Adjusted Viterbi training for hidden Markov models

Adjusted Viterbi training for hidden Markov models

Jüri Lember*

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

Alexey Koloydenko

formerly with Eurandom, Eindhoven, the Netherlands.
Presently with Division of Statistics
The University of Nottingham. University Park. Nottingham NG7 2RD, UK.
Tel: +44(0)115.951.4937, alexey.koloydenko@nottingham.ac.uk

http://www.eurandom.tue.nl/EURANDOMreports.htm

December 2, 2005

Abstract

We consider estimation of the emission parameters in hidden Markov models. Commonly, one uses the EM algorithm for this purpose. However, our primary motivation 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. For this reason 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 with relevant simulations ascertaining the theory. This paper shows how the adjusted Viterbi training is also possible for more general hidden Markov models.

*Estonian Science Foundation Grant 5694

1 Introduction

We consider a set of procedures to estimate the emission parameters of a finite state hidden Markov model given observations $x_1, \ldots, x_n$. Thus, Y is a Markov chain with (finite) state space S, transition matrix (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, 7, 8, 15, 16]. 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 [7, 12, 16, 17, 18, 19], natural language modeling [13], image analysis [11], bioinformatics [4, 14]. We are also motivated by connections with constrained vector quantization [3, 5]. 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 [9, 17] (see also [15, 16]). 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 [3], 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 procedure? 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 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 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

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

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 P_l^n. First of all, one needs to know if there exists any limiting probability measures Q_l such that for every $l \in S$

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

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

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

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 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

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

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 pass 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 subalignments each for each segment enclosed by two consequitive 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. Now, the definition of the node guarantees the existence of state l such that any alignment passes through l at time u independently of the observations outside the barrier.

In Lemma 3.1, we prove (under certain assumptions) the existence of a special subsequence, or a block, of Y states such first, the subsequence itself occurs with a positive probability, and second, with a positive probability, it emmits a barrier. 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 passing throught 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 furthemore, 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 dividors, 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 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 $Q_l^n \Rightarrow Q_l$, a.s and $P_l^n \Rightarrow Q_l$ (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 [10], VA is described for the mixture case. The simulations in [10] 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, computations 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 $P = (p_{ij})$ and initial distribution π that is also the stationary distribution of Y. We consider the 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 $\mathcal{B}$ to be a separable metric space and the corresponding Borel σ-algebra, respectively. Let f_l be the density of P_l with respect to some reference measure λ on $(\mathcal{X}, \mathcal{B})$, which one for concreteness may want to specialize to the Lebesgue measure.

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 f such that for each n,*

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

Hence, the emission distribution P_l is the distribution of $f(l, e_n)$. 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:

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

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 (π, 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 P and π are known, but the emission densities are known only up to the parametrization $f_l(\cdot; \theta_l)$, $\theta_l \in \Theta_l$. In this 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

$$\hat{P}_l^n(A; \theta^j, x_{1\ldots n}) := \frac{\sum_{i=1}^n I_{A\times l}(x_i, v_i)}{\sum_{i=1}^n I_l(v_i)}, \quad A \in \mathcal{B}, \quad l \in S. \tag{5}$$

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), $l \in S$, such that for every l

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

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 $\mathcal{F}_l = \{f_l(\theta_l) : \theta_l \in \Theta_l\}$, 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

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

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 [10], 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 pointwise 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 [10] 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 [7, 15, 16], 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 $1 \leq u_1 < u_2 < \ldots < u_k \leq n$. 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:

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

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

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

Next, define the *scores*

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

and notice the trivial case: $\delta_l(1) = \pi_l f_l(x_1)$. Then, we have the following recursion (see, for example, [16]):

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

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:

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

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

It can also be shown that

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

We shall also need the following notation:

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

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

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

Similarly, $\mathcal{W}_{(l)u_1\ldots u_k}^{l_1\ldots l_k}(x_{1\ldots n})$ 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 $1 \leq u \leq n$, then*

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

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

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

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

where

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

Thus, (20) follows from the equivalence between maximizing $\Lambda(q; x_{1\ldots n})$ over $S_u^l(n)$ on one hand, and maximizing $\Lambda(q_{1\ldots u}; x_{1\ldots u})$ and $\Lambda(q_{u+1\ldots n}; x_{u\ldots n}|l)$ 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

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

Proof. The hypotheses of (22) together with (21) imply $\mathcal{V}_u^l(x_{1\ldots n}) = \mathcal{W}_u^l(x_{1\ldots n})$ and $\mathcal{V}_u^l(x_{1\ldots u}) = \mathcal{W}_u^l(x_{1\ldots u})$. 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:

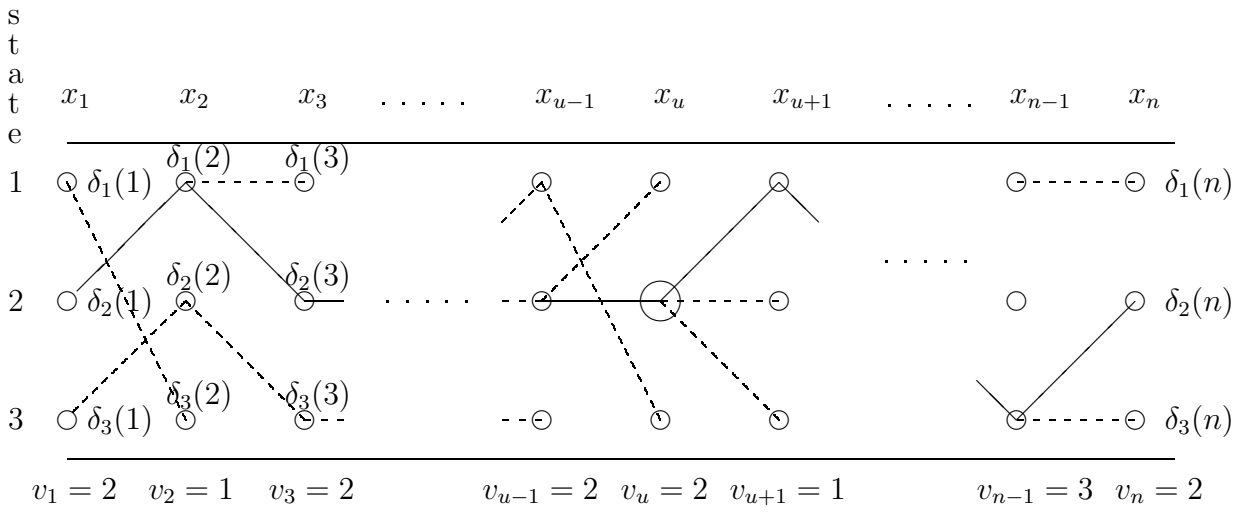

Figure 1: An example of the Viterbi algorithm in action. The solid line corresponds to the final alignment $v_{1\ldots n}$. The dashed links are of the form $(k,l)-(k+1,j)$ with $l \in t(k,j)$ and are not part of the final alignment. E.g., $(1,3)-(2,2)-(3,3)$ is because $3 \in t(1,2)$, $2 \in t(2,3)$. 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*.

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

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

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.

Proposition 3.2

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

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

$$\delta_l(u) \ge \delta_i(u), \quad \forall i \in S, \tag{29}$$

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 $1 \leq u_1 < u_2 < \cdots < u_k \leq n$, 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 $\{w^{ql}(x_{1\ldots m}):\mathbb{R}^m\to\mathcal{W}_{(q)m}^{\ \ l}(x_{1\ldots m}),\ m=1,\ldots,n,\ q,l\in S\}$:

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

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* $i=2,\ldots,k-1$

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

where $\{w^{ql}(x_{1\ldots m}):\mathbb{R}^m\to\mathcal{W}_{(q)m}^{\ \ l}(x_{1\ldots m}),\ m=1,\ldots,n,\ q,l\in S\}$ *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

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

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 $w^{\pi l}(x_{1\ldots m})\stackrel{\mathrm{def}}{=}\max_{\prec}\mathcal{W}_m^l(x_{1\ldots m})$, $1\leq m\leq n$, 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* $\sum_{i=1}^K p_if_i$. *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*

$$\delta_l(u)\geq\delta_i(u),\quad\forall i.$$

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

$$\delta_i(u+1) = \max_j(\delta_j(u))p_i f_i(x_{u+1}) \propto p_i f_i(x_{u+1}), \quad \forall i$$

and

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

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 pointwise: $v(x_{1\ldots n}) = (v(x_1), \ldots, v(x_n))$, 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 $\delta_i(u) > 0$ for every i. In this case, (23) is equivalent to

$$\delta_l(u) \geq \max_i\Big(\max_j\Big(\frac{p_{ij}}{p_{lj}}\Big)\delta_i(u)\Big)$$

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 P in general can have a zero in every row, violating the above positivity constraint on the l^{th} row of 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

$$p_{ij}^{(r)}(u) = \max_{q \in S} p_{iq}(u) f_q(x_{u+1}) p_{qj}^{(r-1)}(u+1).$$

The recursion (16) then generalizes to

$$\delta_j(u+1) = \max_i\big(\delta_i(u-r)p_{ij}^{(r)}(u-r)\big) f_j(x_{u+1}), \ r < u.$$

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

$$\begin{aligned} t^{(r)}(u,j) &= \{l \in S : \forall i \in S \ \delta_l(u) p_{lj}^{(r-1)} \geq \delta_i(u) p_{ij}^{(r-1)}\}, \\ t^{(r)}(u,J) &= \{t^{(r)}(u,j) : \ j \in J\}, \quad J \subset S. \end{aligned} \tag{37}$$

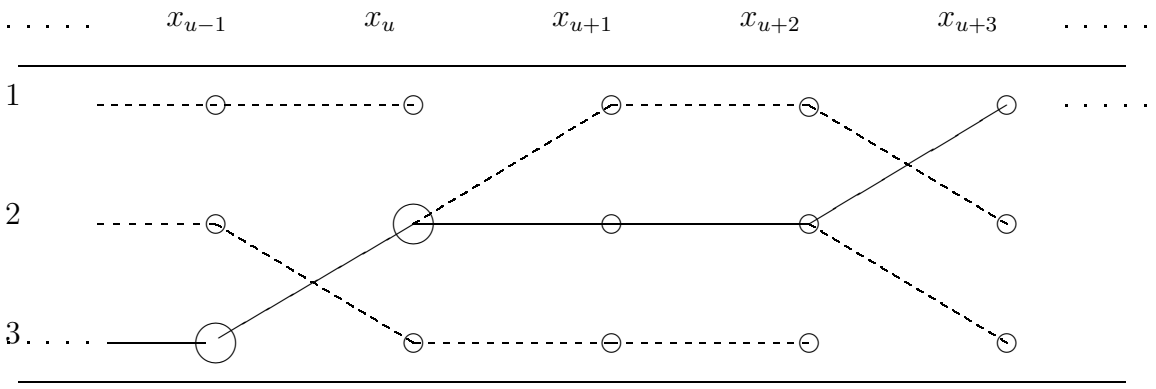

Figure 2: In this example, x_u is a 2^d order 2-node, x_{u-1} is a 3^d-order 3-node. Thus, for given $x_{1...n}$, the alignment includes $v_u = 2$. However, unlike in the case of ordinary nodes (of order 0), x_{u+1} can now destroy the property of x_u being the (second order) node.

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

$$t^{(r+q)}(u, j) = t^{(q)}(u, t^{(r)}(u + q, j)), \tag{38}$$

where $t^{(1)}(u, j)$ coincides with $t(u, j)$ (18). Thus, $l_1 \in t^{(q)}(u, t^{(r)}(u + q, j))$ in (38) 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:

$$\mathcal{V}(x_1, \ldots, x_n) = \{v \in S^n : \delta_{v_n}(n) \geq \delta_i(n)\ \forall i \in S, v_u \in t^{(n-u)}(u, v_n)\ 1 \leq u < n\} \tag{39}$$

$$\mathcal{W}^l_u(x_1, \ldots, x_n) = \{v \in S^l_n(n) :\ v_u \in t^{(n-u)}(u, l)\ 1 \leq u < n\} \tag{40}$$

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*

$$\delta_l(u) p^{(r)}_{lj}(u) \geq \delta_i(u) p^{(r)}_{ij}(u), \quad \forall i, j \in S. \tag{41}$$

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

$$x_u \text{ is an } l\text{-node of order } r \iff l \in t^{(r+1)}(u,j)\ \forall j \in S, \tag{42}$$

$$u+r<n, x_u \text{ is an } l\text{-node of order } r \Rightarrow \forall v \in \mathcal{V}(x_{1\ldots n}), \forall v^* \in \mathcal{W}_u^l(x_{1\ldots u})$$

$$\exists v' \in \mathcal{W}_{u\ u+r+1}^{l\ v_{u+r+1}}(x_{1\ldots u+r+1}) : v^* = v'_{1\ldots u}, (v', v_{u+r+1\ldots n}) \in \mathcal{V}_u^l(x_{1\ldots n}), \tag{43}$$

$$\Rightarrow \mathcal{V}_u^l(x_{1\ldots n}) \neq \emptyset, \tag{44}$$

$$\Rightarrow \mathcal{V}_u^l(x_{1\ldots n}) = \mathcal{W}_u^l(x_{1\ldots u}) \times \mathcal{V}_{(l)}(x_{u+1\ldots n}). \tag{45}$$

Finding $v'_{u+1\ldots u+r}$ and $v^* \in \mathcal{W}_u^l(x_{1\ldots u})$ in (43) 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). (42) also follows immediately from Definition 3.5 and (37). In order to see (43), note that applying (38) 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 (38) 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 (40), $\tilde{v} = v'_{u+1\ldots u+r}$ for some $v' \in \mathcal{W}_{u\ u+r+1}^{l\ v_{u+r+1}}(x_{1\ldots u+r+1})$. 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 $\delta_{v_n}(n) \geq \delta_j(n)\ \forall j \in S$ ($v \in \mathcal{V}(x_{1\ldots n})$ 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, (43) implies (44). Finally, given (44), Proposition 3.1 yields (45). ■

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*

$$\emptyset \neq \mathcal{V}_{u_1 u_2 \cdots u_k}^{l_1 l_2 \cdots l_k}(x_{1\ldots n}) = \\ = \mathcal{W}_{u_1}^{l_1}(x_{1\ldots u_1}) \times \mathcal{W}_{(l_1)u_2}^{\ l_2}(x_{u_1+1\ldots u_2}) \times \cdots \times \mathcal{W}_{(l_{k-1})u_k}^{\ l_k}(x_{u_{k-1}+1\ldots u_k}) \times \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n}). \tag{46}$$

Proof. By (44), 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 (45),

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

Apply Corollary 3.3 to get

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

and repeat similarly to get

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

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

$$v = (v_1, \ldots, v_{k+1}) \in \mathcal{V}(x_{1\ldots n}), \tag{47}$$

where $v_1 \in \mathcal{W}^{l_1}_{u_1}(x_{1\ldots u_1})$, $v_2 \in \mathcal{W}^{\ l_2}_{(l_1) u_2}(x_{u_1+1\ldots u_2}), \ldots,\ v_k \in \mathcal{W}^{\ l_k}_{(l_{k-1}) u_k}(x_{u_{k-1}+1\ldots u_k}),\ v_{k+1} \in \mathcal{V}_{(l_k)}(x_{u_k+1\ldots n})$. Moreover, for every $i = 1, \ldots, k$, the vectors $w(i) \stackrel{def}{=} (v_1, \ldots, v_i)$ satisfy $w(i) \in \mathcal{W}^{l_i}_{u_i}(x_{1\ldots u_i})$ 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* (47) *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

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

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 class of HMMs with irreducible and aperiodic hidden layers as possible. Having 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 $\forall u\ 0 < u \leq n$, $\delta_j(u) > 0\ \forall j \in S$ (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 $p^{(r)}_{lj}(u) > 0\ \forall j \in S$. The latter in turn implies that $r_{lj} > 0$ for every $j \in S$, where r_{lj} is the entry lj of P^r. Thus, having an l-node of order r for **some** r does not impose any restriction on P: by virtue of irreducibility and aperiodicity of P, there always exists r_0 such that P^r 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 $x'_{1\dots w} \in S^w$, $w+u$-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 $y_{1\dots k} \in S^k$ $(k \geq r)$ can be such that for any $w > 0$ and for any $x'_{1\dots w} \in S^w$, $w+k-r$-th element of $(x'_{1\dots w}, y_{1\dots k})$ is a node of order r. $y_{1\dots k}$ in that case will be called *a barrier*.

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

Definition 3.9 *We say that a subset of states $C \subset S$ is a cluster, if, simultaneously,*

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

Hence, a cluster is a maximal subset of states such that the corresponding emission distributions have a "detectable" intersection of their supports. The clusters are not necessarily disjoint and a cluster can consist of a single state. In this latter case the state is not hidden: Any emission from this state indicates that state. 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 the Appendix.

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

$$P_l\left(x : f_l(x) \max_j\{p_{jl}\} > \max_{i, i \neq l}\{f_i(x) \max_j\{p_{ji}\}\}\right) > 0. \tag{48}$$

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 $Q = (p_{ij})_{i,j \in C}$ has all of its entries non-zero. Then, for some integer M and r, $M > r \geq 0$, there exists a set $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_M \subset \mathcal{X}^M$, an M-tuple of states $q_{1\dots M} \in S^M$, and a state $l \in S$, such that every vector $y_{1\dots M} \in \mathcal{Y}$ 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 $\mathbf{P}\big((X_1, \dots, X_M) \in \mathcal{Y}\big) > 0$. Every element of $\mathcal{Y}$ is a (r-order) barrier. By the ergodicity of X, a.e. realization of X therefore has infinitely many l-barriers of order r, hence each such realization 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 some 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 also occur infinitely often. Let $\mathcal{Y} \subset \mathcal{X}^M$ and M and r be as in Lemma 3.1 and $x_{j\ldots j+M-1} \in \mathcal{Y}$, i.e. is an l-barrier of order r for some $l \in S$ and some $j > 0$, 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 $\mathcal{Y}$. 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 $\mathcal{Y} \subset \mathcal{X}^k$ such that all its elements are l-barriers of order r ($r \le k$), and let $y_{1\ldots k} \in \mathcal{Y}$. We say that $y_{1\ldots k}$ is a separatedr^{th} order l-barrier relative to $\mathcal{Y}$ if for any w, $1 \le w \le r$, and for any $y'_{1\ldots w} \in \mathcal{X}^w$ the concatenated block $(y'_{1\ldots w}, y_{1\ldots k-w}) \notin \mathcal{Y}$.*

In other words, a barrier is separated, if the distance from the previous barrier is at least $r + 1$. Next, Lemma 3.2 (proven in Appendix) shows that a.e. realization of X has infinitely many separated barriers.

Lemma 3.2 *Suppose the assumptions of Lemma 3.1 are satisfied. Let M, r, $\mathcal{Y} \subset \mathcal{X}^M$ and $q_{1\ldots M}$ be as promised by Lemma 3.1. Then, for some $N \ge M$, there exist a set $\mathcal{Y}^* = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{N-M} \times \mathcal{Y} \subset \mathcal{X}^N$, an N-tuple of states $q^* = (q^*_{1\ldots N-M}, q_{1\ldots M}) \in S^N$, and a state $l \in S$ such that every vector $y_{1\ldots N} \in \mathcal{Y}^*$ is an l-barrier of order r, and moreover $y_{N-M+1\ldots N} \in \mathcal{Y}$ is a separated r^{th}-order l-barrier relative to $\mathcal{Y}$, $q_{M-r} = l$ and*

$$\mathbf{P}\Big((X_1, \ldots, X_N) \in \mathcal{Y}^* \Big| (Y_1, \ldots, Y_N) = q^*\Big) > 0, \quad \mathbf{P}\Big((Y_1, \ldots, Y_N) = q^*\Big) > 0.$$

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*

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

Let the emission distributions be such that (48) 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 Q_i corresponding to C_i, $i = 1, 2$ are

$$Q_1 = Q_2 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

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 $t(u, j)^{(r)}$ (38), 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 (42)). 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 (42); 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 $\mathcal{Y} = (0, +\infty) \times (\log(2), +\infty) \times (0, +\infty)$ 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 (48) holds. Now, the model has only one cluster that is $S = \{1, \ldots, 4\}$. Since the matrix 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,*

$$z_1, z_2, z_3, y_1, \ldots, y_k, z'_1, z'_2, z'_3, \tag{49}$$

where $z_i, z_i' \in \cap_{j=1}^{4} G_j$, $i = 1, 2, 3$, $y_i \in \mathcal{X}$, $i = 1, \ldots k$ *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 $M \geq 0$, $\mathcal{Y} \subset \mathcal{X}^M$, $r \geq 0$, and $l \in S$ as promised by Lemma 3.1. Then, by Lemma 3.2, for some $N > r \geq 0$, there exist $\mathcal{Y}^* = \mathcal{Y}' \times \mathcal{Y} \subset \mathcal{X}^N$ for some $\mathcal{Y}' \subset \mathcal{X}^{N-M}$, and an N-tuple of states $q^* = q_{1\ldots N} \in S^N$ such that for every n,

$$\mathbf{P}\Big((Y_n, \ldots, Y_{n+N-1}) = q^*\Big) > 0, \quad \mathbf{P}\Big((X_n, \ldots, X_{n+N-1}) \in \mathcal{Y}^* \Big| (Y_n, \ldots, Y_{n+N-1}) = q^*\Big) > 0$$

hence every $x_{n\ldots n+N-1} \in \mathcal{Y}^*$ is (ends with) a separated barrier from $\mathcal{Y}$.
Denote $\mathbf{P}\Big((X_n, \ldots, X_{n+N-1}) \in \mathcal{Y}^*\Big)$ by γ^*, thus $\gamma^* > 0$, and define

$$U_n = (X_n, \ldots, X_{n+N-1}), \quad D_n = (Y_n, \ldots, Y_{n+N-1}). \tag{50}$$

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

$$\nu_o := \min\{n \geq 1 : U_n \in \mathcal{Y}^*, D_n = q^*\}, \quad \nu_i := \min\{n > \nu_{i-1} : U_n \in \mathcal{Y}^*, D_n = q^*\}; \tag{51}$$

$$\vartheta_o := \min\{n \geq 1 : U_n \in \mathcal{Y}^*\}, \quad \vartheta_i := \min\{n > \vartheta_{i-1} : U_n \in \mathcal{Y}^*\}; \tag{52}$$

$$R_0 := \min\{n \geq 1 : \ D_n = q^*\}, \quad R_i := \min\{n > R_{i-1} : \ D_n = q^*\}. \tag{53}$$

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 by looking at the X process only; the stopping times R are observable by looking at the Y process only; the sopping times ν require the knowledge of the full two-dimensional process (X, Y). Clearly $\vartheta_i \leq \nu_i$, and $R_i \leq \nu_i$.

From (53), 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 $\nu_0, (\nu_1 - \nu_0), (\nu_2 - \nu_1), \ldots$. In Appendix, we prove that all the latter ones have finite expectations, and, therefore, are stationary renewal times:

Proposition 4.1 *$E(\nu_1 - \nu_0) < \infty$ and for any π' (not necessarily equal to π, the invariant distribution) initial distribution of Y, $E_{\pi'}\nu_0 < \infty$.*

To every ν_i, $i = 0, 1, \ldots$ there corresponds an r-order l-barrier. This barrier occupies the interval $[\nu_i + N - M, \nu_i + N - 1]$. By Definition 3.8, X_{τ_i} is an l-node of order r, where

$$\tau_i = \nu_i + (N - 1) - r, \quad i = 0, 1, \ldots$$

Define

$$T_0 \stackrel{\text{def}}{=} \tau_0, \quad T_i \stackrel{\text{def}}{=} \tau_i - \tau_{i-1} = (\nu_i - \nu_{i-1}),\ i = 1, 2, \ldots. \tag{54}$$

From Proposition 4.1, it immediately follows $E_{\pi'} T_1 < \infty$, $E_{\pi'} T_0 < \infty$, where π' is any initial distribution of Y. Thus T_i, $i = 0, 1, \ldots$ correspond to a delayed renewal process [6].

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

$$u_i = \vartheta_i + (N-1) - r, \quad i = 0, 1, 2, \ldots. \tag{55}$$

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

4.2 Alignments

We next specify the alignments $v(x_{1\ldots n}) \in \mathcal{V}(x_{1\ldots n})$, and define $v(x_{1\ldots\infty})$ and the measures 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:

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

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 (56) below). Namely, we state

Definition 4.1

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

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 $v(x_{1\ldots\infty}) \stackrel{\text{def}}{=} (\vee \mathcal{W}^l_{u_0}(x_{1\ldots u_0}), \vee \mathcal{W}^{\ l}_{(l)u_1}(x_{u_0+1\ldots u_1}), \ldots, \vee \mathcal{W}^{\ l}_{(l)u_{k-1}}(x_{u_{k-2}+1\ldots u_{k-1}}), 1, 1, \ldots)$, 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 $(x_1, \ldots, x_n)$

$$v(x_1^\infty)_{1\cdots u_i} = v(x_1, \ldots, x_n)_{1\cdots u_i}. \tag{56}$$

Let us now formally define the empirical measures 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*

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

For other $l \in S$ (i.e. such that $l \neq V_i'$ for $i = 1, \ldots, n$), define P_l^n to equal some 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

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

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

4.3 Regenerativity

To prove our main theorem, we use the fact that Z is a regenaritive process:

Proposition 4.2 *The processes V, X, and Z are regenerative with respect to τ.*

The proof is given in the Appendix.

Recall $\mathcal{Y}^* = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_{N-M} \times \mathcal{Y}_{N-M+1} \times \cdots \times \mathcal{Y}_N$ and $q^* = q_{1\ldots N}$ (from §4.1), and let

$$P_{q_i}^r \propto P_{q_i} I_{\mathcal{Y}_i}, \quad i = 1, \ldots, N.$$

Thus, $P_{q_i}^r$ is the measure $P_{q_i}^r$ conditioned on $\mathcal{Y}_i$. Recall also that $q_{N-r} = l$.

Define new processes

$$Y^r \stackrel{\text{def}}{=} (Y_i^r)_{i=1}^\infty, \quad \text{where} \quad Y_1^r = q_{N-r+1}, \ldots, Y_r^r = q_N, \quad \text{and} \quad Y_{r+1}^r, Y_{r+2}^r, \ldots \tag{57}$$
is an S-valued Markov chain with transition probability matrix P and initial distribution $(p_{q_N j})_{j\in S}$;

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

$$V^r \stackrel{\text{def}}{=} (V_i^r)_{i=1}^\infty \stackrel{\text{def}}{=} v(X^r), \text{ where } v \text{ is as in Definition 4.1;} \tag{58}$$

$$Z^r \stackrel{\text{def}}{=} (X^r, V^r). \tag{59}$$

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 rest. 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}, $P_{q_2}, \ldots, P_{q_n}$ (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 fro the initial distribution being $(p_{lj})_{j\in S}$ (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 $(p_{lj})_{j\in S}$ as the initial distribution of its Y-component.

Finally we define analogues of ν_0 and τ_0:

$$\begin{aligned} \nu_0^r &\stackrel{\text{def}}{=} \min\left\{ n \geq 1 : (Y_n^r, \ldots, Y_{n+N-1}^r) = q^*, \quad (X_n^r, \ldots, X_{n+N-1}^r) \in \mathcal{Y}^* \right\} \\ \tau_0^r &\stackrel{\text{def}}{=} \nu_0^r + (N-1) - r. \end{aligned} \tag{60}$$

Note that the random variable τ_0^r has the same law as T_i (54), $i \geq 1$. The barriers in $\mathcal{Y}^*$ end by (length M) l barriers of order r from $\mathcal{Y} = \mathcal{Y}_{N-M+1} \times \mathcal{Y}_{N-M+2} \times \cdots \times \mathcal{Y}_N$ that are separated relative to $\mathcal{Y}$ (Definition 3.10, Lemma 3.2). Therefore, $\nu_0^r > r$. This means that the laws of ν_0^r, τ_0^r, $\nu_0 + r$, and $\tau_0 + r$ would all be equal if Y had $(p_{q_M l})_{l\in S}$ for the initial distribution. Recalling that any initial distribution π of Y yields $E_\pi(\nu_0) < \infty$ (Proposition 4.1), we obtain

$$ET_1 = E\tau_0^r = E_{q_M}(\nu_0 + (N-1) - r + r) < \infty. \tag{61}$$

The above observations will allow us to prove (see Appendix) 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*

$$P_l^n \Rightarrow Q_l, \quad a.s., \quad Q_l^n \Rightarrow Q_l, \quad a.s.$$

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

$$Q_l(A) = \frac{\sum_{i=1}^\infty \mathbf{P}(Z_i^r \in A \times l, \tau_o^r \geq i)}{\sum_{i=1}^\infty \mathbf{P}(V_i^r = l, \tau_o^r \geq i)}. \tag{62}$$

where V^r, Z^r, and τ_o^r are defined in (58), (59), *and* (60), *respectively.*

Corollary 4.1 *Suppose X satisfies the assumptions of Lemma 3.1 with $r = 0$. Then, for each $l \in S$ (62) takes form*

$$Q_l(A) = \frac{\sum_{j=1}^\infty \mathbf{P}_l(Z_j \in A \times i, \tau_o \geq j)}{\sum_{j=1}^\infty \mathbf{P}_l(V_j = i, \tau_o \geq j)}, \tag{63}$$

where P_l stands for initial distribution of Y equal $(p_{lj})_{j\in S}$.

5 Appendix

5.1 Proof of Lemma 3.1

Work in progress.

5.2 Proof of Lemma 3.2

Work in progress.

5.3 Proof of Proposition 4.1

We define blocked processes

$$U_m^b = (X_{(m-1)N+1}, \dots, X_{mN}), \quad D_m^b = (Y_{(m-1)N+1}, \dots, Y_{mN}), \quad m = 1, 2, \dots,$$

and stopping times

$$\begin{aligned} \nu_o^b &:= \min\{m \geq 1 : U_m^b \in \mathcal{Y}^*, D_m^b = Q^*\} \\ \nu_i^b &:= \min\{m > \nu_{i-1}^* : U_m^b \in \mathcal{Y}^*, D_m^b = Q^*\}; \end{aligned} \tag{64}$$

$$\begin{aligned} R_0^b &:= \min\{m > 1 : \ D_m^b = Q^*\}, \\ R_i^b &:= \min\{m > R_{i-1} : \ D_m^* = Q^*\}. \end{aligned} \tag{65}$$

The process D^b is a finite MC, since Y is non-periodic and irreducible, the same holds for D^b. Hence (D^b, U^b) is a HMM.
Since Y is stationary, Q^* occurs in every possible integer-interval with the same probability; so Q^* belongs to the state space of D^b. Since D^b is irreducible and finite, then every state is visited infinitely often, a.s. This means that for every $i = 1, 2, \dots$, $\mathbf{P}(R_i^b - R_{i-1}^b < \infty) = 1$ and then, of course, $\mathbf{P}(R_i - R_{i-1} < \infty) = 1$.
If $D_m^b = Q^*$, then by Lemma 3.2, $U_m^b \in \mathcal{Y}^*$ with probability $\gamma^* > 0$. This means that for every $i = 1, 2, \dots$, $\mathbf{P}(\nu_i^b - \nu_{i-1}^b < \infty) = 1$ and this means $\mathbf{P}(\nu_i - \nu_{i-1} < \infty) = 1$.
The chain D^b has finite state space and is ergodic. Hence, for any initial distribution of D^*, we have $E(R_0) \leq NE(R_0^b) < \infty$. Every initial distribution π of Y induces an initial distribution of D^b. Thus, for any π, $E_\pi(R_0) < \infty$. Now, obviously, $E\nu_0 = \gamma^{*-1} E_\pi(R_0) < \infty$. This proves the second statement.

To prove the first statement, consider the blocked Markov chain D^b with initial distribution concentrated on Q^*. The fact that $E(R_0) \leq NE(R_0^b) < \infty$ with any initial distribution implies that $E(R_1 - R_0) < \infty$ and, hence, $E(\nu_1 - \nu_0) = (\gamma^*)^{-1} E(R_1 - R_0) < \infty$.

5.4 Proof of Proposition 4.2

Recall the definition of stopping times τ. By definition, for each i 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. With the fact that T_i are renewal, this establishes the regenerativity of

X. To every τ_i corresponds a r-order l-node. Moreover, since τ_i is always a u_j, for some $j > i$, it means that all the nodes corresponding to τ_i-s are also used to fix the alignment in Definition 4.1. Therefore, the alignment up to τ_i does not depend on the alignment after τ_i. In other words, the piece of alignment process corresponding to the T_i is a function of the piece of X corresponding to the T_i. Formally

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

So, the process Z is regenerative with respect toτ.

5.5 Proof of Theorem 4.4

At first note that the right side of (62) defines a measure.
The proof uses the regenerativity of Z in the most standard way. For every $n \geq \tau_o$ and $A \in \mathcal{B}$, and for every $l \in S$.

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

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

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

Let $\mu := E\tau_0^r$. By (61), $\mu < \infty$. Then

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

Finally, since Z is a regenerative process with respect to the $\tau_o, \tau_1, \dots$, we have

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

where

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

are i.i.d. random variables. Denote $m_l(A) := E\xi_k \leq \mu < \infty$ (and drop the A, because it is fixed). Then, as $n \to \infty$, we have

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

Let us calculate m_l. Clearly,

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

Now

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

Similarly,

$$
\frac{1}{n}\sum_{i=1}^{n} I_l(V_i^r) \to \frac{1}{\mu}\sum_{i=1}^{\infty} P(V_i^r = l, \tau_0^r \geq i) =: \frac{w_l}{\mu} \leq 1, \quad \text{a.s.} \tag{66}
$$

Hence, we have shown that for each $l \in S$ and for every $A \in \mathcal{B}$

$$
Q_l^n(A) \to \frac{m_l(A)}{w_l} = \frac{\sum_{i=1}^{\infty} \mathbf{P}(Z_i^r \in A\times l, \tau_0^r \geq i)}{\sum_{i=1}^{\infty} \mathbf{P}(V_i^r = l, \tau_0^r \geq i)}, \quad \text{a.s.} \tag{67}
$$

The theory of weak convergence of measures (recall $\mathcal{X}$ was assumed to be a separable metric space) now establishes $Q_l^n \Rightarrow Q_l$, a.s..

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

$$
P_l^n(A) \to \frac{m_l(A)}{w_l}, \quad \text{a.s..} \tag{68}
$$

For this consider the sum $\sum_{i=1}^{n} I_{A\times l}(X_i, V_i')$. Since $V_i' = V_i$, if $i \leq \tau_{k(n)}$, we get, as $n \to \infty$

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

Similarly,

$$
\frac{1}{n}\sum_{i=1}^{n} I_l(V_i') \to \frac{1}{\mu}\sum_{i=1}^{\infty} P_1(V_i = l, \tau_0^r \geq i) =: \frac{w_l}{\mu}, \quad \text{a.s..} \tag{69}
$$

These convergences prove (68).

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. Chou, T. Lookbaugh, and R. Gray. Entropy-Constrained Vector Quantization. *IEEE Transaction on Acoustic Speech and Signal Processing*, 37(1):31–42, 1989.

[4] G. Ehret, P. Reichenbach, U. Schindler, C. Horvath, S. Fritz, M. Nabholz, and P. Bucher. DNA Binding Specificity of Different STAT Proteins. *The Journal of Biological Chemistry*, 276(9):6675–6688, 2001.

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

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

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

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

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

[10] J. Lember and A Koloydenko. Adjusted Viterbi training. a proof of concept. Technical Report 000, Eurandom, Eindhoven, The Netherlands. http://www.eurandom.tue.nl/EURANDOMreports.htm, 2005.

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

[12] H. Ney, V. Steinbiss, R. Haeb-Umbach, B. Tran, and U. Essen. An overview of the Philips research system for large vocabulary continuous speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(1):33–70, 1994.

[13] F. Och and H. Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, A digital archive of research papers in computational linguistics: http://acl.ldc.upenn.edu/P/P00/P00-1056.pdf, 2000.

[14] 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.

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

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

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

[18] V. Steinbiss, H. Ney, X. Aubert, S. Besling, C. Dugast, U. Essen, D. Geller, R. Haeb-Umbach, R. Kneser, H. Meyer, M. Oerder, and B. Tran. The Philips research system for continuous-speech recognition. *Philips Journal of Research*, 49:317–352, 1995.

[19] 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*, MIT Comp. Sci. and AI Lab., Spoken Language Systems http://www.sls.lcs.mit.edu/sls/publications/1999/asru99-strom.pdf, 1999.