} (\vee\mathcal{W}^l_{u_0}(x_{1\ldots u_0}), \vee\mathcal{W}_{(l)u_1}^{\ l}(x_{u_0+1\ldots u_1}),\ldots,)$$

for all $x_{1\infty}$ that contain infinitely many x_{u_0}, x_{u_1}, ..., l-nodes of order r, which is the case a.s. (Lemmas 3.1 and 3.2). (For all other realizations, let us adopt $v(x_{1\ldots\infty})\stackrel{\text{def}}{=}(\vee\mathcal{W}^l_{u_0}(x_{1\ldots u_0}), \vee\mathcal{W}_{(l)u_1}^{\ l}(x_{u_0+1\ldots u_1}),\ldots,\vee\mathcal{W}_{(l)u_{k-1}}^{\ l}(x_{u_{k-2}+1\ldots u_{k-1}}),1,1,\ldots)$, where k is the total number of l nodes of order r in the given realization.)

Note that for every x_{u_i} observed in $(x_1,\ldots,x_n)$

$$v(x_1^\infty)_{1\cdots u_i} = v(x_1,\ldots,x_n)_{1\cdots u_i}. \tag{59}$$

Let us now formally define the empirical measures $\hat{P}_l^n$ which are central to this theory:

Definition 4.2 *Let $=V'_{1\ldots n}=v(X_1,\ldots,X_n)$ (where v is as in Definition 3.7). For each state $l\in S$ that appears in V'_1, V'_2, ... V'_n define the empirical l-measure*

$$\hat{P}_l^n(A,X_{1\ldots n})\stackrel{\text{def}}{=}\frac{\sum_{i=1}^n I_{A\times l}(X_i,V'_i)}{\sum_{i=1}^n I_l(V'_i)},\quad A\in\mathcal{B}.$$

For other $l\in S$ (i.e. such that $l\neq V'_i$ for $i=1,\ldots,n$), define $\hat{P}_l^n$ to be an arbitrarily chosen (probability) measure P^*.

The infinite alignment allows us to define the *alignment process*:

Definition 4.3 *The encoded process $V\stackrel{\text{def}}{=}v(X)$ will be called the* alignment process.

(Of course, the definition of V above is sensible only because X has infinitely many u_i-s a.s..) We shall also consider the 2-dimensional process

$$Z\stackrel{\text{def}}{=}(X,V).$$

Using Z, we define a related quantity Q_l^n as follows: Let $V_1,\ldots,V_n$ be the first n elements of the alignment process. In general

$$v(x_1^\infty)_{1\ldots n}\neq v(x_1,\ldots,x_n),$$

hence V'_i need not equal V_i. For every $l\in S$, we define

$$\hat{Q}_l^n(A,Z_{1\ldots n})\stackrel{\text{def}}{=}\frac{\sum_{i=1}^n I_{A\times l}(X_i,V_i)}{\sum_{i=1}^n I_l(V_i)}=\frac{\sum_{i=1}^n I_{A\times l}(Z_i)}{\sum_{i=1}^n I_l(V)},\quad A\in\mathcal{B}.$$

(As in Definition 4.2, if $l\neq V_i$, $i=1,\ldots,n$, then $\hat{Q}_l^n\stackrel{\text{def}}{=}P^*$.)

4.3 Regenerativity

To prove our main theorem, we use the fact that Z is a regenerative process (for a general reference, see, for example, [3]):

Proposition 4.2 *The processes V, X, and Z are regenerative with respect to the sequence of stopping times τ_i.*

A proof is given in Appendix §5.4.

Recall (§4.1) B, the set of separated l-barriers of order r, and the corresponding state sequence q. Let

$$P^r_{q_i} \propto P_{q_i} I_{B_i}, \quad i = 1, \ldots, M.$$

Thus, $P^r_{q_i}$ is the measure $P^r_{q_i}$ conditioned on B_i, i-th component of B. Recall also that $q_{M-r} = l$.

Define new processes

$$Y^r \stackrel{\text{def}}{=} (Y^r_i)_{i=1}^{\infty}, \quad \text{where} \quad Y^r_1 = q_{M-r+1}, \ldots, Y^r_r = q_M, \quad \text{and} \quad Y^r_{r+1}, Y^r_{r+2}, \ldots \tag{60}$$
is an S-valued Markov chain with transition probability matrix $\mathbb{P}$ and initial distribution $(p_{q_N j})_{j \in S}$;

$X^r \stackrel{\text{def}}{=} (X^r_i)_{i=1}^{\infty}$ is a modified HMM with Y^r as its underlying Markov chain and $P_{Y^r_i}$ if $i > r$, and $P^r_{q_{N-r+i}}$ if $1 \le i \le r$, as its emission distributions;

$$V^r \stackrel{\text{def}}{=} (V^r_i)_{i=1}^{\infty} \stackrel{\text{def}}{=} v(X^r), \text{ where } v \text{ is as in Definition 4.1;} \tag{61}$$

$$Z^r \stackrel{\text{def}}{=} (X^r, V^r). \tag{62}$$

Note that the process X^r is not exactly an HMM as defined in Definition 2.1 because the first r-emissions are generated from distributions that differ from the distributions of the subsequent emissions. However, conditioned on the underlying Markov Chain Y^r, all emissions are still independent. Also note that in the definition of V^r, the alignment is still based on the original HMM X, i.e. the definition of $v(x_1, \ldots, x_n)$ relies on the distributions P_{q_1}, P_{q_2},..., P_{q_n} (given $Y_{1\ldots n} = q_{1\ldots n}$).

Finally, note that for $r = 0$, the process Y^0 is essentially our original Markov chain except for the initial distribution that is now $(p_{l\,j})_{j \in S}$ instead of π. Similarly, X^0 is the HMM in the sense of Definition 2.1 with Y^0 as its underlying Markov chain. Therefore, Z^0 is the process Z with $(p_{l\,j})_{j \in S}$ as the initial distribution of its Y-component.

Finally we define analogues of ν_0 and τ_0:

$$\nu^r_0 \stackrel{\text{def}}{=} \min\Big\{n \ge 1 : (Y^r_n, \ldots, Y^r_{n+M-1}) = q, \quad (X^r_n, \ldots, X^r_{n+M-1}) \in B\big)\Big\}$$
$$\tau^r_0 \stackrel{\text{def}}{=} \nu^r_0 + (M-1) - r. \tag{63}$$

Note that the random variable τ^r_0 has the same law as T_i (57), $i \ge 1$. Since the barriers from B are separated (Definition 3.10, Lemma 3.2), then $\nu^r_0 > r$. This means that the

laws of ν_0^r, τ_0^r, $\nu_0 + r$, and $\tau_0 + r$ would all be equal if the initial distribution of Y were $(p_{q_M\, l})_{l\in S}$. Recalling that any initial distribution π' of Y yields $E_{\pi'}(\nu_0) < \infty$ (Proposition 4.1), we obtain

$$ET_1 = E\tau_0^r = E_{q_M}(\nu_0 + (M-1) - r + r) < \infty. \tag{64}$$

The above observations will allow us to prove (see Appendix §5.5) the following theorem which is *the main result of the paper*:

Theorem 4.4 *If X satisfies the assumptions of Lemma 3.1, then there exist probability measures Q_l, $l \in S$, such that*

$$\hat{P}_l^n \Rightarrow Q_l, \quad a.s., \quad \hat{Q}_l^n \Rightarrow Q_l, \quad a.s.$$

and for each $A \in \mathcal{B}$,

$$Q_l(A) = \frac{\sum_{i=1}^{\infty} \mathbf{P}(Z_i^r \in A \times l, \tau_0^r \geq i)}{\sum_{i=1}^{\infty} \mathbf{P}(V_i^r = l, \tau_0^r \geq i)}. \tag{65}$$

where V^r, Z^r, and τ_0^r are defined in (61), (62), *and* (63), *respectively.*

Corollary 4.1 *Suppose X satisfies the assumptions of Lemma 3.1 with $r = 0$. Then, for each $l \in S$ (65) takes form*

$$Q_l(A) = \frac{\sum_{j=1}^{\infty} \mathbf{P}_l(Z_j \in A \times i, \tau_0 \geq j)}{\sum_{j=1}^{\infty} \mathbf{P}_l(V_j = i, \tau_0 \geq j)}, \tag{66}$$

where $\mathbf{P}_l$ corresponds to the Y process initialized with $(p_{l\, j})_{j\in S}$ instead of π.

5 Appendix

5.1 Proof of Lemma 3.1

Proof. The proof below is a rather direct construction which is, however, technically involved. In order to facilitate the exposition of this proof, we have divided it into 18 short parts as outlined below:

I. - §5.1.1 - Maximal probability transitions p_i^* and maximal likelihood ratio A.

II. Construction of

- (§5.1.2) auxiliary subsets $\mathcal{X}_l \in \mathcal{X}$, (68);
- (§5.1.3) a special set $Z \subset \mathcal{X}$, (70), (71);
- (§5.1.4) auxiliary sequences **s** (72), **a** (73), and **b** (5.1.4) of states in S;
- (§5.1.5) k, the number of **s** cycles inside the s-path;
- (§5.1.6) the s-path (78), a prototype of the required sequence $q_{1\ldots M}$;
- (§5.1.7) the required barrier (79).

III. Proving the barrier construction (79):

(§5.1.8) $\alpha, \beta, \gamma, \eta$-notation for commonly used maximal partial likelihoods;

(§5.1.9) a bound (85) on β;

(§5.1.10) bounds (86), (87), (88), and (89) on common likelihood ratios;

(§5.1.11)$\gamma_j \leq const \times \gamma_1$;

(§5.1.12) Further bounds (104), (105) on likelihoods;

(§5.1.13) $\eta_j \leq const \times \eta_1$;

(§5.1.14) a special representation of η_1 (107);

(§5.1.15) an implication of (103) and (107) for $\delta_1(x_{lL})$;

x_{kL} is a $(kL + m + P)$-order 1-node:

(§5.1.16) proof

(§5.1.17) proof of an auxiliary inequality (114).

IV. (§5.1.18) Completion of the s-path to $q_{1\ldots M}$ and conclusion.

5.1.1 Maximal probability transitions p_i^* and maximal likelihood ratio A.

Let

$$p_i^* = \max_{j \in S}\{p_{j\,i}\},\ i \in S, \quad \text{and} \quad A = \max_{i \in S} \max_{j \in S} \left\{ \frac{p_i^*}{p_{ji}} : p_{ji} > 0 \right\} \tag{67}$$

be defined as above.

5.1.2 $\mathcal{X}_l \subset \mathcal{X}$.

It follows from the assumption (50) and finiteness of S that there exists an $\epsilon > 0$ such that for all $l \in S$

$$P_l(\mathcal{X}_l) > 0, \quad \text{where } \mathcal{X}_l \stackrel{\text{def}}{=} \left\{ x \in \mathcal{X} : \max_{i, i \neq l}\{p_i^* f_i(x)\} < (1-\epsilon) p_l^* f_l(x) \right\}. \tag{68}$$

(Note that $p_l^* > 0$ for all $l \in S$ by irreducibility of Y.) Also note that the sets $\mathcal{X}_l, l \in S$ are disjoint and have positive reference measure $\lambda(\mathcal{X}_l) > 0$.

5.1.3 $\mathcal{Z} \subset \mathcal{X}$ and $\delta - K$ bounds on cluster densities f_i, $i \in C$

Let C be a cluster as in the assumptions of the Lemma with the corresponding sub-stochastic matrix $\mathbb{Q}$. The existence of C implies the existence of a set $\hat{\mathcal{Z}} \subset G_C (= \cap_{i \in C} G_i)$ and $\delta > 0$, such that $\lambda(\hat{\mathcal{Z}}) > 0$, and $\forall z \in \hat{\mathcal{Z}}$, the following statements hold:

(i) $\min_{i \in C} f_i(z) > \delta$;

(ii) $\max_{j \notin C} f_j(z) = 0$.

Indeed, if no $\hat{\mathcal{Z}}$ and $\delta > 0$ existed with property (i), we would have $\lambda\left(\cap_{i\in C}(G_i \cap \{z \in \mathcal{X} :\ f_i(z) > 0\})\right) = 0$, contradicting the first defining property of cluster: $P_j(G_C \cap_{i\in C} \{x \in \mathcal{X} :\ f_i(x) > 0\}) > 0$ (with any $j \in C$). Now, if $\hat{\mathcal{Z}}$ did not satisfy (ii), we would remove from it $\hat{\mathcal{Z}} \cap \cup_{i\notin C}\{z \in \mathcal{X} :\ f_i(z) > 0\}$ as this would not reduce its λ measure. This is due to the second condition in the definition of cluster which implies $\lambda(G_C \cap \{z \in \mathcal{X} :\ f_i(z) > 0\}) = 0$ for all $i \notin C$.

Evidently, $K > 0$ can be chosen sufficiently large to make $\lambda(\{z \in \mathcal{X} :\ f_i(z) \geq K\})$ arbitrarily small, and in particular, to guarantee that $\lambda(\{z \in \mathcal{X} :\ f_i(z) \geq K\}) < \frac{\lambda(\hat{\mathcal{Z}})}{|C|}$. Clearly then, redefining $\hat{\mathcal{Z}} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \cap \{z \in \mathcal{X} :\ f_i(z) < K,\ i \in C\}$ preserves $\lambda(\hat{\mathcal{Z}}) > 0$. Next, consider

$$\lambda(\hat{\mathcal{Z}} \backslash (\cup_{l\in S}\mathcal{X}_l)). \tag{69}$$

If (69) is positive, then define

$$\mathcal{Z} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \backslash (\cup_{l\in S}\mathcal{X}_l). \tag{70}$$

If (69) is zero, then there must be $s \in C$ such that

$$\lambda(\hat{\mathcal{Z}} \cap \mathcal{X}_s) > 0$$

and in this case, let

$$\mathcal{Z} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \cap \mathcal{X}_s. \tag{71}$$

Such $s \in S$ must clearly exist since $\lambda(\hat{\mathcal{Z}}) > 0$ but $\lambda(\hat{\mathcal{Z}} \backslash (\cup_{l\in S}\mathcal{X}_l)) = 0$. To see that s must necessarily be in the cluster C, note $\forall s \notin C$, $f_s(z) = 0\ \forall z \in \hat{\mathcal{Z}}$, which implies $\hat{\mathcal{Z}} \cap \mathcal{X}_s = \emptyset$.

5.1.4 Sequences s, a, and b of states in S

Let us define an auxiliary sequence of states q_1, q_2, and so on, as follows: If (69) is zero, that is, if $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$ for some $s \in C$, then define $q_1 = s$, otherwise let q_1 be an arbitrary state in C. Let q_2 be a state with maximal probability of transition to q_1, i.e.: $p_{q_2\, q_1} = p^*_{q_1}$ (see (67) for the p^* notation). Suppose $q_2 \neq q_1$. Then find q_3 with $p_{q_3\, q_2} = p^*_{q_2}$. If $q_3 \notin \{q_1, q_2\}$, find $q_4 :\ p_{q_4\, q_3} = p^*_{q_3}$, and so on. Let U be the first index such that $q_U \in \{q_1, \ldots, q_{U-1}\}$, that is, $q_U = q_T$ for some $T < U$. This means that there exists a sequence of states $\{q_T, \ldots, q_U\}$ such that

- $q_T = q_U$
- $q_{T+i} = \arg\max_j p_{j q_{T+i-1}}, \quad i = 1, \ldots, U - T.$

To simplify the notation and without loss of generality, assume $q_U = 1$. Reorder and rename the states as follows:

$$s_1 := q_{U-1},\ s_2 := q_{U-2}, \ldots, s_i := q_{U-i}, \ldots, s_L := q_T = 1 \quad i = 1, \ldots, L \stackrel{\text{def}}{=} U - T, \tag{72}$$

$$a_1 := q_{T-1},\ a_2 := q_{T-2}, \ldots, a_P := q_1, \quad P \stackrel{\text{def}}{=} T - 1. \tag{73}$$

Hence,

$$\{q_1, \ldots, q_{T-1}, q_T, q_{T+1}, \ldots, q_{U-1}, q_U\} = \{a_P, \ldots, a_1, 1, s_{L-1}, \ldots, s_1, 1\}.$$

Note that if $T = 1$, then $P = 0$ and $\{q_1, \ldots, \ldots, q_{U-1}, q_U\} = \{1, s_{L-1}, \ldots, s_1, 1\}$. We have thus introduced special sequences $\mathbf{a} = (a_1, a_2, \ldots, a_P)$ and $\mathbf{s} = (s_1, s_2, \ldots, s_{L-1}, 1)$. Clearly,

$$\begin{aligned} p_{s_{i-1}\, s_i} &= p^*_{s_i}, \quad i = 2, \ldots, L, \quad p^*_{s_1} = p_{1\, s_1} \\ p_{a_{i-1}\, a_i} &= p^*_{a_i}, \quad i = 2, \ldots, P, \quad p^*_{a_1} = s_L = 1. \end{aligned} \tag{74}$$

Next, we are going to exhibit $\mathbf{b} = (b_1, \ldots b_R)$, another auxiliary sequence for some $R \geq 1$, characterized as follows:

(i) $b_R = 1$;

(ii) there exists $b_0 \in C$ such that $p_{b_0\, b_1} p_{b_1\, b_2} \cdots p_{b_{R-1}\, b_R} > 0$

(iii) if $R > 1$, then $b_{i-1} \neq b_i$ for every $i = 1, \ldots, R$.

Thus, the path $b_1, \ldots b_{R-1}, b_R$ connects cluster C to state 1 in R steps. Let us also require that R be minimal such. Clearly such $\mathbf{b}$ and b_0 do exist due to irreducibility of Y. Specifically, for any two states in S in general, and for any state in C and state 1 in particular, there exists a (finite) transition path of a positive probability. Note also that minimality of R guarantees (iii) (in the special case of $R = 1$ it may happen that $b_1 = 1 \in S$ and $p_{1\,1} > 0$, in which case b_0 can be taken to be also 1).

5.1.5 k, the number of s cycles inside the s-path

Let $\mathbb{Q}^m$ be the m-th power of the sub-stochastic matrix $\mathbb{Q} = (p_{ij})_{i,j \in C}$; let q_{ij} be the entries of $\mathbb{Q}^m$. By the assumption, $q_{ij} > 0$ for every $i, j \in C$. This means that for every $i, j \in C$, there exists a path from i to j of length m that has a positive probability. Let q^*_{ij} be the probability of a maximum probability path from i to j. In other words, for every $i, j \in C$, there exist states $w_1, \ldots, w_{m-1} \in C$ such that

$$p_{i w_1} p_{w_1 w_2} \cdots p_{w_{m-1} w_{m-1}} p_{w_{m-1} j} = q^*_{ij} > 0. \tag{75}$$

Denote by q

$$\min_{i,j \in C} q^*_{i\,j} > 0. \tag{76}$$

Next, choose k sufficiently large for the following to hold:

$$(1 - \epsilon)^{k-1} < q^2 (\frac{\delta}{K})^{2m} A^{-R}, \tag{77}$$

where A and ϵ are as in (67) and (68), respectively, and δ and K are introduced in §5.1.3.

5.1.6 The s-path

We now fix the state-sequence

$$b_0, b_1, \ldots, b_R, s_1, s_2, \ldots, s_{2Lk}, a_1, \ldots, a_P, \tag{78}$$

where $s_{Lj+i} = s_i,\ j = 1, \ldots, 2k-1,\ i = 1, \ldots, L,$ (and in particular $s_{Lj} = 1,\ j = 1, \ldots, 2k$). The sequence (78) will be called the *s-path*. The s-path is a concatenation of $2k$ **s** cycles $s_1, \ldots, s_L$, the beginning and the end of which are connected to the cluster C via positive probability paths **b** and **a**, respectively (recall that $a_P = q_1 \in C$ and $b_R = 1$ by construction). Additionally, the $b_R, s_1, s_2, \ldots, s_{2Lk}, a_1, \ldots, a_P$-segment of the s-path (78) has the important property (74), i.e. every consecutive transition along this segment occurs with the maximal transition probability given its destination state. (However, **b**, the beginning of the s-path, need not satisfy this property.) The s-path comes very close to being the sequence $q_{1\ldots M}$ required by the Lemma and will be completed to $q_{1\ldots M}$ in §5.1.18. In fact, the idea of the Lemma and its proof is to exhibit (a cylinder subset of) observations such that once emitted along the s-path, these observations would trap the Viterbi backtracking so that the latter winds up on the s-path. That will guarantee that an observation corresponding to the beginning of the s-path, is, as desired, a node.

5.1.7 The barrier

Consider the following sequence of observations

$$z_0, z_1, \ldots, z_m, x'_1, \ldots, x'_{R-1}, x_0, x_1, \ldots, x_{2Lk}, x''_1, \ldots, x''_P, z'_1, \ldots, z'_m, \tag{79}$$

where $z_0, z_i, z'_i \in \mathcal{Z},\ i = 1, \ldots, m$; $x'_i \in \mathcal{X}_{b_i},\ i = 1, \ldots, R-1$; and

$$x_0 \in \mathcal{X}_1, \quad x_{i+Lj} \in \mathcal{X}_{s_i},\ j = 1, \ldots, 2k-1,\ i = 1, \ldots, L; \quad x''_i \in \mathcal{X}_{a_i}, i = 1, \ldots, P.$$

From this point on throughout §5.1.16, we shall be proving that x_{Lk} is a 1-node of order $(kL + m + P)$, and, therefore, that (79) is a 1-barrier of order $(kL + m + P)$.

Let $u \geq 2Lk + 2m + 1 + P + R$ and let $x_1, \ldots, x_u$ be any sequence of observations terminating in the $2Lk + 2m + 1 + P + R$ observation long sequence of (79).

5.1.8 α, β, γ, η

Recall the definition of the scores $\delta_u(i)$ (15) and the maximum partial likelihoods $p^{(r)}_{ij}(u)$ (36). Now, we need to abbreviate some of the notation as follows. Namely, we denote by $\delta_i(x_l)$ (resp. $\delta_i(z_l)$) the scores corresponding to the observation x_l (resp. z_l). Similarly, we denote by $p^{(r)}_{ij}(x_l)$ (resp. $p^{(r)}_{ij}(z_l)$) the maximum partial likelihoods corresponding to the observation x_l (resp. z_l). Formally, for any $i, j \in S$ and appropriate $r \geq 0$, the abbreviated notation is as follows:

$$\begin{aligned}
&\delta_i(x_l) := \delta_{u-P-m-2kL+l}(i),\ p^{(r)}_{ij}(x_l) := p^{(r)}_{ij}(u - P - m - 2kL + l),\ 0 \leq l \leq 2kL; \qquad (80)\\
&p^{(r)}_{ij}(x'_l) := p^{(r)}_{ij}(u - P - m - 2kL - R + l),\ 1 \leq l \leq R-1;\\
&\delta_i(z_l) := \delta_{u-2Lk-2m-P-R+l}(i),\ p^{(r)}_{ij}(z_l) := p^{(r)}_{ij}(u - 2Lk - 2m - P - R + l), 0 \leq l \leq m;\\
&\delta_i(z'_l) := \delta_{u-m+l}(i),\ p^{(r)}_{ij}(z'_l) := p^{(r)}_{ij}(u - m + l),\ 1 \leq l \leq m.
\end{aligned} \tag{81}$$

Also, we will be frequently using the scores corresponding to z_0, x'_1, x_{Lk}, and x_{2Lk}, hence the following further abbreviations:

$$\alpha_i := \delta_i(z_0), \quad \beta_i := \delta_i(z_m), \quad \gamma_i := \delta_i(x_0), \quad \eta_i := \delta_i(x_{Lk}).$$

Note that $\forall j \notin C$, $f(z_0) = f_j(z'_l) = f_j(z_l) = 0$, $l = 1, \ldots, m$ by construction of $\mathcal{Z}$ (§5.1.3). Hence, $\alpha_j = \beta_j = 0$ $\forall j \notin C$, and a more general implication is that for every $j \in S$

$$\beta_j = \max_{i \in C} \alpha_i p_{ij}^{(m-1)}(z_0) f_j(z_m) = \alpha_{i_\beta(j)} p_{i_\beta(j)\, j}^{(m-1)}(z_0) f_j(z_m) \text{ for some } i_\beta(j) \in C; \tag{82}$$

$$\gamma_j = \max_{i \in C} \beta_i p_{ij}^{(R-1)}(z_m) f_j(x_0) = \beta_{i_\gamma(j)} p_{i_\gamma(j)\, j}^{(R-1)}(z_m) f_j(x_0) \text{ for some } i_\gamma(j) \in C. \tag{83}$$

Also note the following representation of η_j in terms of γ that we will use:

$$\eta_j = \max_{i \in S} \gamma_i p_{i\, j}^{(kL-1)}(x_0) f_j(x_{kL}) = \gamma_{i_\eta(j)} p_{i_\eta(j)\, j}^{(kL-1)}(x_0) f_j(x_{kL}) \text{ for some } i_\eta(j) \in S. \tag{84}$$

5.1.9 Bounds on β

Recall (§5.1.4) that $b_0 \in C$. We show that for every $j \in S$

$$\beta_j < q^{-1} \Big(\frac{K}{\delta} \Big)^m \beta_{b_0}. \tag{85}$$

Fix $j \in S$ and consider $\alpha_{i_\beta(j)}$ from (82). Let $v_1, \ldots, v_{m-1}$ be a path that realizes $p_{ij}^{(m-1)}(z_0)$. Then

$$\beta_j = \alpha_{i_\beta(j)} p_{i_\beta(j)\, v_1} f_{v_1}(z_1) p_{v_1\, v_2} f_{v_2}(z_2) \cdots p_{v_{m-1}\, j} f_j(z_m) < \alpha_{i_\beta(j)} K^m.$$

(The last inequality follows from the definition of $\mathcal{Z}$, §5.1.3.) Let $w_1, \ldots, w_{m-1}$ be a maximum probability path from $i_{\beta(j)}$ to b_0 as in (75). Thus,

$$\beta_{b_0} \geq \alpha_{i_\beta(j)} p_{i_\beta(j)\, b_0}^{(m-1)}(z_0) f_{b_0}(z_m) \geq \alpha_{i_\beta(j)} p_{i_\beta(j)\, w_1} f_{w_1}(z_1) p_{w_1\, w_2} f_{w_2}(z_2) \cdots p_{w_{m-1}\, b_0} f_{b_0}(z_m) \geq \alpha_{i_\beta(j)} q \delta^m.$$

(The last inequality again follows from the definition of $\mathcal{Z}$, §5.1.3.) Since $q > 0$ (76), we thus obtain:

$$\beta_j < \alpha_{i_\beta(j)} K^m \leq \frac{\beta_{b_0}}{q \delta^m} K^m,$$

as required.

5.1.10 Likelihood ratio bounds

We prove the following claims

$$p_{i1}^{(L-1)}(x_{lL}) \leq p_{11}^{(L-1)}(x_{lL}), \quad \forall i \in S, \quad \forall l = 0, \ldots, 2k-1; \tag{86}$$

$$\frac{p_{ij}^{(L-1)}(x_{lL}) f_j(x_{(l+1)L})}{p_{11}^{(L-1)}(x_{lL}) f_1(x_{(l+1)L})} < 1 - \epsilon, \quad \forall i, j \in S,\, j \neq 1, \quad \forall l = 0, \ldots, 2k-1; \tag{87}$$

$$p_{ij}^{(R-1)}(z_m) f_j(x_0) \leq A^R p_{b_0 1}^{(R-1)}(z_m) f_1(x_0), \quad \forall i, j \in S; \tag{88}$$

$$\frac{p_{ij}^{(m+P-1)}(x_{2kL})}{p_{1j}^{(m+P-1)}(x_{2kL})} \le q^{-1}\Big(\frac{K}{\delta}\Big)^{m-1}, \quad \forall j \in C, \forall i \in S. \tag{89}$$

If $L = 1$, then (86) becomes $p_{i\,1} \le p_{1\,1}$ for all $i \in S$, which is true by the assumption $p_1^* = p_{1\,1}$ made in the course of constructing the $\mathbf{s}$ sequence (§5.1.4). If $L = 1$, then (87) becomes

$$\frac{p_{ij} f_j(x_{l+1})}{p_{11} f_1(x_{l+1})} < 1 - \epsilon, \quad \forall i, j \in S, j \ne 1,$$

and thus, since $x_{l+1} \in \mathcal{X}_1$, $0 \le l < 2k$ in this case, (87) is true by the definition of $\mathcal{X}_1$ (§5.1.2) (and the fact that $p_1^* = p_{1\,1}$). Let us next prove (86) and (87) for the case $L > 1$. Consider any $l = 0, 1, \ldots, 2k-1$. Note that the definitions of the s-path (78), $\mathcal{X}_{s_i}$ (§5.1.2), and the fact that $x_{lL+i} \in \mathcal{X}_{s_i}$ for $1 \le i < L$ imply that given observations $x_{Ll+1}, \ldots, x_{L(l+1)-1}$, the path $s_1, \ldots, s_{L-1}$ realizes the maximum in $p_{11}^{(L-1)}(x_{Ll})$, i.e.

$$p_{11}^{(L-1)}(x_{lL}) = p_{1\,s_1} f_{s_1}(x_{lL+1}) p_{s_1\,s_2} \cdots p_{s_{L-2}\,s_{L-1}} f_{s_{L-1}}(x_{(l+1)L-1}) p_{s_{L-1}\,1}. \tag{90}$$

(Indeed,

$$p_{1\,s_1} f_{s_1}(x_{lL+1}) p_{s_1\,s_2} \cdots p_{s_{L-2}\,s_{L-1}} f_{s_{L-1}}(x_{(l+1)L-1}) p_{s_{L-1}\,1} = p_{s_1}^* f_{s_1}(x_{lL+1}) p_{s_2}^* \cdots p_{s_{L-1}}^* f_{s_{L-1}}(x_{(l+1)L-1}) p_1^*,$$

and for $i = 1, 2, \ldots, L-1$, $p_{s_i}^* f_{s_i}(x_{lL+i}) \ge p_{hj} f_j(x_{lL+i})$ for any $h, j \in S$.) Suppose $j \ne 1$ and $t_1, \ldots, t_{L-1}$ realizes $p_{ij}^{(L-1)}(x_{lL})$, i.e.

$$p_{ij}^{(L-1)}(x_{lL}) = p_{i\,t_1} f_{t_1}(x_{lL+1}) p_{t_1\,t_2} \cdots p_{t_{L-2}\,t_{L-1}} f_{t_{L-1}}(x_{(l+1)L-1}) p_{t_{L-1}\,j}. \tag{91}$$

Hence, with t_0 and t_L standing for i and j, respectively (and $s_0 = s_L = 1$), the left-hand side of (87) becomes

$$\Big(\frac{p_{t_0\,t_1} f_{t_1}(x_{lL+1})}{p_{s_0\,s_1} f_{s_1}(x_{lL+1})}\Big)\Big(\frac{p_{t_1\,t_2} f_{t_2}(x_{lL+2})}{p_{s_1\,s_2} f_{s_2}(x_{lL+2})}\Big) \cdots \Big(\frac{p_{t_{L-2}\,t_{L-1}} f_{t_{L-1}}(x_{(l+1)L-1})}{p_{s_{L-2}\,s_{L-1}} f_{s_{L-1}}(x_{(l+1)L-1})}\Big)\Big(\frac{p_{t_{L-1}\,t_L} f_j(x_{(l+1)L})}{p_{s_{L-1}\,s_L} f_1(x_{(l+1)L})}\Big).$$

For $h = 1, \ldots, L$ such that $t_h \ne s_h$,

$$\frac{p_{t_{h-1}\,t_h} f_{t_h}(x_{lL+h})}{p_{s_{h-1}\,s_h} f_{s_h}(x_{lL+h})} < 1 - \epsilon, \quad \text{since } x_{lL+h} \in \mathcal{X}_{s_h}. \tag{92}$$

For all other h, $s_h = t_h$ and therefore, the left-hand side of (92) becomes $\frac{p_{t_{h-1}\,t_h}}{p_{s_{h-1}\,s_h}} = \frac{p_{t_{h-1}\,s_h}}{p_{s_h}^*} \le 1$ (by property (74)). Since the last term of the product above does satisfy (92) ($j \ne 1$), (87) is thus proved. Suppose next that $t_1, \ldots, t_{L-1}$ realizes $p_{i1}^{(L-1)}(x_{lL})$. With $s_0 = 1$ and $t_0 = i$, similarly to the previous arguments, we have

$$\frac{p_{i\,1}^{(L-1)}(x_{lL})}{p_{1\,1}^{(L-1)}(x_{lL})} = \prod_{h=1}^{L-1}\Big(\frac{p_{t_{h-1}\,t_h} f_{t_h}(x_{lL+h})}{p_{s_{h-1}\,s_h} f_{s_h}(x_{lL+h})}\Big)\frac{p_{t_{L-1}\,1}}{p_{s_{L-1}\,1}} \le 1,$$

implying (86).

Let us now prove (88). To that end, note that for all states $h, i, j \in S$ such that $p_{jh} > 0$, it follows from the definitions (67) that

$$\frac{p_{ih}}{p_{jh}} \leq \frac{p^*_h}{p_{jh}} \leq A. \tag{93}$$

If $R = 1$, then (88) becomes

$$p_{ij} f_j(x_0) \leq A p_{b_0 1} f_1(x_0).$$

By the definition of $\mathcal{X}_1$ (recall that $x_0 \in \mathcal{X}_1$), we have that for every $i, j \in S$ $p_{ij} f_j(x_0) \leq p^*_1 f_1(x_0)$. Using (93) with $h = 1$ and $j = b_0$, we get $p^*_1 f_1(x_0) \leq A p_{b_0 1} f_1(x_0)$ ($p_{b_0 1} > 0$ by the construction of **b** §5.1.4). Putting these all together, we obtain

$$p_{ij} f_j(x_0) < p^*_1 f_1(x_0) \leq A p_{b_0 1} f_1(x_0), \text{ as required.}$$

Consider the case $R > 1$. Let $t_1, \ldots, t_{R-1}$ be a path that realizes $p^{(R-1)}_{ij}(z_m)$, i.e.

$$p^{(R-1)}_{ij}(z_m) = p_{i\,t_1} f_{t_1}(x'_1) p_{t_1\,t_2} f_{t_2}(x'_2) \cdots p_{t_{R-2}\,t_{R-1}} f_{t_{R-1}}(x'_{R-1}) p_{t_{R-1} j}.$$

By the definition of $\mathcal{X}_l$ (§5.1.2) and the facts that $x'_r \in \mathcal{X}_{b_r}$, $r = 1, 2, \ldots, R-1$, and $x_0 \in \mathcal{X}_1$, we have

$$p^{(R-1)}_{ij}(z_m) f_j(x_0) \leq p^*_{b_1} f_{b_1}(x'_1) p^*_{b_2} f_{b_2}(x'_2) \cdots p^*_{b_{R-1}} f_{b_{R-1}}(x'_{R-1}) p^*_1 f_1(x_0). \tag{94}$$

Now, by the construction of **b** (§5.1.4), $p_{b_{r-1}\,b_r} > 0$ for $r = 1, \ldots, R$, ($b_R = 1$). Thus, the argument behind (93) applies here to bound the right-hand side of (94) from above by

$$A p_{b_0\,b_1} f_{b_1}(x'_1) A p_{b_1\,b_2} f_{b_2}(x'_2) \cdots A p_{b_{R-2}\,b_{R-1}} f_{b_{R-1}}(x'_{R-1}) A p_{b_{R-1}\,1} f_1(x_0) = A^R p^{(R-1)}_{b_0 1}(z_m) f_1(x_0),$$

as required.

Let us now prove (89). If $m = 1$ then (89) becomes

$$p^{(P)}_{ij}(x_{2kL}) \leq p^{(P)}_{1j}(x_{2kL}) q^{-1}, \quad \forall j \in C, \forall i \in S. \tag{95}$$

If $P = 0$, then (95) reduces to $p_{ij} \leq p_{1j} q^{-1}$ which is true, because in this case the state $q_1 = q_T = 1$ belongs to C (§5.1.4) and $p_{1j} q^{-1} \geq 1$ ((75), (76) with $m = 1$). To see why (95) is true with $P \geq 1$, note that by the same argument as used to prove (86) and (87), we now get

$$p^{(P-1)}_{1\,a_P}(x_{2kL}) f_{a_P}(x''_P) \geq p^{(P-1)}_{h',l}(x_{2kL}) f_l(x''_P), \quad \forall h, l \in S. \tag{96}$$

Also, since $a_P = q_1 \in C$ (§5.1.4), $p_{a_P\,j} q^{-1} \geq 1$ ((75), (76) with $m = 1$). Thus

$$p^{(P)}_{i\,j}(x_{2kL}) \overset{\text{by (37)}}{=} \max_{l \in S} p^{(P-1)}_{i\,l}(x_{2kL}) f_l(x''_P) p_{l\,j} \overset{\text{by (96)}}{\leq} p^{(P-1)}_{1 a_P}(x_{2kL}) f_{a_P}(x''_P) \max_{l \in S} p_{l\,j} \leq$$

$$\leq \quad p^{(P-1)}_{1\,a_P}(x_{2kL}) f_{a_P}(x''_P) \leq p^{(P-1)}_{1\,a_P}(x_{2kL}) f_{a_P}(x''_P) p_{a_P\,j} q^{-1} \overset{\text{by (37)}}{\leq} p^{(P)}_{1\,j}(x_{2kL}) q^{-1}.$$

For $m > 1$, let $t_1, t_2, \ldots, t_{m-1}$ be a path realizing $p^{(m-1)}_{h\,j}(x''_P)$. Thus,

$$p^{(m-1)}_{h\,j}(x''_P) = p_{h\,t_1} f_{t_1}(z'_1) p_{t_1\,t_2} f_{t_2}(z'_2) \cdots f_{t_{m-1}}(z'_{m-1}) p_{t_{m-1}j} < K^{m-1}. \tag{97}$$

(This is true since $z'_r \in \mathcal{Z}$ for $r = 1, 2, \ldots, m-1$ (§5.1.3) and thus, for $p^{(m-1)}_{h\,j}(x''_P)$ to be positive it is necessary that $t_r \in C$, $r = 1, \ldots, m-1$, implying $f_{t_r}(z'_r) < K$.) Now, let $t_1, t_2, \ldots, t_{m-1}$ realize $p^{(m-1)}_{a_P\,j}(x''_P)$, which is clearly positive, with $t_r \in C$, $r = 1, \ldots, m-1$ ($z'_r \in \mathcal{Z}$ for $r = 1, 2, \ldots, m-1$), and $a_P, j \in C$ (recall the positivity assumption on Q^m, §5.1.5). We thus have $p^{(m-1)}_{a_P\,j}(x''_P) = p_{a_P\,t_1} f_{t_1}(z'_1) p_{t_1\,t_2} f_{t_2}(z'_2) \cdots f_{t_{m-1}}(z'_{m-1}) p_{t_{m-1}j} \geq$

$$\geq q^*_{a_P\,j} f_{t_1}(z'_1) f_{t_2}(z'_2) \cdots f_{t_{m-1}}(z'_{m-1}) > q\delta^{m-1}. \tag{98}$$

Combining the bounds of (97) and (98) ($q > 0$, (76)), we obtain :

$$p^{(m-1)}_{h\,j}(x''_P) < p^{(m-1)}_{a_P\,j}(x''_P) \Big(\frac{K}{\delta}\Big)^{m-1} q^{-1}. \tag{99}$$

Finally,

$$\begin{aligned} p^{(P+m-1)}_{ij}(x_{2kL}) &\overset{\text{by (37)}}{=} \max_{l \in S} p^{(P-1)}_{il}(x_{2kL}) f_l(x''_P) p^{(m-1)}_{lj}(x''_P) \overset{\text{by (96), (99)}}{<} \\ &\overset{\text{by (96), (99)}}{<} p^{(P-1)}_{1\,a_P}(x_{2kL}) f_{a_P}(x''_P) p^{(m-1)}_{a_P\,j}(x''_P) \Big(\frac{K}{\delta}\Big)^{m-1} q^{-1} \overset{\text{by (37)}}{\leq} \\ &\overset{\text{by (37)}}{\leq} p^{(P+m-1)}_{1j}(x_{2kL}) \Big(\frac{K}{\delta}\Big)^{m-1} q^{-1}. \end{aligned}$$

5.1.11 $\gamma_j \leq const \times \gamma_1$

Combining (83), (85), and (88), we get that for every state $j \in S$,

$$\begin{aligned} \gamma_j &\overset{\text{by (83)}}{=} \beta_{i_\gamma(j)} p^{(R-1)}_{i_\gamma(j)\,j}(z_m) f_j(x_0) \overset{\text{by (88)}}{\leq} \beta_{i_\gamma(j)} p^{(R-1)}_{b_0\,1}(z_m) f_1(x_0) A^R \overset{\text{by (85)}}{\leq} \\ &\overset{\text{by (85)}}{\leq} q^{-1} \Big(\frac{K}{\delta}\Big)^m A^R \beta_{b_0} p^{(R-1)}_{b_0\,1}(z_m) f_1(x_0) \leq U \max_{i \in S} \beta_i p^{(R-1)}_{i\,1}(z_m) f_1(x_0) \overset{\text{by (83)}}{=} U\gamma_1, \end{aligned}$$

where

$$U \overset{\text{def}}{=} q^{-1} \Big(\frac{K}{\delta}\Big)^m A^R. \tag{100}$$

Hence

$$\gamma_j \leq U\gamma_1, \quad \forall j \in S. \tag{101}$$

5.1.12 Further bounds on likelihoods

Let $l \geq 0$ and $n > 0$ be integers such that $l + n \leq 2k$ but arbitrary otherwise. Expanding $p_{1\,1}^{(nL-1)}(x_{lL})$ recursively according with (37), we obtain

$$p_{1\,1}^{(nL-1)}(x_{lL}) = \max_{i_1,i_2,\ldots,i_{n-1}\in S} p_{1\,i_1}^{(L-1)}(x_{lL}) f_{i_1}(x_{(l+1)L}) p_{i_1\,i_2}^{(L-1)}(x_{(l+1)L}) f_{i_2}(x_{(l+2)L}) \cdots \\ \cdots p_{i_{n-2}\,i_{n-1}}^{(L-1)}(x_{(l+n-2)L}) f_{i_{n-1}}(x_{(l+n-1)L}) p_{i_{n-1}\,1}^{(L-1)}(x_{(l+n-1)L}). \tag{102}$$

Since $p_{1\,i_1}^{(L-1)}(x_{lL}) f_{i_1}(x_{(l+1)L}) \leq p_{1\,1}^{(L-1)}(x_{lL}) f_1(x_{(l+1)L})$ for any $i_1 \in S$, as well as

$$p_{i_{r-1}\,i_r}^{(L-1)}(x_{(l+r-1)L}) f_{i_r}(x_{(l+r)L}) \leq p_{1\,1}^{(L-1)}(x_{(l+r-1)L}) f_1(x_{(l+r)L}) \; r = 2, \ldots, n-1, \text{ by } (87),$$

and since $p_{i_{n-1}\,1}^{(L-1)}(x_{(l+n-1)L}) \leq p_{1\,1}^{(L-1)}(x_{(l+n-1)L})$ for any $i_{n-1} \in S$ by (86), maximization (102) above is achieved as in (103) below:

$$p_{1\,1}^{(nL-1)}(x_{lL}) = p_{1\,1}^{(L-1)}(x_{lL}) f_1(x_{(l+1)L}) p_{1\,1}^{(L-1)}(x_{(l+1)L}) f_1(x_{(l+2)L}) \cdots \\ \cdots p_{1\,1}^{(L-1)}(x_{(l+n-2)L}) f_1(x_{(l+n-1)L}) p_{1\,1}^{(L-1)}(x_{(l+n-1)L}). \tag{103}$$

Now, we replace state 1 by generic states $i, j \in S$ on the both ends of the paths in (102) and repeat the above arguments. Thus, also using (103), we arrive at bound (104) below:

$$p_{ij}^{(nL-1)}(x_{lL}) f_j(x_{(l+n)L}) \leq \prod_{u=l+1}^{l+n} p_{1\,1}^{(L-1)}(x_{(u-1)L}) f_1(x_{uL}) \overset{\text{by (103)}}{=} \\ \overset{\text{by (103)}}{=} p_{1\,1}^{(nL-1)}(x_{lL}) f_1(x_{(l+n)L}), \; \forall i, j \in S. \tag{104}$$

In particular, (104) states

$$p_{ij}^{(kL-1)}(x_0) f_j(x_{kL}) \leq p_{1\,1}^{(kL-1)}(x_0) f_1(x_{kL}), \quad \forall i, j \in S. \tag{105}$$

5.1.13 $\eta_j \leq const \times \eta_1$

In order to see

$$\eta_j \leq U \eta_1, \quad \forall j \in S, \tag{106}$$

note that:

$$\eta_j \overset{(84)}{=} \max_{i\in S} \gamma_i p_{i\,j}^{(kL-1)}(x_0) f_j(x_{kL}) \overset{\text{by (105)}}{\leq} \max_{i\in S} \gamma_i p_{1\,1}^{(kL-1)}(x_0) f_1(x_{kL}) \overset{\text{by (101)}}{\leq} \\ \overset{\text{by (101)}}{\leq} U \gamma_1 p_{1\,1}^{(kL-1)}(x_0) f_1(x_{kL}) \overset{\text{by (84)}}{\leq} U \eta_1.$$

5.1.14 A representation of η_1

Recall that k, the number of cycles in the s-path, was chosen sufficiently large for (77) to hold (in particular, $k > 1$). We now prove that there exists $\kappa \in \{1, \ldots, k-1\}$ such that

$$\eta_1 = \delta_1(x_{\kappa L}) p_{11}^{((k-\kappa)L-1)}(x_{\kappa L}) f_1(x_{kL}). \tag{107}$$

The relation (107) states that (given observations $x_1, x_2, \ldots, x_u$) a maximum-likelihood path (from time 1, observation x_1) to time $u-m-P-kL$ (observation x_{kL}) goes through state 1 at time $u - m - P - 2kL + \kappa L$, that is when $x_{\kappa L}$ is observed.

To see this, suppose no such κ exists to satisfy (107). Then, applying (37) to (84) and recalling that $\delta_1(x_{\kappa L})$ is introduced by (80), we would have

$$\eta_1 = \gamma_{j_\eta(1)} p_{j_\eta(1)\, j_1}^{(L-1)}(x_0) f_{j_1}(x_L) p_{j_1\, j_2}^{(L-1)}(x_L) f_{j_2}(x_{2L}) p_{j_2\, j_3}^{(L-1)}(x_{2L}) \cdots p_{j_{k-1}\, 1}^{(L-1)}(x_{(k-1)L}) f_1(x_{kL})$$

for some $j_1 \neq 1, \ldots, j_{k-1} \neq 1$. Furthermore, this would imply

$$\begin{aligned}
\eta_1 &\overset{\text{by (87), (86)}}{<} \gamma_{j_\eta(1)} (1-\epsilon)^{k-1} \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL}) \overset{\text{by (77)}}{<} \\
&\overset{\text{by (77)}}{<} \gamma_{j_\eta(1)} q^2 \left(\frac{\delta}{K}\right)^{2m} A^{-R} \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL}) \overset{\text{by (101)}}{\leq} \\
&\overset{\text{by (101)}}{\leq} \gamma_1 U q^2 \left(\frac{\delta}{K}\right)^{2m} A^{-R} \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL}) \overset{\text{by (100)}}{=} \\
&\overset{\text{by (100)}}{=} \gamma_1 q \left(\frac{\delta}{K}\right)^{m} \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL}) < \gamma_1 \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL}).
\end{aligned} \tag{108}$$

(The last inequality follows from $q \leq 1$ (76) and $\delta < K$, §5.1.3.) On the other hand, by definition (84) (and $k-1$-fold application of (37)), $\eta_1 \geq \gamma_1 \prod_{i=1}^{k} p_{11}^{(L-1)}(x_{(i-1)L}) f_1(x_{iL})$, which evidently contradicts (108) above. Therefore, κ satisfying (107) and $1 \leq \kappa < k$, does exist.

5.1.15 An implication of (103) and (107) for $\delta_1(x_{lL})$

Clearly, the arguments of the previous section (§5.1.14) are valid if k is replaced by any $l \in \{k, \ldots, 2k\}$. Hence the following generalization of (107):

$$\delta_1(x_{lL}) = \delta_1(x_{\kappa(l)L}) p_{11}^{((l-\kappa(l))L-1)}(x_{\kappa(l)L}) f_1(x_{lL}) \text{ for some } \kappa(l) < l. \tag{109}$$

We apply (109) recursively, starting with $\kappa^{(0)} := l$ and returning $\kappa^{(1)} := \kappa(l) < l$. If $\kappa^{(1)} \leq k$, we stop, otherwise we substitute $\kappa^{(1)}$ for l, and obtain $\kappa^{(2)} := \kappa(l) < \kappa^{(1)}$, and so, on until $\kappa^{(j)} \leq k$ for some $j > 0$. Thus,

$$\delta_1(x_{lL}) = \delta_1(x_{\kappa^{(j)}L}) p_{11}^{((\kappa^{(j-1)}-\kappa^{(j)})L-1)}(x_{\kappa^{(j)}L}) f_1(x_{\kappa^{(j-1)}L}) \cdots p_{11}^{((l-\kappa^{(1)})L-1)}(x_{\kappa^{(1)}L}) f_1(x_{lL}). \tag{110}$$

Applying (103) to the appropriate factors of the right-hand side of (110) above, we get:

$$\begin{aligned}\delta_1(x_{lL}) =&\delta_1(x_{\kappa^{(j)}L})p_{11}^{(L-1)}(x_{\kappa^{(j)}L})f_1(x_{(\kappa^{(j)}+1)L})\cdots p_{11}^{(L-1)}(x_{(k-1)L})f_1(x_{kL})\cdots \\ &\cdots p_{11}^{(L-1)}(x_{kL})f_1(x_{(k+1)L})\cdots p_{11}^{(L-1)}(x_{(\kappa^{(j-1)}-1)L})f_1(x_{\kappa^{(j-1)}L})\cdots \\ &\cdots p_{11}^{(L-1)}(x_{(\kappa^{(1)}-1)L})f_1(x_{\kappa^{(1)}L})\cdots p_{11}^{(L-1)}(x_{(l-1)L})f_1(x_{lL}).\end{aligned} \tag{111}$$

Also, according to (103),

$$\delta_1(x_{\kappa^{(j)}L})p_{11}^{(L-1)}(x_{\kappa^{(j)}L})f_1(x_{(\kappa^{(j)}+1)L})\cdots p_{11}^{(L-1)}(x_{(k-1)L}) = \delta_1(x_{\kappa^{(j)}L})p_{11}^{((k-\kappa^{(j)})L-1)}(x_{\kappa^{(j)}L}).$$

At the same time,

$$\delta_1(x_{\kappa^{(j)}L})p_{11}^{((k-\kappa^{(j)})L-1)}(x_{\kappa^{(j)}L})f_1(x_{kL}) \overset{\text{by (38)}}{\leq} \eta_1. \tag{112}$$

However, we cannot have the strict inequality in (112) above since that, via (111), would contradict maximality of $\delta_1(x_{lL})$. We have thus arrived at

$$\delta_1(x_{lL}) = \eta_1 p_{11}^{(L-1)}(x_{kL})f_1(x_{(k+1)L})\cdots p_{11}^{(L-1)}(x_{(l-1)L})f_1(x_{lL}). \tag{113}$$

In summary, for any $l \geq k$ and $l \leq 2k$ there exists a realization of $\delta_1(x_{lL})$ that goes through state 1 every time when x_{iL}, $i = k, \ldots, l$, is observed.

5.1.16 x_{kL} is a $(kL+m+P)$-order 1-node

When we prove in §5.1.17 that for any $i \in S, i \neq 1$, and any$j \in C$,

$$\eta_i p_{ij}^{(kL+m+P-1)}(x_{kL}) \leq \eta_1 p_{1j}^{(kL+m+P-1)}(x_{kL}), \tag{114}$$

this will immediately imply that x_{kL} is a 1-node of order $kL+m+P$. Indeed, let $l \in S$ be arbitrary. Since $f_j(z'_m) = 0$ for every $j \in S \setminus C$, any maximum likelihood path to state l at time $u+1$ (observation x_{u+1}) must go through a state in C at time u (observation $x_u = z'_m$.) Formally,

$$\begin{aligned}\eta_i p_{il}^{(kL+m+P)}(x_{kL}) &= \max_{j\in S} \eta_i p_{ij}^{(kL+m+P-1)}(x_{kL})f_j(z'_m)p_{jl} = \max_{j\in C} \eta_i p_{ij}^{(kL+m+P-1)}(x_{kL})f_j(z'_m)p_{jl} \\ &\overset{\text{by (114)}}{\leq} \max_{j\in C} \eta_1 p_{1j}^{(kL+m+P-1)}(x_{kL})f_j(z'_m)p_{jl} \overset{\text{by (37)}}{=} \eta_1 p_{1l}^{(kL+m+P)}(x_{kL}).\end{aligned}$$

Therefore, by Definition 3.5 x_{kL} is a 1-node of order $kL+m+P$.

5.1.17 Proof of (114)

Let $i \in S$ and $j \in C$ be arbitrary. Let state $j^* \in S$ be such that

$$p_{i\,j}^{(kL+m+P-1)}(x_{kL}) = p_{i\,j^*}^{(kL-1)}(x_{kL})f_{j^*}(x_{2kL})p_{j^*\,j}^{(m+P-1)}(x_{2kL}) = \nu(i,j^*)p_{j^*\,j}^{(m+P-1)}(x_{2kL}),$$

where

$$\nu(i,j) \overset{\text{def}}{=} p_{ij}^{(kL-1)}(x_{kL})f_j(x_{2kL}), \quad \text{for all } i,j \in S.$$

We consider the following two cases separately:

1. There exists a path realizing $p_{i\,j^*}^{(kL-1)}(x_{kL})$ and going through state 1 at the time of observing x_{lL} for some $l \in \{k, \ldots, 2k\}$.

$$p_{i\,j^*}^{(kL-1)}(x_{kL}) = p_{i\,1}^{((l-k)L-1)}(x_{kL}) f_1(x_{lL}) p_{1\,j^*}^{((2k-l)L-1)}(x_{lL}). \tag{115}$$

Equation (115) above together with the fundamental recursion (38) yields the following:

$$\begin{aligned}\eta_i p_{i\,j^*}^{(kL-1)}(x_{kL}) &\overset{\text{by (115)}}{=} \eta_i p_{i\,1}^{((l-k)L-1)}(x_{kL}) f_1(x_{lL}) p_{1\,j^*}^{((2k-l)L-1)}(x_{lL}) \overset{\text{by (80), (38)}}{\leq} \\ &\overset{\text{by (80), (38)}}{\leq} \delta_1(x_{lL}) p_{1\,j}^{((2k-l)L-1)}(x_{lL}).\end{aligned} \tag{116}$$

At the same time, the right hand-side of (116) can be expressed as follows:

$$\begin{aligned}\delta_1(x_{lL}) p_{1\,j^*}^{((2k-l)L-1)}(x_{lL}) &\overset{\text{by (113)}}{=} \eta_1 p_{1\,1}^{((l-k)L-1)}(x_{kL}) f_1(x_{lL}) p_{1\,j^*}^{((2k-l)L-1)} \overset{\text{by (103)}}{=} \\ &\overset{\text{by (103)}}{=} \eta_1 p_{1\,j^*}^{(kL-1)}(x_{kL}).\end{aligned} \tag{117}$$

Therefore, if there exists $l \in \{k, \ldots, 2k\}$ such that (115) holds, we have by virtue of (116) and (117):

$$\eta_i p_{i\,j^*}^{(kL-1)}(x_{kL}) \leq \eta_1 p_{1\,j^*}^{(kL-1)}(x_{kL}), \quad \text{that is,} \quad \eta_i \nu(i, j^*) \leq \eta_1 \nu(1, j^*). \tag{118}$$

Hence,

$$\begin{aligned}\eta_i p_{i\,j}^{(kL+m+P-1)}(x_{kL}) &\overset{\text{by (115)}}{=} \eta_i \nu(i, j^*) p_{j^*\,l}^{(m+P-1)}(x_{2kL}) \overset{\text{by (118)}}{\leq} \\ &\overset{\text{by (118)}}{\leq} \eta_1 \nu(1, j^*) p_{j^*\,j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (37)}}{\leq} \eta_1 p_{1\,j}^{(kL+m+P-1)}(x_{kL})\end{aligned}$$

and (114) holds.

2. Assume now that no path exists to satisfy (115). Argue as for (108) to get

$$\nu(i, j^*) < (1-\epsilon)^{k-1} \prod_{n=k+1}^{2k} p_{1\,1}^{(L-1)}(x_{(n-1)L}) f_1(x_{nL}). \tag{119}$$

By 103, the (partial likelihood) product in the right-hand side of (119) equals $\nu(1, 1)$. Thus,

$$\begin{aligned}\eta_i \nu(i, j^*) p_{j^*\,j}^{(m+P-1)}(x_{2kL}) &\overset{\text{by (119)}}{<} \eta_i (1-\epsilon)^{k-1} \nu(1,1) p_{j^*\,j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (77)}}{<} \\ &\overset{\text{by (77)}}{<} \eta_i q^2 \left(\frac{\delta}{K}\right)^{2m} A^{-R} \nu(1,1) p_{j^*\,j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (100), (106)}}{\leq} \\ &\overset{\text{by (100), (106)}}{\leq} \eta_1 q \left(\frac{\delta}{K}\right)^{m} \nu(1,1) p_{j^*\,j}^{(m+P-1)}(x_{2kL}).\end{aligned} \tag{120}$$

Hence, for every $j' \in S$,

$$\eta_i \nu(i, j') p_{j'\,j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (115)}}{\leq} \eta_i \nu(i, j^*) p_{j^*\,j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (120)}}{<}$$

$$\overset{\text{by (120)}}{<} \eta_1 q \left(\frac{\delta}{K}\right)^m \nu(1,1) p_{j^* j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (89)}}{\leq}$$
$$\overset{\text{by (89)}}{\leq} \eta_1 \left(\frac{\delta}{K}\right) \nu(1,1) p_{1j}^{(m+P-1)}(x_{2kL}) < \eta_1 \nu(1,1) p_{1j}^{(m+P-1)}(x_{2kL}) \overset{\text{by (37)}}{\leq}$$
$$\overset{\text{by (37)}}{\leq} \eta_1 p_{1j}^{(kL+m+P-1)}(x_{kL}),$$

which, by virtue of (37), implies (114).

5.1.18 Completion of the s-path to $q_{1\ldots M}$ and conclusion

Finally, let

$$M = 2m + 2Lk + P + R + 2, \quad r = kL + P + m, \quad l = 1.$$

Recall from §5.1.4 that $b_0 \in C$. Since all the entries of Q^m are positive, there exists a path $v_0, v_1, \ldots, v_{m-1}, b_0 \in C$ such that $p_{v_i\, v_{i+1}} > 0$ and $p_{v_{m-1} b_0} > 0$. Similarly, there must exist a path $u_1, \ldots, u_m \in C$ such that $p_{u_i\, u_{i+1}} > 0\ \forall i = 1, \ldots, m-1$ and $p_{a_P\, u_1} > 0$ (recall that $a_P \in C$). Hence, by these and the constructions of §5.1.6, all of the transitions of the following sequence occur with positive probabilities.

$$q_{1\ldots M} \overset{\text{def}}{=} v_0, v_1, \ldots, v_{m-1}, b_0, b_1, \ldots, b_{R-1}, b_R, s_1, \ldots, s_{2Lk}, a_1, \ldots, a_P, u_1, \ldots, u_m. \quad (121)$$

Clearly, the actual probability of observing $q_{1\ldots M}$ is positive, as required. By the constructions of §§5.1.2-5.1.4, the conditional probability of B below, given $q_{1\ldots M}$, is evidently positive, as required.

$$B \overset{\text{def}}{=} \mathcal{Z}^{m+1} \times \mathcal{X}_{b_1} \times \cdots \times \mathcal{X}_{b_{R-1}} \times \mathcal{X}_1 \times \mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_{2kL-1}} \times \mathcal{X}_1 \times \mathcal{X}_{a_1} \times \cdots \times \mathcal{X}_{a_P} \times \mathcal{Z}^m.$$

Finally, since the sequence (79) below was chosen from B arbitrarily (§5.1.7) and has been shown to be an l-barrier of order r, this completes the proof of the Lemma.

$$z_0, z_1, \ldots, z_{m-1}, z_m, x_1', \ldots, x_{R-1}', x_0, x_1, \ldots, x_{2Lk}, x_1'', \ldots, x_P'', z_1', \ldots, z_m'. \quad (79)$$

■

5.2 Proof of Lemma 3.2

Proof. We use the notation of the previous proof in §5.1. We deal with the following two different situations: First (§5.2.1), all barriers from B as constructed in the proof of Lemma 3.2 are already separated. Obviously, there is nothing to do in this case. The second situation (§5.2.2) is complementary, in which case a simple extension will immediately ensure separation.

5.2.1 All $x^b \in B$ are already separated

Recall the definition of $\mathcal{Z}$ from §5.1.3. Consider the two cases in the definition separately. First, suppose $\mathcal{Z} = \hat{\mathcal{Z}} \backslash (\cup_{l \in S} \mathcal{X}_l)$, in which case $\mathcal{Z}$ and $\mathcal{X}_l$ are disjoint for every $l \in S$.

This implies that every barrier (79) is already separated. Indeed, for any w, $1 \leq w \leq r$, and for any $x^b \in B$, the fact that $x^b_{M-\max(m,w)} \notin \mathcal{Z}$, for example, makes it impossible for $(x'_{1\ldots w}, x^b_{1\ldots M-w}) \in B$ for any $x'_{1\ldots w} \in \mathcal{X}^w$. Consider now the case when $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$ for some $s \in C$. Then

$$B \subset \mathcal{X}_s^{m+1} \times \mathcal{X}_{b_1} \times \cdots \times \mathcal{X}_{b_{R-1}} \times \mathcal{X}_1 \times \mathcal{X}_{s_1} \times \cdots \times \mathcal{X}_{s_{2kL-1}} \times \mathcal{X}_1 \times \mathcal{X}_{a_1} \times \cdots \times \mathcal{X}_{a_{P-1}} \times \mathcal{X}_s^{m+1}.$$

Let $x^b \in B$ be arbitrary. Assume first $L > 1$. By construction (§5.1.4), the states $s_1, \ldots, s_L$ are all distinct. We now show that $(x'_{1\ldots w}, x^b_{1\ldots M-w}) \notin B$ for any $x'_{1\ldots w} \in \mathcal{X}^w$ when $1 \leq w \leq r$. Note that the sequence

$$q_{m+2\ldots m+R+2kL+P+1} = (b_1, \ldots, b_{R-1}, 1, s_1, \ldots, s_{2kL-1}, 1, a_1, \ldots, a_{P-1}, s)$$

is such that no two consecutive states are equal. It is straightforward to verify that there exist indices j, $0 \leq j \leq m-1$, such that, when shifted w positions to the right, the pair $x_{j+1\,j+2} \in \mathcal{X}_s^2$ would at the same time have to belong to $\mathcal{X}_{q_{j+1+w}} \times \mathcal{X}_{q_{j+2+w}}$ with $m+1 \leq j+1+w < j+2+w \leq m+R+2kL+1+P$. This is clearly a contradiction since $\mathcal{X}_{q_{j+1+w}}$ and $\mathcal{X}_{q_{j+2+w}}$ are disjoint for that range of indices j. A verification of the above fact simply amounts to verifying that the inequality $\max(0, m-w) \leq j \leq \min(m-1, m+R+2kL-1+P-w)$ is consistent for any w from the admissible range:

i.) When $0 \geq m-w$, $m-1 \leq m+R+2kL-1+P-w$ ($m \leq w \leq \min(r, R+2kL+P)$), $0 \leq j \leq m-1$ is evidently consistent.

ii.) When $0 \geq m-w$, $m-1 > m+R+2kL-1+P-w$ ($\max(m, R+2kL+P) \leq w \leq r$), $0 \leq j \leq m+R+2kL-1+P-w$ is also consistent since $m+R+2kL-1+P-r = R+kL-1 \geq 0$.

iii.) When $0 < m-w$, $m-1 \leq m+R+2kL-1+P-w$ ($1 \leq w \leq \min(m-1, R+2kL+P)$), $m-w \leq j \leq m-1$ is consistent since $w \geq 1$.

iv.) When $0 < m-w$, $m-1 > m+R+2kL-1+P-w$ ($\max(1, R+2kL+P-1) \leq w < m$), $m-w \leq j \leq m+R+2kL-1+P-w$ is consistent since $R+2kL-1 \geq 0$.

Next consider the case of $L = 1$ but $s \neq 1$ (that is, $P > 0$). Then

$$B \subset \mathcal{X}_s^{m+1} \times \mathcal{X}_{b_1} \times \cdots \times \mathcal{X}_{b_{R-1}} \times \mathcal{X}_1^{2k+1} \times \mathcal{X}_{a_1} \times \cdots \times \mathcal{X}_{a_{P-1}} \times \mathcal{X}_s^{m+1}.$$

If $s \neq 1$, then also $b_i \neq 1$, $i = 1, \ldots, R-1$ and $a_i \neq 1$, $i = 1, \ldots, P-1$. To see that y is separated in this case, simply note that $x_{M-max(w,m+1)} \notin \mathcal{X}_s$ for any admissible w.

5.2.2 Barriers $x^b \in B$ need not be separated

Finally, we consider the case when $L = 1$ and $s = 1$ (where $s \in C$ is such that $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$). This implies that $P = 0$, $1 \in C$, and $p_{1\,1} > 0$, which in turn implies that $R = 1$, and

$$B \subset \mathcal{X}_1^{m+1} \times \mathcal{X}_1^{2k+1} \times \mathcal{X}_1^{m+1} = \mathcal{X}_1^{2m+2k+3}.$$

Clearly, the barriers from B need not be, and indeed, are not separated. It is, however, easy to extend them to separated ones. Indeed, let $q_0 \neq 1$ be such that $p_{q_0\,1} > 0$ and

redefine $B \stackrel{\text{def}}{=} \mathcal{X}_{q_0} \times B$. Evidently, any shift of any $x^b \in B$ by w $(1 \le w \le r)$ positions to the right makes it impossible for x_1^b to be simultaneously in $\mathcal{X}_{q_0}$ and in $\mathcal{X}_1$ (since the latter sets are disjoint, §5.1.2). ■

5.3 Proof of Proposition 4.1

Proof. Let us additionally define the following non-overlapping block-valued processes

$$U_m^b \stackrel{\text{def}}{=} (X_{(m-1)M+1}, \ldots, X_{mM}), \quad D_m^b \stackrel{\text{def}}{=} (Y_{(m-1)M+1}, \ldots, Y_{mM}), \quad m = 1, 2, \ldots,$$

and stopping times

$$\begin{aligned} \nu_0^b &\stackrel{\text{def}}{=} \min\{m \ge 1 : U_m^b \in B, D_m^b = q\}, \\ \nu_i^b &\stackrel{\text{def}}{=} \min\{m > \nu_{i-1}^b : U_m^b \in B, D_m^b = q\}; \end{aligned} \tag{122}$$

$$\begin{aligned} R_0^b &\stackrel{\text{def}}{=} \min\{m > 1 : \ D_m^b = q\}, \\ R_i^b &\stackrel{\text{def}}{=} \min\{m > R_{i-1}^b : \ D_m^b = q\}. \end{aligned} \tag{123}$$

The process D^b is clearly a time homogeneous, finite state Markov chain. Since Y is aperiodic and irreducible, so is D^b. Hence (D^b, U^b) is also an HMM.
Since Y is stationary (under π), q occurs in every interval of length M with the same positive probability (Lemma 3.2). In particular, q belongs to the state space of D^b. Since D^b is irreducible and its state space is finite, all of its states, including q, are positive recurrent. Hence $E_{\pi'}(R_0^b) < \infty$ and $E_{\pi'}(R_1^b - R_0^b) < \infty$ and recall that $R_i^b - R_{i-1}^b$, $i = 1, 2, \ldots$ are i.i.d. (These and the statements below hold for any initial distribution π'.) The following two equalities are straightforward to verify and ultimately yield the second statement: $E_{\pi'}(\nu_1 - \nu_0) \le E_{\pi'}(\nu_1^b - \nu_0^b) = \frac{1}{\gamma^*} E_{\pi'}(R_1^b - R_0^b) < \infty$. The second equality above is also a simple extension of the Wald's equation (for a general reference see, for example, [7]).

It can similarly be verified that $E_{\pi'}\nu_0^b = \gamma^* E_{\pi'} R_0^b + \frac{1-\gamma^*}{\gamma^*} E_{\pi'}(R_1^b - R_0^b)$, which is again finite. Finally, $E_{\pi'}\nu_0 \le M(E_{\pi'}\nu_0^b - 1) + 1 < \infty$. ■

5.4 Proof of Proposition 4.2

Proof. Recall (56), the definition of stopping times τ, according to which, for each $i = 1, 2, \ldots$ the underlying Markov chain satisfies $Y_{\tau_i} = l$. Hence, the behavior of X after τ_i does not depend of the behavior of X up to τ_i. Together with the fact that T_i are renewal, this establishes regenerativity of X. Next, to every τ_i there corresponds a r-order l-node and τ_i is always u_j for some $j > i$. This means that all the nodes corresponding to τ_i's are also used to define the alignment as in Definition 4.1. Therefore, the alignment up to τ_i does not depend on the alignment after τ_i. In other words, the segment of the alignment process that corresponds to T_i is a function of the segment of X corresponding to the same T_i. Formally

$$(V_s : s \in \tau_{i-1} + 1, \ldots, \tau_i) = v_{(l)}(X_s : s \in \tau_{i-1} + 1, \ldots, \tau_i).$$

Thus, the process Z is regenerative with respect to τ. ■

5.5 Proof of Theorem 4.4

Proof. First note that the right-hand side of (65) does define a measure.
The proof below uses regenerativity of Z in a standard way. For every $n \geq \tau_0$ and $A \in \mathcal{B}$, and for every $l \in S$, we have

$$\frac{1}{n}\sum_{i=1}^{n} I_{A\times l}(Z_i) = \frac{1}{n}\sum_{i=1}^{\tau_o} I_{A\times l}(Z_i) + \frac{1}{n}\sum_{i=\tau_o+1}^{\tau_{k(n)}} I_{A\times l}(Z_i) + \frac{1}{n}\sum_{i=\tau_{k(n)}+1}^{n} I_{A\times l}(Z_i)$$

where $k(n) = \max\{k : \tau_k \leq n\}$ stands for the renewal process. Now, since $\tau_0 < \infty$ a.s., we have

$$\frac{1}{n}\sum_{i=1}^{\tau_0} I_{A\times l}(Z_i) \leq \frac{\tau_0}{n} \to 0, \quad \text{a.s..}$$

Let $\mu \stackrel{\text{def}}{=} E\tau_0^r$. By (64), $\mu < \infty$. Then

$$\frac{n - \tau_{k(n)}}{n} \leq \frac{T_{k(n)+1}}{n} \to 0, \quad \text{a.s.}$$

Finally, since Z is regenerative with respect to $\tau_0, \tau_1, \ldots$, we have

$$\frac{1}{n}\sum_{i=\tau_0+1}^{\tau_{k(n)}} I_{A\times l}(Z_i) = \frac{k(n)}{n}\frac{1}{k(n)}\sum_{k=1}^{k(n)} \xi_k,$$

where

$$\xi_k \stackrel{\text{def}}{=} \sum_{i=\tau_{k-1}+1}^{\tau_k} I_{A\times l}(Z_i), \quad k = 1, 2, \ldots$$

are i.i.d. random variables. Let $m_l(A)$ stand for $E\xi_k$. Thus, $m_l(A) \leq \mu < \infty$. Then, as $n \to \infty$, we have

$$\frac{n}{k(n)} \to \mu \quad \text{and} \quad \frac{1}{k(n)}\sum_{k=1}^{k(n)} \xi_k \to m_l(A), \quad \text{a.s.}$$

Let us calculate $m_l(A)$. Clearly,

$$m_l(A) = E\sum_{i=1}^{\tau_0^r} I_{A\times l}(Z_i^r).$$

Now

$$m_l(A) = E\sum_{i=1}^{\tau_0^r} I_{A\times l}(Z_i^r) = \sum_{j=1}^{\infty} E(\sum_{i=1}^{j} I_{A\times l}(Z_i^r)|\tau_0^r = j)\mathbf{P}(\tau_0^r = j)$$

$$= \sum_{j=1}^{\infty} \sum_{i=1}^{j} \mathbf{P}(Z_i^r \in A \times l | \tau_0^r = j) \mathbf{P}(\tau_0^r = j)$$
$$= \sum_{j=1}^{\infty} \mathbf{P}(Z_1^r \in A \times l | \tau_0^r = j) \mathbf{P}(\tau_0^r = j) + \sum_{j=2}^{\infty} \mathbf{P}(Z_2^r \in A \times l | \tau_0^r = j) \mathbf{P}(\tau_0^r = j) + \cdots$$
$$= \mathbf{P}(Z_1^r \in A \times l, \tau_0^r \geq 1) + \mathbf{P}(Z_2^r \in A \times l, \tau_0^r \geq 2) + \cdots$$
$$= \sum_{i=1}^{\infty} \mathbf{P}(Z_i^r \in A \times l, \tau_0^r \geq i) \leq \sum_{i=1}^{\infty} \mathbf{P}(\tau_0^r \geq i) = \mu < \infty$$

Similarly,

$$\frac{1}{n} \sum_{i=1}^{n} I_l(V_i^r) \to \frac{w_l}{\mu} \leq 1, \quad \text{a.s.}, \tag{124}$$

where $w_l \stackrel{\text{def}}{=} \sum_{i=1}^{\infty} P(V_i^r = l, \tau_0^r \geq i)$. Hence, we have shown that for each $l \in S$ and for every $A \in \mathcal{B}$

$$Q_l^n(A) \to \frac{m_l(A)}{w_l} = \frac{\sum_{i=1}^{\infty} \mathbf{P}(Z_i^r \in A \times l, \tau_0^r \geq i)}{\sum_{i=1}^{\infty} \mathbf{P}(V_i^r = l, \tau_0^r \geq i)}, \quad \text{a.s.} \tag{125}$$

Recalling that $\mathcal{X}$ is a separable metric space and envoking the theory of weak convergence of measures now establishes $Q_l^n \Rightarrow Q_l$, a.s..

It remains to show that for all $l \in S$ and $A \in \mathcal{B}$

$$P_l^n(A) \to \frac{m_l(A)}{w_l}, \quad \text{a.s.}. \tag{126}$$

To see this, consider $\sum_{i=1}^{n} I_{A \times l}(X_i, V_i')$. Since $V_i' = V_i$, if $i \leq \tau_{k(n)}$, we obtain

$$\frac{1}{n} \sum_{i=1}^{n} I_{A \times l}(X_i, V_i') = \frac{1}{n} \sum_{i=1}^{\tau_0} I_{A \times l}(Z_i) + \frac{1}{n} \sum_{i=\tau_0+1}^{\tau_{k(n)}} I_{A \times l}(Z_i) + \frac{1}{n} \sum_{i=\tau_{k(n)}+1}^{n} I_{A \times l}(X_i, V_i') \to \frac{m_l(A)}{\mu} \quad \text{a.s.}$$

Similarly,

$$\frac{1}{n} \sum_{i=1}^{n} I_l(V_i') \to \frac{1}{\mu} \sum_{i=1}^{\infty} P_1(V_i = l, \tau_0^r \geq i) = \frac{w_l}{\mu}, \quad \text{a.s.}. \tag{127}$$

These convergences prove (126). ■

Acknowledgement

The first author has been supported by the Estonian Science Foundation Grant 5694 at the later stages of the work. The authors are also thankful to *Eurandom* (The Netherlands) and Professors R. Gill and A. van der Vaart for their support.

References

[1] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37:1554–1563, 1966.

[2] J. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Technical Report 97–021, International Computer Science Institute, Berkeley, CA, USA, 1998.

[3] P. Brémaud. *Markov Chains: Gibbs fields, Monte Carlo simulation, and Queues*. Springer, 1999.

[4] P. Chou, T. Lookbaugh, and R. Gray. Entropy-Constrained Vector Quantization. *IEEE Trans. Acoust. Speech Signal Process.*, 37(1):31–42, 1989.

[5] G. Ehret, P. Reichenbach, U. Schindler, and *et. al.* DNA Binding Specificity of Different STAT Proteins. *J. Biol. Chem.*, 276(9):6675–6688, 2001.

[6] R. Gray, T. Linder, and J. Li. A Lagrangian formulation of Zador's entropy-constrained quantization theorem. *IEEE Trans. Inf. Theory*, 48(3):695–707, 2000.

[7] G. Grimmet and D. Stirzaker. *Probability and Random Processes*. Oxford University Press Inc., 2 edition, 1995.

[8] X. Huang, Y. Ariki, and M. Jack. *Hidden Markov models for speech recognition*. Edinburgh University Press, Edinburgh, UK, 1990.

[9] F. Jelinek. *Statistical methods for speech recognition*. The MIT Press, Cambridge, MA, USA, 2001.

[10] B.-H. Juang and L.R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Signal Proc.*, 38(9):1639–1641, 1990.

[11] A. Koloydenko, M. Käärik, and J. Lember. On Adjusted Viterbi training. *Acta Appl. Math.*, 96(1—3):309—326, 2007.

[12] J. Lember and A. Koloydenko. Adjusted Viterbi training: A proof of concept. *Probab. Eng. Inf. Sci.*, 21(3):451—475, 2007.

[13] J. Lember and A. Koloydenko. The Adjusted Viterbi training for hidden Markov models. *Bernoulli*, 2007. Invited revision submitted.

[14] J. Li, R. Gray, and R. Olshen. Multiresolution image classification by hierarchical modeling with two dimensional hidden Markov models. *IEEE Trans. Inf. Theory*, 46(5):1826–1841, 2000.

[15] H. Ney, V. Steinbiss, R. Haeb-Umbach, and *et. al.* An overview of the Philips research system for large vocabulary continuous speech recognition. *Int. J. Pattern Recognit. Artif. Intell.*, 8(1):33–70, 1994.

[16] F. Och and H. Ney. Improved Statistical Alignment Models. In *Proceedings of 38th Annual Meeting of the Association for Computational Linguistics*, 2000.

[17] U. Ohler, H. Niemann, G. Liao, and G. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17(Suppl. 1):S199–S206, 2001.

[18] L. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[19] L. Rabiner and B. Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[20] L. Rabiner, J. Wilpon, and B. Juang. A segmental K-means training procedure for connected word recognition. *AT&T Tech. J.*, 64(3):21–40, 1986.

[21] V. Steinbiss, H. Ney, X. Aubert, and *et. al.* The Philips research system for continuous-speech recognition. *Philips J. Res.*, 49:317–352, 1995.

[22] N. Ström, L. Hetherington, T. Hazen, E. Sandness, and J. Glass. Acoustic Modeling Improvements in a Segment-Based Speech Recognizer. In *Proc. IEEE ASRU Workshop Keystone, CO, USA*, 1999.