A constructive and implementation-aware proof of the existence of Viterbi processes

Jüri Lember, Alexey Koloydenko

***Abstract*—Since the early days of digital communication, hidden Markov models (HMMs) have now been also routinely used in speech recognition, processing of natural languages, images, and in bioinformatics. In an HMM $(X_t, Y_t)_{t\geq 1}$, observations $X_1, X_2, \ldots$ are assumed to be conditionally independent given an "explanatory" Markov process $Y_1, Y_2, \ldots$, which itself is not observed; moreover, the conditional distribution of X_t depends solely on Y_t. Central to the theory and applications of HMM is the Viterbi algorithm to find *a maximum a posteriori* (MAP) estimate $v(x_{1:T}) = (v_1, v_2, \ldots, v_T)$ of $Y_{1:T}$ given observed data $x_{1:T}$. *Maximum a posteriori* paths are also known as Viterbi paths, or alignments. Recently, attempts have been made to study the behavior of Viterbi alignments when $T \to \infty$. Thus, it has been shown that in some special cases a well-defined limiting Viterbi alignment exists. While innovative, these attempts have relied on rather strong assumptions and involved proofs which are existential. This work proves the existence of infinite Viterbi alignments in a more constructive manner and for a very general class of HMMs.**

***Index Terms*—Asymptotic, HMM, maximum a posteriori path, Viterbi algorithm, Viterbi extraction, Viterbi training.**

I. Introduction

LET $Y = (Y_t)_{t\geq 1}$ be a Markov chain with state space $S = \{1, \ldots, K\}$, $K > 1$, and transition matrix $\mathbb{P} = (p_{ij})_{i,j\in S}$. Suppose that Y is irreducible and aperiodic, hence a unique stationary distribution $\pi = \pi\mathbb{P}$ exists; suppose further that $Y_t \sim \pi$ from time $t = 1$. To every state $i \in S$, let us assign an *emission distribution* P_i on $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is usually $\mathbb{R}^d$, the d-dimensional Euclidean space for some $d \geq 1$, and $\mathcal{B}$ is the Borel σ-algebra of $\mathcal{X}$. Let f_i be the density of P_i with respect to a suitable reference measure λ on $(\mathcal{X}, \mathcal{B})$. Most commonly, λ is either the Lebesgue measure (continuously distributed X_t) or the counting measure (discretely distributed X_t).

Definition 1.1: The stochastic process (X, Y) is a hidden Markov model if there is a (measurable) function g such that for each $t = 1, 2, \ldots$, $X_t = g(Y_t, \xi_t)$, where $\xi_1, \xi_2, \ldots$ are i.i.d. and independent of Y.

Hence, the emission distribution P_i is the distribution of $g(i, \xi_1)$. The distribution of X is completely determined by $\mathbb{P}$ and the emission distributions P_i, $i \in S$. It can be shown that X is also ergodic [1], [2], [3]. Let $x_{1:T} = (x_1, \ldots, x_T)$ and $y_{1:T} = (y_1, \ldots, y_T)$ be fixed observed and unobserved realizations, respectively, of the HMM $(X_t, Y_t)_{t\geq 1}$ up to time $T \geq 1$. Treating $y_{1:T}$ as missing data [4], or parameters to be estimated, let $\Lambda(y'_{1:T}; x_{1:T})$ be the complete likelihood function, henceforth simply the likelihood, $\mathbf{P}(Y_{1:T} = y'_{1:T}) \prod_{t=1}^{T} f_{y'_t}(x_t)$ of $y'_{1:T} \in S^T$, and let $\mathcal{V}(x_{1:T})$ be the set of the corresponding maximum likelihood estimates $v(x_{1:T}) \in S^T$ of $y_{1:T}$. The elements of $\mathcal{V}(x_{1:T})$ are called *(Viterbi) alignments* and are commonly computed by the Viterbi algorithm [5], [4]. If $\mathbf{P}(Y_{1:T} = y'_{1:T})$ is thought of as the prior distribution of $Y_{1:T}$, then the $v(x_{1:T})$ also maximize the probability mass function of the posterior distribution of $Y_{1:T}$, hence the term *maximum a posteriori (MAP) paths*.

At the same time, the Viterbi alignments differ, often significantly, from the actual realization $y_{1:T}$. Hence, the natural question: Are these deviations "purely random", or is there anything systematic in their (asymptotic, i.e. as $T \to \infty$) behavior?

Besides their direct significance for prediction of Y from X, the Viterbi alignments, or MAP paths, are also central to the theory and applications of HMMs [6] in the more general setting of model estimation. Namely, the emission distributions P_i and the transition probabilities p_{ij}, $i, j \in S$ are usually parameterized in practice, and some, or all, of the model parameters would be unknown and of interest. Thus, for example, the commonly used Viterbi Training (VT), or extraction, algorithm (also known as the *Baum-Viterbi* algorithm) [7] estimates the model parameters ignoring the discrepancies between the alignment and $y_{1:T}$. Hence the question about the incurred bias, and, subsequently, significance of the asymptotic behavior of the Viterbi alignments for the inference about the unknown parameters [6], [8].

Attempts to gain insight into the asymptotics of the alignments might immediately stall as the formal definition of $v(x_{1:T})$ does not automatically extend to the infinite sequence $x_{1:\infty}$. Indeed, it is then not clear whether any limiting, infinite Viterbi alignment $v(x_{1:\infty})$ exists at all. To appreciate that the question of extending $v(x_{1:T})$ *ad infinitum* is not a trivial one (even if the problem of non-uniqueness of $v(x_{1:T})$ is disregarded), suffice it to say that an additional observation x_{T+1} can in principle change the entire alignment based on $x_{1:T}$, i.e. $v(x_{1:T})$ and $v(x_{1:T+1})_{1:T}$ can disagree significantly, if not fully. Fortunately, the situation is not hopeless and in this paper we prove that in most HMMs the Viterbi alignments can be consistently extended *piecewise*. Specifically, motifs of (contiguous) observations $z_{1:M}$, called *barriers*, are observed with positive probability, forcing the Viterbi alignments based on extended observations $(x_{1:t}, z_{1:M}, x_{t+M+1:t+M+u})$ to stabilize as follows: Roughly, $\forall t, u \geq 0$ and for some τ such

J. Lember is with the Institute of Mathematical Statistics, Tartu University, J. Liivi 2-507, 50409, Estonia.
E-mail: jyri@ut.ee
A. Koloydenko is with the Mathematics Department of Royal Holloway University of London, Egham, TW20 0EX, UK.
E-mail: alexey.koloydenko@rhul.ac.uk
Manuscript received April 8, 2008; revised April 1, 2009

that $t < \tau \leq t + M$, and for all prefixes $x_{1:t} \in \mathcal{X}^t$ and all extensions $x_{t+M+1:t+M+u} \in \mathcal{X}^u$

$$v(x_{1:t}z_{1:M}x_{t+M+1:t+M+u})_{1:\tau} = v(x_{1:\tau}). \tag{1}$$

To be more specific, a particular state $l \in S$ and an element $z_k \in \mathcal{X}$, $1 \leq k \leq M$, called an *l-node*, can be found inside the *l-barrier* $z_{1:M}$, such that regardless of the observations before and after $z_{1:M}$, the alignment has to go through l at time $\tau = t + k$. The optimality principle then insures the stabilization (1) and in particular $v_\tau = l$.

There are further benefits of barriers. Namely, suppose now that $x_{1:T}$ contains $N \geq 1$ l-barriers with nodes occurring at times $\tau_1 < \cdots < \tau_N \leq T$. Then the Viterbi alignment $v(x_{1:T})$ can be constructed piecewise as follows: Let $v(x_{1:T}) = (v^1, v^2, \ldots, v^N, v^{N+1})$, where v^1 is the alignment based on $x_{1:\tau_1}$ and ending in l, and let v^n, for $n = 2, 3, \ldots, N+1$, be the conditional alignment based on $x_{\tau_{n-1}+1:\tau_n}$ given that $Y_{\tau_{n-1}} = l$; note that the alignments v^n, $n = 2, 3, \ldots, N$ also end in l, which is determined by the type of the barrier (node) being l. Now, if a new observation x_{T+1} is adjoined, then the last segment v^{N+1} can change, but the segments $v^1, \ldots, v^N$ remain intact. Suppose now that a realization $x_{1:\infty}$ contains infinitely many l-barriers, and hence also infinitely many nodes. Then the *(piecewise) infinite* alignment $v(x_{1:\infty})$ is defined naturally as the infinite succession of the segments v^1, v^2, ...

In this paper, we prove that for any HMM from a very wide class, there exists an integer $M > 0$, such that the probability that $X_{1:M}$ emits a barrier, is positive. Since X is ergodic, almost every realization $x_{1:\infty}$ has infinitely many barriers and, therefore, the infinite piecewise alignment is well-defined. Apparently, the piecewise alignment gives rise to a decoding process $v : \mathcal{X}^\infty \mapsto S^\infty$ via $V_{1:\infty} = v(X_{1:\infty})$, which we shall call the *Viterbi alignment process*. Thus, the results of this paper ensure that, *for a large class of HMMs the Viterbi alignment process V exists.* The piecewise construction also ensures that V is regenerative and ergodic. Besides the marginal process V, the joint processes (Y, V) and (X, Y, V) are also of interest as their asymptotic properties determine the systematic behavior of the alignment, including the deviations of the alignment V from the "truth" Y. Note also how this piecewise construction naturally calls for a buffered on-line implementation in which the memory used to store $x_{\tau_{n-1}:\tau_n}$ can be released once v^n has been computed.

A. Previous related work and contribution of this work

The problem of constructing infinite Viterbi processes has been brought to the attention of the IEEE Information Theory community fairly recently by [9] and [10]. Although the piecewise structure of the Viterbi alignments was already hinted at in [11] ('merge phenomenon') and acknowledged in [12], to our best knowledge, the subject has been first seriously considered in [9], [10]. In these latter works, the existence of infinite alignments for certain special cases, such as $K = 2$ and Markov chains with additive white Gaussian noise, has been proved. In particular, in these cases the authors of [9], [10] have proved the existence of 'meeting times' and 'meeting states', which are a special (stronger) type of nodes. While innovative, the main result of [9] (Theorem 2) makes several restrictive assumptions and is proved in an existential manner, which prevents its extension beyond the $K = 2$ case.

Independently of these works, [13], [8], [14] have developed a more general theory to address the problem of estimating unknown parameters (usually consisting of the emission parameters and the transition probabilities p_{ij}, $i, j \in S$). Namely, the focus of this theory has been the Viterbi *training* (VT) algorithm. Competing with EM-based procedures, this algorithm provides computationally and intuitively appealing estimates which, on the other hand, are biased, even in the limit when $T \to \infty$. In order to reduce this bias, the *adjusted Viterbi training* (VA) has been introduced in [13], [8], [14]. Naturally, VA relies on the existence of the infinite alignments and their ergodic properties. Although the general theory has been presented in [14], [8], some of the main results of the theory (Lemma 3.1 and 3.2 of [8]) have appeared without proof due to the limitations of scope and size. *This paper slightly refines these results, emphasizes their constructive and implementation-aware character, places them in a wider context of asymptotic (not necessarily Viterbi) alignments, touches on some new machine learning aspects, and, most importantly, presents the complete proofs of the main results.*

Whereas the present results are formulated for general HMMs ($K \geq 2$), [15] has most recently considered in full detail the special case of $K = 2$, generalizing similar results of [9], [10]. Specifically, it has been proved in [15] that infinitely many barriers (and hence the infinite Viterbi alignment) *exist for any aperiodic and irreducible 2-state HMM*. Thus, the results presented here extend the ones of [15] and [9], [10] to $K \geq 2$. It turns out that this extension is far from being straightforward and requires a more advanced analysis and tools, such as generalized, or higher order "weak" nodes mentioned above (§I) and first introduced in [8]. This generalization is not absolute in the sense that when $K > 2$, certain *aperiodic and irreducible HMMs can still fail to have infinitely many nodes*, undermining the piecewise construction of the infinite alignments for those models. Therefore, to guarantee that the infinite alignments can be constructed piecewise, certain assumptions, such as the cluster assumption in our main result Lemma 3.1, are indeed needed when $K > 2$. While the cluster condition can be further relaxed in obvious ways (cf. discussion below), we believe that its present version is a reasonable compromise between the generality of the Lemma and technical complexity of the proof.

The "disappearance" of the nodes has to do with the fact that an aperiodic and irreducible Markov chain can have zeros in the transition matrix. If this possibility is excluded, as is the case in [9], [10], then the 'meeting times' and 'meeting states' of [9], [10] are sufficient to prove the existence of infinite Viterbi alignments for many HMMs used in practice. In their recent communication with us, the authors of [9], [10] have corrected those statements in their aforementioned works where the strict positivity of the transition matrix is implicitly assumed but formally omitted (see [8] for details).

Models with forbidden transitions are indeed abundant in practice, and thankfully for a large class of such models the

generalized notion of nodes does effectively remove the limitations of the 'meeting times' and 'meeting states'. However, the price (e.g. length of the presented proofs) for this achievement has been rather high mostly due to the interfering issue of non-uniqueness of the (finite) Viterbi alignments. Specifically, ties in the Viterbi algorithm cause two complications. First, higher order nodes need to be a sufficiently large (but fixed) distance apart from each other. If they are not sufficiently separated, then breaking a tie in favor of one node can make it impossible to brake a tie involving another node in favor of that latter node, implying that the alignment cannot go through the both nodes. The separation of the nodes is not not an issue in [9], [10], [15] since for the special cases treated there, special ('strong' or 0th order) nodes are sufficient. Second, ties in distinct segments should be broken consistently. This would not be an issue if our goal were solely to have an infinite alignment. However, we also require the obtained alignment to be "proper" in the sense that the resulting Viterbi alignment process be *regenerative*. This additional property is crucial in the adjusted Viterbi training application ([8]) and should also be helpful in future analysis of various alignment-based statistics (e.g. barrier/node inter-arrival times). For a detailed treatment of the piecewise construction of the proper infinite alignment and regenerative Viterbi process in general HMMs, and the role of the infinite Viterbi process for the adjusted Viterbi training theory, we refer to the state-of-the-art article [8].

As far as we are aware, the presented constructive and implementation-aware approach is the first of its kind, especially given the level of its generality. For example, to see why it is problematic to extend the arguments of [9] to $K > 2$, note that the proof of the main existence result there (Theorem 2) is based on a contradiction. Namely, assuming that $K = 2$ and two Viterbi paths never meet, can be shown to contradict the Central Limit Theorem. The thereby exhibited meeting point corresponds to what we call a node of order 0. Unfortunately, that argument does not apply in the case $K > 2$, since for a node of order 0 to occur, $K > 2$ previously non-intersecting paths would have to coalesce at the same point. Moreover, in the more general situation allowing for zeros in the transition probabilities, a higher order node would be needed, requiring in turn that all possible pairs of the K (previously non-intersecting) paths meet in particular ways. Thus, the mere fact that any two paths have to meet *almost surely* does not any more guarantee the existence of the nodes.

By contrast the present approach is constructive and implementation-aware. To appreciate the latter feature, note that in order to determine whether an observation x_τ is a node, in general the entire history $x_{1:\tau}$ would need to be processed, which is, of course, not attractive in practice. At the same time, the presented results and their proof guarantee that the nodes encapsulated in the barriers are detectable by a sliding window filter of fixed width M (which, of course, depends on the HMM at hand). Presently, we do not set a goal to minimize M or to maximize the probability of detection (cf. §IV-B); this should of course be reconsidered in practice for any model individually. Note also that if we were merely concerned with the existence of the piecewise alignments, we would be content with the fact that any infinite sequence of nodes contains an infinite subsequence of separated nodes, and thus stop after proving Lemma 3.1. However, we are also concerned with applications, mainly the adjusted Viterbi training, and implementing such construction efficiently online. For these reasons, we make sure that the barrier detector does not respond to a new barrier unless the node inside this barrier is sufficiently separated from its predecessor. Surely, one can simply be checking this condition dynamically by keeping track of the distance from the most recently detected node while suppressing any "premature" detection. Instead, we propose to achieve node "anti-aliasing" by slightly adjusting the barrier detecting filter and making it more selective. Hence, Lemma 3.2 extends Lemma 3.1 by constructing 'separated' barriers. Besides the aesthetic advantage of "unconditional sliding", the separated barriers are also attractive since their times, including inter-arrival ones, are more "stochastically regular" and their analysis is somehow less complicated. In particular, the separated barriers are essential if one wants to stationarize the Viterbi process V (or the joint processes (X, Y, V), (X, V), (Y, V)) by embedding these semi-infinite processes into suitable doubly-infinite extensions. For the reasons of size limitation, [8] instead of stationarization took an overall shorter approach based on regenerativity in which the separated barriers are not really necessary. At the same time, unlike their purely regenerative counterparts, the stationary versions of V and the joint processes would automatically deliver the limiting measures needed for asymptotic inference such as in the adjusted Viterbi training [8].

In summary, the presented proofs of Lemmas 3.1 and 3.2 give detailed instructions on how to construct a prototype of an efficient online barrier/node detector for any HMM from a very large class. The detector can then immediately be used to parse a virtually infinite observation sequence and to output the piecewise Viterbi alignment.

Note that in the context of pulse amplitude modulated sequence detection, the possibility of online piecewise construction of the Viterbi alignment was already noticed in [11]. In that work, the occurrence of special (0th order) nodes, referred to as 'merge phenomenon', was briefly discussed and possibly observed empirically. At the same time, the authors of [11] remarked that "a merge is a random phenomenon and consequently may be of limited value in practical applications". It is therefore not clear if the authors realized that *almost every* realization of many HMMs used in practice exhibit infinitely many merges. Several other points make us more optimistic than the authors of [11] about the utility of merges. First, our research shows that despite being non-determinisitic, merges follow a certain stochastic pattern, hence can be studied theoretically, and subsequently can be efficiently computed (e.g. the fixed support barrier-based merge detectors discussed above). Moreover, merges make it possible to define the infinite alignments, which have several practical applications, such as the adjusted Viterbi training [13], [8], [14], and possibly more as discussed in §I-B below.

At the same time, the question "how often do merges occur?" is not easy to answer. Indeed, the merge inter-arrival times are, in general, neither independent nor identically

distributed. (Note that these random times are not the same as the renewal times on which regenerativity in [8] is based and which are essentially i.i.d..) On the other hand, in i.i.d. mixture models, which are a special case of HMMs, every observation is a node. The degenerate case of HMMs with the identity transition matrix is the opposite extreme in the sense that nodes cannot occur in those models at all. Based on these two extremes, one can loosely argue that the weaker the dependence on the past, the more often the nodes occur. It is also discussed in §4 of [8] how the expected barrier inter-arrival time can be reasonably bounded from above. This, obviously, also gives an upper bound on the expected merge inter-arrival time. Since typically there can be several merges between two consecutive barriers, the obtained bound might be rather crude. At the same time, it is the nodes sitting inside the barriers that we suggest to be of more interest, particularly for applications, and their probability can be explicitly computed based on the constructions below. However, without further optimization, that probability generally might be very low (cf. §IV-B for a numerical example).

B. Further motivation

The main motivation of this work is the frequent use of the Viterbi alignment $v(x_{1:T})$ for prediction of the hidden realization $Y_{1:T}$, also known in this context as *segmentation*. Examples of segmentation include HMM-based speech recognition as well as DNA sequencing [16] (e.g. segmenting coding regions from non-coding ones, or detecting CpG-islands) and thus substantially depart from Viterbi's original framework of convolutional coding-decoding [5], [17]. As already pointed out above, the Viterbi alignment is also at the core of the Viterbi training algorithm that iteratively, and concurrently with segmentation, estimates the unknown parameters of the HMM.

Since the Viterbi alignment can deviate significantly from the true hidden sequence, the Viterbi alignment is hardly representative of a typical realization. Hence, using the Viterbi alignment in further inference is conceptually problematic. Indeed, when estimating, say, the probability of heads from i.i.d. tosses of a biased coin, we naturally hope to observe a typical realization and not the constant one of maximum probability.

However, in models where the deviations between the maximum likelihood (Viterbi-like) alignment and the true sequence follow a distinct pattern, at least as $T \to \infty$, further inference based on the alignment can be sometimes made more accurate with little computational overhead. This has been the rationale behind the adjusted Viterbi training and can possibly be extended to other, i.e. non-Viterbi, types of alignments (see below). If known — possibly estimated — these adjustments might also be appreciated when the Viterbi paths are used merely for prediction, or segmentation, of $Y_{1:T}$. Indeed, in segmentation of DNA sequences, the underlying chain Y has few, often two, states (e.g. coding and non-coding regions, or CpG islands and non-CpG regions) and the probabilities of transitions between the states are very low. Therefore, the true and predicted hidden paths tend to consist of long constant blocks. At the same time, the predicted constant blocks tend to be somewhat longer than what the chain parameters would suggest. With the help of the infinite Viterbi process $V_{1:\infty}$ it is now clear that this discrepancy is not simply due to the random fluctuations but is systematic, does not vanish asymptotically, and is a direct consequence of that the transition probabilities of $V_{1:\infty}$ do indeed tend to underestimate the true ones. Note that in these examples, unlike in the estimation of the HMM emission parameters, the overall performance is directly linked to the accuracy of the transition probability estimates. Thus, finding the differences between the processes (X, Y) and (X, V) in this case might help find better alignments.

The Viterbi alignment process V also makes it possible to define the risk of (the Viterbi alignment based) segmentation. Specifically, let $\mathcal{L} : S \times S \to \mathbb{R}^+$ be a loss function, i.e. $\mathcal{L}(i, j)$ measures the loss of classifying true state i as j. Perhaps the most common loss-function is the symmetric one given below

$$\mathcal{L}(i,j) = \begin{cases} 0, & \text{if } i = j; \\ 1, & \text{if } i \neq j. \end{cases}$$

Given a loss function $\mathcal{L}$, quality of segmentation $v(x_{1:T})$ can be naturally measured by the empirical risks given below

$$\begin{aligned} R_T(x_{1:T}, y_{1:T}) &\stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}\left(y_t, v\left(x_{1:T}\right)_t\right), \\ R(x_{1:\infty}, y_{1:\infty}) &\stackrel{\text{def}}{=} \limsup_{T\to\infty} R_T(x_{1:T}, y_{1:T}). \end{aligned} \tag{2}$$

Since $y_{1:T}$ is usually hidden (unsupervised learning), the empirical risk R_T cannot be calculated. However, provided that regenerative $v(X_{1:\infty})$ exists, it is not hard to show that R_T converges *almost surely* to R (i.e. $\limsup$ in (2) can be replaced by $\lim$ for *almost every* realization) where R is (*almost surely*) a constant. The limit R will be called the asymptotic risk of the Viterbi segmentation. Thus, for T large, $R_T(X_{1:T}, Y_{1:T}) \approx R$. (Note the difference with the usual machine learning framework where the empirical risk is used to approximate the unknown "true" risk R. Here, on the contrary, the would-be known R is being used to approximate the unknown empirical risk R_T.) If $\mathcal{L}$ is symmetric, then R_T is simply the proportion of the misclassified states and R is the asymptotic misclassification rate. It should be noted that R_T is not minimum over all possible segmentations. Indeed, given $x_{1:T}$, the empirical risk R_T is minimized by the segmentation $s(x_{1:T}) \in S^T$ given for each $t = 1, 2, \ldots, T$ below:

$$s(x_{1:T})_t \stackrel{\text{def}}{=} \arg\min_{j\in S} \sum_{i\in S} \mathcal{L}(i,j) P(Y_t = i | X_{1:T} = x_{1:T}).$$

Therefore, for the symmetric loss

$$s(x_{1:T})_t = \arg\max_{j\in S} P(Y_t = j | X_{1:T} = x_{1:T}),$$

hence the segmentation $s(x_{1:T})$ returns an individually most likely state (cf. [4] and also 'optimal symbol-by-symbol detection' in [11]). We will call the resulting alignment *pointwise MAP (PMAP)*. Since the PMAP alignment appears to be the main alternative to the Viterbi alignment, it would be interesting to know if an infinite PMAP alignment can be defined and what asymptotic properties it would have. On

one hand, the situation is similar to the Viterbi case since the two algorithms share the forward-backward feature (the main difference is that maximization used in the Viterbi algorithm is now replaced by summation). Consequently, the concept of nodes extends immediately to the PMAP case. In fact, the condition (18) of [11] defines a PMAP node. A node x_τ (in the PMAP sense) fixes the PMAP alignment at time τ, i.e. fixes a single optimal symbol v_τ. However, since the dynamic programming optimality no longer holds in the PMAP case, this need not fix the PMAP alignment before τ. In fact, it is not hard to find counterexamples where observations adjoined after a node change the PMAP alignment before the node. Thus, in the language of [11], in symbol-by-symbol detection, the merge phenomenon appears to be of less value than in the case of optimum sequence detection.

The existence of a well-defined infinite PMAP alignment can be immediately established with the help of martingale convergence theory. Indeed, the smoothing probabilities $P(Y_t = i|X_{1:T})$, $i \in S$, converge as $T \to \infty$ [18]. However, the existence of an infinite PMAP alignment with additional properties such as regenerativity or stationarity, is, to the best of our knowledge, an open question. Using the barriers as in the present paper, one can possibly show that *almost every* realization has infinitely many PMAP nodes. However, since the nodes do not anymore fix the alignment, they cannot be used for constructing a piecewise limiting alignment. Nonetheless, if $s_{1:\infty}$ is an infinite PMAP alignment with suitable properties, then the best possible empirical risk

$$R^*_T(x_{1:T}, y_{1:T}) \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(y_t, s_t) \le R_T(x_{1:T}, y_{1:T})$$

can possibly be also shown to converge (*almost surely*) as $T \to \infty$ to some (constant) limit $R^* \le R$, a generalization of the Bayesian risk. Despite being inferior to the pointwise MAP alignment in terms of risk based on the symmetric loss, segmentation by the Viterbi alignment still seems to be popular in practice. A commonly cited justification is that in certain models with forbidden transitions the PMAP alignment may in principle turn out to be also forbidden (i.e. of zero probability) [4].

Interestingly, for alignments that admit infinite regenerative extensions, such as the Viterbi alignments, it may be possible to assess quality of their segmentation as function of observable $x \in \mathcal{X}$ using, for example, regenerativity of the three-dimensional process (X, Y, V). Indeed, assume for the moment that $y_{1:T}$ is known and split the sample $x_{1:T}$ into two subsamples as follows: observation x_t belongs to subsample "correct" if $v_t = y_t$, and otherwise it belongs to the subsample "incorrect". The subsamples give rise to the empirical measures $\hat{P}_T^{\text{correct}}$ and $\hat{P}_T^{\text{incorrect}}$, i.e.:

$$\hat{P}_T^{\text{correct}}(A) \stackrel{\text{def}}{=} \frac{\sum_{t=1}^{T} \mathbb{I}_{A\times\{0\}}\left(x_t, \mathcal{L}(v_t, y_t)\right)}{\sum_{t=1}^{T} \mathbb{I}_{\{0\}}\left(\mathcal{L}(v_t, y_t)\right)} \quad \forall A \in \mathcal{B},$$

where $\mathcal{L}$ is the symmetric loss function and $\mathbb{I}_A$ is the indicator function of set A. Again, in practice these measures would be unknown, but using regenerativity, it is possible to show that there exist certain limiting probability measures $P^{\text{(in)correct}}$ such that

$$\hat{P}_T^{\text{(in)correct}} \underset{T\to\infty}{\Rightarrow} P^{\text{(in)correct}} \textit{ almost surely},$$

where '$\Rightarrow$' refers to weak convergence of probability measures, e.g. for every (Borel) set A

$$\frac{\sum_{t=1}^{T} \mathbb{I}_{A\times\{0\}}\left(X_t, \mathcal{L}(V_t, Y_t)\right)}{\sum_{t=1}^{T} \mathbb{I}_{\{0\}}\left(\mathcal{L}(V_t, Y_t)\right)} \underset{T\to\infty}{\to} P^{\text{correct}}(A) \textit{ almost surely}.$$

(For a similar proof, see, for example, the proof of Theorem 4.1 in [8].) Thus, knowing the model, the measures $P^{\text{(in)correct}}$ can in principle be found. Note that, with $P = \sum_{i\in S} \pi_i P_i$ and R being the law of X_t and the risk based on the symmetric loss, respectively, it holds that

$$P = P^{\text{correct}}(1 - R) + P^{\text{incorrect}} R.$$

Denoting by $f^{\text{(in)correct}}$ the densities of $P^{\text{(in)correct}}$ with respect to λ, it follows that the ratio

$$P(\text{incorrect}|x) \stackrel{\text{def}}{=} \frac{f^{\text{incorrect}}(x)R}{f^{\text{incorrect}}(x)R + f^{\text{correct}}(x)(1-R)}$$

can be interpreted as the probability that the segmentation at a given time is incorrect, given that x is observed at that time. Again, $P(\text{incorrect}|x)$ may not be easy to find analytically, but, knowing the model, it can be estimated off-line for a sufficiently dense mesh of $\mathcal{X}$ by simulations. A possible use of $P(\text{incorrect}|x)$ in practice can be in a flexible, or "active", semi-supervised learning regime in which for some observations their hidden states can be also revealed, say, at a very high cost (e.g. expert annotation of genetic sequences). It then makes sense to consult the *oracle* for those observations x_t for which $P(\text{incorrect}|x_t)$ is relatively high. For a trivial example, in the case of i.i.d. mixture models, this would happen when x_t falls in regions of significant overlaps of the densities of competing states. With the "purchased" additional information, the constrained Viterbi alignment can be obtained which, in general, would lower the empirical risk considerably. Thus, a trade-off between the cost of the revealed states and alignment risk is conceivable. Finally, the function $P(\text{incorrect}|x)$ provides but one example for such type of a learning scenario, and other, e.g. entropy-based functions, are also conceivable.

Recall also that in Viterbi's original context of convolutional coding-decoding, decoders in practice would often force suboptimal alignments after a certain fixed delay T which is proportionate to the encoder memory [11], [17, §11.4]. Thus, one could consider the following block-stationary process $(v(x_{1:T}), v(x_{T+1:2T}), \ldots)$ or its modification, which can be made regenerative, and in which $v(x_{nT+1:(n+1)T})$, $n = 1, 2, \ldots$, are the conditional alignments given that $Y_{nT} = v(x_{(n-1)T+1:nT})_T$. Clearly, limiting (as $n \to \infty$) characteristics of these processes, such as risk R, or its modification with $Y_{1:\infty}$ replaced by the Viterbi alignment $v(X_{1:\infty})$, will depend on T. It might then be interesting to examine this dependence in light of the practically employed values of T between four and six multiples of the encoder memory [17, §11.4].

C. Organization of the rest of the paper

In §II we briefly outline the construction of the infinite Viterbi alignments (cf. §II-B) based on the nodes (cf. §II-A) and barriers (cf. §II-C) that were introduced in [8]. Next, §III states Lemmas 3.1 and 3.2, our main results. In §IV, we present a complete and detailed proof of the main results. In particular, Lemma 3.1 is proved in §IV-A, followed in §IV-B by an illustration of the central construction with a concrete numerical example; §IV-C is the proof of Lemma 3.2. We conclude in §V by explaining the technical assumptions of the main results and indicating further generalizations.

II. Construction

A. Nodes

Let an observable realization $x_{1:\infty} \in \mathcal{X}^\infty$ be given. First, for all $j \in S$ and for any time $t \geq 1$, consider the *scores*

$$\delta_t(j) \stackrel{\text{def}}{=} \max_{y'_{1:t-1} \in S^{t-1}} \Lambda\big((y'_{1:t-1}, j); x_{1:t}\big) \quad \text{and} \tag{3}$$

the *back-pointers*

$$l(t,j) \stackrel{\text{def}}{=} \{l \in S : \forall i \in S \ \delta_t(l)p_{lj} \geq \delta_t(i)p_{ij}\}. \tag{4}$$

Thus, $\delta_t(j)$ is the maximum of the likelihood of the paths terminating at time t in state j. Note that $\delta_1(j) = \pi_j f_j(x_1)$ and the recursion below

$$\delta_{t+1}(j) = \max_{i \in S}(\delta_t(i)p_{ij})f_j(x_{t+1}) \quad \forall\ t \geq 1, \forall j \in S,$$

helps to verify that for any $T \geq 1$, $\mathcal{V}(x_{1:T})$, the set of all the Viterbi alignments, can be written as follows: $\mathcal{V}(x_{1:T}) = \big\{v \in S^T :\ \forall i \in S, \delta_T(v_T) \geq \delta_T(i)$ and
$\forall t :\ 1 \leq t < T,\ v_t \in l(t, v_{t+1})\big\}$.

Next, we introduce $p_{ij}^{(r)}(t)$, the maximum of the likelihood realized along the paths connecting states i and j at times t and $t+r+1$, respectively. Thus, $p_{ij}^{(0)}(t) \stackrel{\text{def}}{=} p_{ij}$, and for all $t \geq 1$ and for all $r \geq 1$, let $p_{ij}^{(r)}(t) \stackrel{\text{def}}{=} \max_{y'_{1:r} \in S^r} p_{iy'_1} f_{y'_1}(x_{t+1}) p_{y'_1 y'_2} f_{y'_2}(x_{t+2}) p_{y'_2 y'_3} \cdots$

$$\cdots p_{y'_{r-1} y'_r} f_{y'_r}(x_{t+r}) p_{y'_r j}. \tag{5}$$

Note also that

$$\begin{aligned} \delta_{t+1}(j) &= \max_{i \in S}\big\{\delta_{t-r}(i)p_{ij}^{(r)}(t-r)\big\}f_j(x_{t+1}) \quad \forall r < t, \\ p_{ij}^{(r)}(t) &= \max_{l \in S} p_{il}^{(r-1)}(t) f_l(x_{t+r}) p_{lj}. \end{aligned} \tag{6}$$

Definition 2.1: Let $\rho \geq 0$ and $\tau \geq 1$ be integers, and let $l \in S$. Given $x_{1:\infty} \in \mathcal{X}^\infty$, x_τ is said to be an *l-node of order ρ* if

$$\delta_\tau(l)p_{lj}^{(r)}(\tau) \geq \delta_\tau(i)p_{ij}^{(r)}(\tau) \quad \forall i, j \in S. \tag{7}$$

Also, x_τ is said to be a node of order ρ if it is an l-node of order ρ for some $l \in S$; x_τ is said to be a strong node of order ρ if the inequalities in (7) are strict for every $i, j \in S, i \neq l$. Certainly, only the starting subsequence $x_{1:\tau+\rho}$ of $x_{1:\infty}$ determines whether x_τ is a node of order ρ. [1] Let $x_{1:\infty}$ be such that x_{τ_n} is an l_n-node of order ρ_n, $1 \leq n \leq N$, for some, possibly infinite, N, and assume that $\tau_{n+1} > \tau_n + \rho_n$ for all $n = 1, 2, \ldots, N-1$. Such nodes are said to be *separated*.

[1] Note that if x_τ is a node of order ρ, it is then also a node of any order higher than ρ. Hence, the order of a node is defined to be the minimum such ρ.

B. Piecewise alignment

Suppose $x_{1:T}$ is indeed such that $x_{\tau_1}, x_{\tau_2}, \ldots x_{\tau_N}$ are separated nodes of type $l \in S$ and order $\rho \geq 0$. It follows then easily from the definition of the node that there exists a Viterbi alignment $v(x_{1:T}) \in \mathcal{V}(x_{1:T})$ that goes through l at τ_n (i.e. $v_{\tau_n} = l$) for each $n = 1, 2, \ldots, N$ (see [8]). It is not difficult to verify that such $v(x_{1:T})$ can actually be computed as follows: Obtain v^1, a path that is optimal among all those that end at τ_1 in l. (Note that unless the order of the node x_{τ_1} is 0, v^1 need not be in $\mathcal{V}(x_{1:\tau_1})$.) Given $x_{\tau_1+1:\tau_2}$, continue on by taking v^2 to be a maximum likelihood path from l back to l. That is, v^2 maximizes the constrained likelihood under the initial distribution $(p_{l\cdot}) \stackrel{\text{def}}{=} (p_{lj})_{j \in S}$ (instead of the stationary π) and the constraint $v^2_{\tau_2 - \tau_1} = l$. Now, (v^1, v^2) maximizes the likelihood given $x_{1:\tau_2}$ over all paths ending with l. Similarly, we define the pieces $v^3, \ldots, v^N$. Finally, v^{N+1} is chosen to maximize the (unconstrained) likelihood given $x_{\tau_{N+1}:T}$ under the initial distribution $(p_{l\cdot})$.

Obviously, the separated nodes assumption $\tau_{n+1} > \tau_n + \rho$, $1 \leq n < N$, is not restrictive at all since it is always possible to choose an infinite subsequence of separated nodes from any infinite sequence of nodes. The reason for including this requirement has to do with the non-uniqueness of alignments and is as follows. The fact that x_{τ_n} is an l-node of order ρ guarantees that when backtracking from $\tau_n + \rho$ down to τ_n, ties (if any) can be broken in such a way that, regardless of the values of $x_{\tau_n+\rho+1:T}$ and how ties are broken in between T and $\tau_n + \rho$, the alignment goes through l at τ_n. At the same time, segment $\tau_n, \ldots, \tau_n + \rho$ is "delicate", that is, unless x_{τ_n} is a strong node, breaking the ties arbitrarily within $\tau_n, \ldots, \tau_n + \rho$ can result in $v_{\tau_n} \neq l$. Hence, when neither x_{τ_n} nor $x_{\tau_{n+1}}$ is strong and $\tau_{n+1} \leq \tau_n + \rho$, breaking the ties in favor of x_{τ_n} can result in $v_{\tau_{n+1}} \neq l$. Clearly, such a pathological situation is impossible if $\rho = 0$ and might also be rare in practice even for $\rho > 0$.

To formalize the piecewise construction, let

$$\begin{aligned} \mathcal{W}^l(x_{1:T}) \stackrel{\text{def}}{=} \{&y'_{1:T} \in S^T :\ y'_T = l \\ &\Lambda(y'_{1:T}; x_{1:T}) \geq \Lambda(y''_{1:T}; x_{1:T}) \quad \forall y''_{1:T} \in S^T : y''_T = l\}, \end{aligned}$$

$\mathcal{V}^l(x_{1:T}) \stackrel{\text{def}}{=} \{y'_{1:T} \in \mathcal{V}(x_{1:T}) : y'_T = l\}$ be the set of maximizers of the constrained likelihood, and the subset of maximizers of the (unconstrained) likelihood, respectively, all elements of which go through l at T. Note that unlike $\mathcal{W}^l(x_{1:T})$, $\mathcal{V}^l(x_{1:T})$ might be empty. It can be shown that $\mathcal{V}^l(x_{1:T}) \neq \emptyset \Rightarrow \mathcal{V}^l(x_{1:T}) = \mathcal{W}^l(x_{1:T})$. Also, let the subscript (i), $i \in S$, in $\mathcal{W}^l_{(i)}(x_{1:T})$ and $\mathcal{V}_{(i)}(x_{1:T})$ refer to the $(p_{i\cdot})$ being used as the initial distribution in place of π. With these notations, the piecewise alignment is $v(x_{1:T}) = (v^1(x_{1:\tau_1}), \ldots, v^{N+1}(x_{\tau_N+1:T})) \in \mathcal{V}(x_{1:T})$, where

$$\begin{aligned} &v^1(x_{1:\tau_1}) \in \mathcal{W}^l(x_{1:\tau_1}),\ v^{N+1}(x_{\tau_N+1:T}) \in \mathcal{V}_{(l)}(x_{\tau_N+1:T}) \\ &v^i(x_{\tau_{n-1}+1:\tau_n}) \in \mathcal{W}^l_{(l)}(x_{\tau_{n-1}+1:\tau_n}),\ 2 \leq n \leq N. \end{aligned} \tag{8}$$

Moreover, for $n = 1, 2, \ldots, N$, the partial paths $w(n) \stackrel{\text{def}}{=} (v^1(x_{1:\tau_1}), \ldots, v^n(x_{\tau_{n-1}+1:\tau_n})) \in \mathcal{W}^l(x_{1:\tau_n})$, where $\tau_0 \stackrel{\text{def}}{=} 0$.

If $x_{1:\infty}$ has infinitely many (separated) nodes $\{x_{\tau_n}\}_{n\geq 1}$ then $v(x_{1:\infty})$, an *infinite piecewise alignment based on the node times* $\{\tau_n(x_{1:\infty})\}_{n\geq 1}$ can be defined as follows: If the sets $\mathcal{W}^l_{(l)}(x_{\tau_{n-1}+1:\tau_n})$, $n = 2, 3 \ldots$, as well as $\mathcal{W}^l(x_{1:\tau_1})$ are singletons, then (8) immediately defines a unique infinite alignment $v(x_{1:\infty}) = (v^1(x_{1:\tau_1}), v^2(x_{\tau_1+1:\tau_2}), \ldots)$. Otherwise, ties must be broken. If we want our infinite alignment process V to be regenerative (see [8]), a natural consistency condition must be imposed on rules to select unique $v(x_{1:T})$ from $\mathcal{W}^l(x_{1:\tau_1}) \times \mathcal{W}^l_{(l)}(x_{\tau_1+1:\tau_2}) \times \cdots \times \mathcal{W}^l_{(l)}(x_{\tau_{N-1}+1:\tau_N}) \times \mathcal{V}_{(l)}(x_{\tau_N+1:T})$. In [8], resulting infinite alignments, as well as decoding $v : \mathcal{X}^\infty \to S^\infty$ based on such alignments, are called *proper*. This condition is, perhaps, best understood by the following example. Suppose for some $x_{1:5} \in \mathcal{X}^5$, $\mathcal{W}^1_{(2)}(x_{1:5}) = \{12211, 11211\}$, and suppose the tie is broken in favor of 11211. Now, whenever $\mathcal{W}^1_{(l)}(x'_{1:4})$ contains $\{1221, 1121\}$, we naturally require that 1221 not be selected. In particular, if for some $x'_{1:4} \in \mathcal{X}^4$ $\mathcal{W}^1_{(1)}(x'_{1:4}) = \{1221, 1121\}$, we would select 1121 from $\mathcal{W}^1_{(1)}(x'_{1:4})$. Subsequently, 112 would be selected from $\mathcal{W}^2_{(1)}(x''_{1:3}) = \{122, 112\}$, and so on. *It can be shown that a decoding by piecewise alignment* (8) *with ties broken in favor of min (or max) under the reverse lexicographic ordering of S^T, $T \in \mathbb{N}$, is a proper decoding.*

Note also that we break ties locally, i.e. within individual intervals $\tau_{n-1}+1, \ldots, \tau_n$, $n \geq 2$, enclosed by adjacent nodes. This is in contrast to global ordering of $\mathcal{V}(x_{1:T})$, an approach taken in [9], [10]. Since a global order need not respect the decomposition (8), it can fail to produce an infinite alignment going through infinitely many nodes unless the nodes are strong.

C. Barriers

Recall (Definition 2.1) that nodes are defined *relative* to the entire realization $x_{1:\infty}$ and, what is actually inconvenient is that whether x_τ is a node (of order ρ) or not, depends, in principle, on the entire sequence $x_{1:\tau+\rho}$, and in particular on all the observations up to the time τ.

We show below that typically a block $z_{1:M} \in \mathcal{X}^M$ ($M > \rho \geq 0$) can be found such that for any realization $x_{1:\infty}$ containing $z_{1:M}$, i.e. $x_{t-M+1:t} = z_{1:M}$ for some $t \geq M$, $x_{t-\rho}$ is a node of order ρ. Sequences $z_{1:M}$ that ensure existence of such persistent nodes are called *barriers* in [8]. Specifically,

Definition 2.2: Given $l \in S$, $z_{1:M} \in \mathcal{X}^M$ is called an (strong) *l-barrier* of order $\rho \geq 0$ and length $M > \rho$, if, for any $x_{1:\infty} \in \mathcal{X}^\infty$ with $x_{t-M+1:t} = z_{1:M}$ for some $t \geq M$, $x_{t-\rho}$ is an (strong) l-node of order ρ.

III. Existence

A. Clusters and main results

For each $i \in S$, let

$$G_i \stackrel{\text{def}}{=} \{x \in \mathcal{X} : f_i(x) > 0\}$$

be the support of f_i.

Definition 3.1: We call a nonempty subset $C \subset S$ *a cluster* if the following conditions are satisfied:

$$\min_{j\in C} P_j(\cap_{i\in C} G_i) > 0, \qquad \text{and}$$
$$\text{either } C = S \quad \text{or } \max_{j\notin C} P_j(\cap_{i\in C} G_i) = 0.$$

Hence, a cluster is a maximal subset of states such that $G_C \stackrel{\text{def}}{=} \cap_{i\in C} G_i$, the intersection of the supports of the corresponding emission densities, is 'detectable'. Distinct clusters need not be disjoint and a cluster can consist of a single state. In this latter case such a state is not hidden, since it is exposed by any observation it emits. When $K = 2$, S is the only cluster possible, since otherwise all observations would expose their states and the underlying Markov chain would cease to be hidden. In practice, many HMMs have the entirety of S as their (necessarily unique) cluster.

Before stating the main results, let us define for every state $j \in S$

$$p^*_j = \max_{i\in S} p_{ij}. \tag{9}$$

Lemma 3.1: Assume that for each state $j \in S$,

$$P_j\left(\left\{x \in \mathcal{X} : \ f_j(x)p^*_j > \max_{i\in S,\ i\neq j} f_i(x)p^*_i\right\}\right) > 0. \tag{10}$$

Moreover, assume that there exists a cluster $C \subset S$ and a positive integer m such that the mth power of the substochastic matrix $\mathbb{Q} = (p_{ij})_{i,j\in C}$ is strictly positive. Then, for some integers M and ρ, $M > \rho \geq 0$, there exist a set $B = B_1 \times \cdots \times B_M \subset \mathcal{X}^M$, an M-tuple of states $y_{1:M} \in S^M$ and a state $l \in S$, such that every $z_{1:M} \in B$ is an l-barrier of order ρ (and length M), $y_{M-\rho} = l$ and

$$\mathbf{P}\left(X_{1:M} \in B, \quad Y_{1:M} = y_{1:M}\right) > 0.$$

Lemma 3.1 implies that $\mathbf{P}(X_{1:M} \in B) > 0$. Also, since every element of B is a barrier of order ρ, the ergodicity of X therefore guarantees that *almost every realization of X contains infinitely many l-barriers of order ρ. Hence, almost every realization of X also has infinitely many l-nodes of order ρ.*

Note that since in two-state HMMs S is the only cluster, we thus have that $\mathbb{Q} = \mathbb{P}$. The irreducibility and aperiodicity in this case imply strict positivity of $\mathbb{P}^2$. Thus, the only condition to be verified is (10), which in this case writes as $P_1\left(\{x \in \mathcal{X} : \ f_1(x)p^*_1 > f_2(x)p^*_2\}\right) > 0$ and $P_2\left(\{x \in \mathcal{X} : \ f_2(x)p^*_2 > f_1(x)p^*_1\}\right) > 0$. In [15], it is shown that in the case of two-state HMMs, one of these two positivity conditions is always met, which, in fact, turns out to be sufficient for the existence of infinitely many strong barriers in this ($K = 2$) case. Thus, *any two-state HMM with irreducible and aperiodic Y has infinitely many strong barriers.* We return to the discussion of the hypotheses of Lemma 3.1 in §V.

Recall (§I-A) that in order to make our barrier detectors useful for online construction of the piecewise alignments that can be made stationary, we need to make those detectors selective, i.e. to insure that they do not respond to nodes which are not sufficiently separated. Thus, instead of simply asserting that any infinite sequence of nodes (such as those guaranteed

by Lemma 3.1) contains an infinite subsequences of separated nodes, we achieve node separation by adjusting the notion of barriers. Namely, note that (given a realization $x_{1:\infty}$) two l-barriers $x_{t-M+1:t}$ and $x_{u-M+1:u}$ of order ρ might be in B with $t < u \leq t+\rho$, implying that the associated nodes $x_{t-\rho}$ and $x_{u-\rho}$ are not separated. In order to prevent this, we require B to be such that for any $x_{1:\infty} \in \mathcal{X}^\infty$ it holds that

$$x_{t-M+1:t}, x_{u-M+1:u} \in B \text{ for some } t,u \geq M,\ t \neq u \quad (11)$$
$$\Rightarrow |u-t| > \rho,$$

where $x_{t-M+1:t}$ and $x_{u-M+1:u}$ are l-barriers of order ρ (and length M). If (11) holds, we say that the barriers in $B \subset \mathcal{X}^M$ are *separated*. This is often easy to achieve by a simple extension of B as shown in the following example. Suppose there exists $x \in \mathcal{X}$ such that $x \notin B_m$, for all $m = 1, 2, \ldots, M$. All elements of $B^* \stackrel{\text{def}}{=} \{x\} \times B$ are evidently barriers, and moreover, they are now separated. The following Lemma incorporates a more general version of the above example.

Lemma 3.2: Suppose the assumptions of Lemma 3.1 are satisfied. Then, for some integers M and ρ, $M > \rho \geq 0$, there exist $B = B_1 \times \cdots \times B_M \subset \mathcal{X}^M$, $y_{1:M} \in S^M$, and $l \in S$, such that every $z_{1:M} \in B$ is a separated l-barrier of order ρ (and length M), $y_{M-\rho} = l$, and $\mathbf{P}\left(X_{1:M} \in B, \quad Y_{1:M} = y_{1:M}\right) > 0$.

IV. Proof of the main results

A. Proof of Lemma 3.1

The proof below is a rather direct construction which is, however, technically involved. A key ingredient of the proof is a prototype of $y_{1:M}$, termed the "s-path". The central idea of the Lemma and its proof is then to exhibit (a cylinder subset of the) 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 middle of the s-path is a node.

In order to facilitate the exposition of this proof, we have divided it into 17 short parts as outlined below:

I. Construction of
- (§IV-A1) emission subsets $\mathcal{X}_i \in \mathcal{X}$, $i \in S$ (12);
- (§IV-A2) a special set $\mathcal{Z} \subset \mathcal{X}$, (14), (15);
- (§IV-A3) auxiliary sequences $\mathbf{s}$, $\mathbf{a}$, and $\mathbf{b}$ of states in S;
- (§IV-A4) k, the number of $\mathbf{s}$ cycles inside the s-path;
- (§IV-A5) the s-path (21), the core of $y_{1:M}$;
- (§IV-A6) the required barrier (22).

II. Proving the barrier construction (22):
- (§IV-A7) $\alpha, \beta, \gamma, \eta, \phi, \psi$-notation for commonly used maximal partial likelihoods;
- (§IV-A8) a bound (27) on β;
- (§IV-A9) bounds (28), (29), (30), and (31) on common likelihood ratios;
- (§IV-A10) $\gamma_j \leq const \times \gamma_1$;
- (§IV-A11) further bounds (47), (48) on likelihoods;
- (§IV-A12) $\eta_j \leq const \times \eta_1$;
- (§IV-A13) a special representation of η_1 (50);
- (§IV-A14) an implication of (46) and (50): $\forall N \in \{k, k+1, \ldots, 2k\}$, $\exists$ a realization of $\phi_1(e_{NL})$ that goes through state 1 $\forall n \in \{k, k+1, \ldots, N\}$;
- (§IV-A15) inequality (57) implies that $x_{u-\rho}$ is a 1-node of order $\rho = kL + m + P$;
- (§IV-A16) proof of the inequality (57);

III. (§IV-A17) Completion of the s-path to $y_{1:M}$.

1) $\mathcal{X}_i \subset \mathcal{X}$, $i \in S$: It follows from the assumption (10) and finiteness of S that there exists an $\epsilon > 0$ such that for all $i \in S$ $P_i(\mathcal{X}_i) > 0$, where

$$\mathcal{X}_i \stackrel{\text{def}}{=} \Big\{ x \in \mathcal{X} : \max_{j \in S,\ j \neq i} p_j^* f_j(x) < (1-\epsilon) p_i^* f_i(x) \Big\}. \quad (12)$$

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

2) $\mathcal{Z} \subset \mathcal{X}$ and $\delta - \Delta$ bounds on cluster densities f_i, $i \in C$: Let C be a cluster as in the assumptions of the Lemma. The existence of C implies the existence of a set $\hat{\mathcal{Z}} \subset \cap_{i \in C} G_i$ 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, $\min_{j \in C} P_j(\cap_{i \in C} G_i) > 0$ implies (and indeed is equivalent to) $\lambda(\cap_{i \in C} G_i) > 0$. The latter implies the existence of $\hat{\mathcal{Z}} \subset \cap_{i \in C} G_i$ with positive λ-measure and $\delta > 0$ such that (i) holds. Since $\lambda(\cap_{i \in C} G_i) > 0$, the condition $P_j(\cap_{i \in C} G_i) = 0$ for $j \notin C$ implies (is equivalent to) $f_j = 0$ λ-*almost everywhere* on $\cap_{i \in C} G_i$. Thus, $\max_{j \notin C} f_j = 0$ λ-*almost everywhere* on $\cap_{i \in C} G_i$, which implies (ii) for any $z \in \hat{\mathcal{Z}}$.

Evidently, $\Delta > \delta$ can be chosen sufficiently large to make $\lambda(\{z \in \mathcal{X} :\ f_i(z) \geq \Delta\})$ arbitrarily small, and in particular, to guarantee that for all $i \in C$ $\lambda(\{z \in \mathcal{X} :\ f_i(z) \geq \Delta\}) < \frac{\lambda(\hat{\mathcal{Z}})}{|C|}$, where $|C|$ is the size of C. Clearly then, redefining $\hat{\mathcal{Z}} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \cap \{z \in \mathcal{X} :\ f_i(z) < \Delta,\ \forall\ i \in C\}$ preserves $\lambda(\hat{\mathcal{Z}}) > 0$. Thus, in summary, $\hat{\mathcal{Z}}$ is a subset of the emission space $\mathcal{X}$ such that for all $z \in \hat{\mathcal{Z}}$ $\delta < f_i(z) < \Delta$ for all $i \in C$ and $f_i(z) = 0$ for all $i \notin C$.

The next modification of $\hat{\mathcal{Z}}$ will actually be needed in §IV-C, the proof of Lemma 3.2. Consider

$$\lambda(\hat{\mathcal{Z}} \backslash (\cup_{i \in S} \mathcal{X}_i)). \quad (13)$$

If (13) is positive, then define

$$\mathcal{Z} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \backslash (\cup_{i \in S} \mathcal{X}_i). \quad (14)$$

If (13) 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

$$\mathcal{Z} \stackrel{\text{def}}{=} \hat{\mathcal{Z}} \cap \mathcal{X}_s. \quad (15)$$

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

3) Sequences $\mathbf{s}$*,* $\mathbf{a}$*, and* $\mathbf{b}$ *of states in* S*:* Let us define an auxiliary sequence of states h_1, h_2, and so on, as follows: If (13) is zero, that is, if $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$ for some $s \in C$, then define $h_1 = s$, otherwise let h_1 be an arbitrary state in C. Let h_2 be a state with maximal probability of transition to h_1, i.e.: $p_{h_2\,h_1} = p^*_{h_1}$ Suppose $h_2 \neq h_1$. Then find h_3 with $p_{h_3\,h_2} = p^*_{h_2}$. If $h_3 \notin \{h_1, h_2\}$, find $h_4 :\ p_{h_4\,h_3} = p^*_{h_3}$, and so on. Let U be the first index such that $h_U \in \{h_1, \ldots, h_{U-1}\}$, that is, $h_U = h_T$ for some $T < U$. This means that there exists a sequence of states $\{h_T, \ldots, h_U\}$ such that

- $h_T = h_U$
- $p_{h_{T+n}\,h_{T+n-1}} = p^*_{h_{T+n-1}}, \quad n = 1, \ldots, U-T.$

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

$$
\begin{aligned}
&s_1 \stackrel{\text{def}}{=} h_{U-1},\ s_2 \stackrel{\text{def}}{=} h_{U-2}, \ldots, s_n \stackrel{\text{def}}{=} h_{U-n}, \ldots,\\
&s_L \stackrel{\text{def}}{=} h_T = 1 \quad n = 1, \ldots, L \stackrel{\text{def}}{=} U - T,\\
&a_1 \stackrel{\text{def}}{=} h_{T-1},\ a_2 \stackrel{\text{def}}{=} h_{T-2}, \ldots, a_P \stackrel{\text{def}}{=} h_1,
\end{aligned}
$$

where $P \stackrel{\text{def}}{=} T - 1$. Hence,

$$
\begin{aligned}
&\{h_1, \ldots, h_{T-1}, h_T, h_{T+1}, \ldots, h_{U-1}, h_U\} =\\
&\{a_P, \ldots, a_1, 1, s_{L-1}, \ldots, s_1, 1\}.
\end{aligned}
$$

Note that if $T = 1$, then $P = 0$ and $\{h_1, \ldots, \ldots, h_{U-1}, h_U\} = \{1, s_{L-1}, \ldots, s_1, 1\}$. 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,

$$
\begin{aligned}
p_{s_{n-1}\,s_n} &= p^*_{s_n},\ n = 2, \ldots, L,\ p^*_{s_1} = p_{1\,s_1}\\
p_{a_{n-1}\,a_n} &= p^*_{a_n},\ n = 2, \ldots, P,\ p^*_{a_1} = s_L = 1.
\end{aligned} \tag{16}
$$

Next, we are going to exhibit $\mathbf{b} = (b_0, b_1, \ldots, b_R)$, another auxiliary sequence for some $R \geq 1$, characterized as follows:

(i) $b_R = 1$;
(ii) $b_0 \in C$ 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 $b_{r-1} \neq b_r$ for every $r = 1, \ldots, R$.

Thus, the path $b_{1:R}$ connects cluster C to state 1 in R steps. Clearly such $\mathbf{b}$ does exist due to irreducibility of Y. Let us also require that the selected path be as short as possible, i.e. $R \leq K - 1$ is the minimal number of steps needed to connect C to 1 in the above sense. Note then that minimality of R guarantees (iii) (in the special case of $R = 1$ it may happen that $b_1 = 1 \in C$ and $p_{1\,1} > 0$, in which case b_0 can be taken to be also 1).

4) Determining k*:* Let $\mathbb{Q}^m$ be the mth 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 hypothesis of the Lemma, $q_{ij} > 0\ \forall i, j \in C$. This means that for every $i, j \in C$, there exists a positive probability path from i to j, which is entirely inside C, and has length m. Let q^*_{ij} be the maximum of the probability of transition from i to j along all such paths. Also, let $c_{1:m-1}(i, j)$, or simply $c_{1:m-1}$ in the absence of ambiguity, be some such maximum probability path, i.e. $c_1, \ldots, c_{m-1} \in C$ are such that

$$
p_{ic_1} p_{c_1 c_2} \cdots p_{c_{m-1} c_{m-1}} p_{c_{m-1} j} = q^*_{ij} > 0. \tag{17}
$$

Let us define

$$
q = \min_{i,j \in C} q^*_{ij} > 0, \quad \text{and} \tag{18}
$$

$$
A = \max_{i \in S} \max_{j \in S} \left\{ \frac{p^*_i}{p_{ji}} : p_{ji} > 0 \right\}, \tag{19}
$$

where p^*_i are as defined in (9). Choose k sufficiently large for the following to hold:

$$
(1 - \epsilon)^{k-1} < q^2 \left(\frac{\delta}{\Delta} \right)^{2m} A^{-R}, \tag{20}
$$

where ϵ is as in (12) and δ and Δ are as introduced in §IV-A2.

5) The s*-path:* We now fix the state sequence

$$
b_0, b_1, \ldots, b_R, s_1, s_2, \ldots, s_{2kL}, a_1, \ldots, a_P, \tag{21}
$$

where $s_{NL+n} = s_n$, $N = 1, \ldots, 2k-1$, $n = 1, \ldots, L$, (and in particular $s_{NL} = 1$, $N = 1, \ldots, 2k$). The sequence in (21) will be called the *s-path*. The s-path is a concatenation of $2k$ $\mathbf{s}$ cycles $s_{1:L}$, the beginning and the end of which are connected to the cluster C via positive probability paths $\mathbf{b}$ and $\mathbf{a}$, respectively (recall that $a_P = h_1 \in C$ and $b_R = 1$ by construction). Additionally, the $b_R, s_1, s_2, \ldots, s_{2kL}, a_1, \ldots, a_P$-segment of the s-path (21) has the important property (16), i.e. every consecutive transition along this segment occurs with the maximal transition probability given its destination state. (However, $\mathbf{b}$, the beginning of the s-path, need not satisfy this property.) The s-path is almost ready to serve as $y_{1:M}$ promised by the Lemma and its completion to $y_{1:M}$ will be acomplished in §IV-A17.

6) The barrier: Let $M = 2m + 2kL + P + R + 1$, and let $\rho = kL + P + m$ and $l = 1$. Let

$$
\begin{aligned}
B &= B_1 \times B_2 \times \cdots \times B_M, \text{ where}\\
B_n &= \mathcal{Z},\ 1 \leq n \leq m+1;\\
B_{m+1+n} &= \mathcal{X}_{b_n},\ 1 \leq n \leq R;\\
B_{m+1+R+n+NL} &= \mathcal{X}_{s_n},\ 1 \leq N < 2k,\ 1 \leq n \leq L;\\
B_{m+1+R+2kL+n} &= \mathcal{X}_{a_n},\ 1 \leq n \leq P;\\
B_{m+1+R+2kL+P+n} &= \mathcal{Z},\ 1 \leq n \leq m,
\end{aligned} \tag{22}
$$

and let $z_{1:M}$ be any sequence from B. We now need to prove that $z_{1:M}$ is a 1-barrier of order ρ as promised by the Lemma. To do this, it might be helpful to give distinct names to some of the key subsequences of $z_{1:M}$ as follows:

$$
z_{1:M} = (z_{1:m+1}, e'_{1:R-1}, e_{0:2kL}, e''_{1:P}, z'_{1:m}),
$$

and also write B more explicitly as follows:

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

Thus,

$$
\begin{aligned}
&z_{m+1}, z_n, z'_n \in \mathcal{Z}, \quad n = 1, \ldots, m;\\
&e'_r \in \mathcal{X}_{b_r}, \quad r = 1, \ldots, R-1;\\
&e_0 \in \mathcal{X}_1, \quad e_{NL+n} \in \mathcal{X}_{s_n},\ N = 1, \ldots, 2k-1,\ n = 1, \ldots, L\\
&e''_n \in \mathcal{X}_{a_n}, \quad n = 1, \ldots, P.
\end{aligned}
$$

Next, let $x_{1:\infty} \in \mathcal{X}^\infty$ be any sequence of observations that contains the subsequence $z_{1:M}$ (22), i.e. $x_{u-M+1:u} = z_{1:M}$ for some $u \geq M$. Henceforth and throughout §IV-A15, we shall be proving that $x_{u-\rho}$, otherwise referred to as e_{kL}, is a 1-node of order $\rho = kL + m + P$, implying that $z_{1:M}$ is a 1-barrier of the same order.

7) α, β, γ, η, ϕ, ψ: Recall the definition of the scores $\delta_t(j)$ (3) and the maximum partial likelihoods $p_{ij}^{(r)}(t)$ (5) for $t = 1, 2, \ldots$, $i, j \in S$, and $r \geq 0$. Now, we need to introduce the following aliases. For any $i, j \in S$ and for any $r \geq 0$, let

$$\begin{aligned}
\phi_i(e_t) &\stackrel{\text{def}}{=} \delta_{u-P-m-2kL+t}(i) \quad \forall t : 0 \leq t \leq 2kL \\
\psi_{ij}^{(r)}(e_t) &\stackrel{\text{def}}{=} p_{ij}^{(r)}(u-P-m-2kL+t), \\
\psi_{ij}^{(r)}(e'_t) &\stackrel{\text{def}}{=} p_{ij}^{(r)}(u-P-m-2kL-R+t) \quad \forall t : \\
&\qquad 1 \leq t \leq R-1, \\
\phi_i(z_t) &\stackrel{\text{def}}{=} \delta_{u-2kL-2m-P-R+t}(i) \quad \forall t : 1 \leq t \leq m+1, \\
\psi_{ij}^{(r)}(z_t) &\stackrel{\text{def}}{=} p_{ij}^{(r)}(u-2kL-2m-P-R+t), \\
\phi_i(z'_t) &\stackrel{\text{def}}{=} \delta_{u-m+t}(i) \quad \forall t : 1 \leq t \leq m, \\
\psi_{ij}^{(r)}(z'_t) &\stackrel{\text{def}}{=} p_{ij}^{(r)}(u-m+t).
\end{aligned} \tag{23}$$

Surely, it is only the index t that is variable in the arguments e_t, e'_t, z_t, and z'_t in the left hand sides of the above aliases, whereas the corresponding elements of $z_{1:M}$ are still fixed. However, we hope that the above redundancy helps to keep track of the individual subsequences $z_{1:m+1}$, $e'_{1:R-1}$, $e_{0:2kL}$, etc.

Also, we will be frequently using the scores corresponding to z_1, z_{m+1}, e_0, and e_{kL}. Hence the following abbreviations:

$$\alpha_i \stackrel{\text{def}}{=} \phi_i(z_1),\ \beta_i \stackrel{\text{def}}{=} \phi_i(z_{m+1}),\ \gamma_i \stackrel{\text{def}}{=} \phi_i(e_0),\ \eta_i \stackrel{\text{def}}{=} \phi_i(e_{kL}).$$

Note that $\forall j \notin C$, $f_j(z_{m+1}) = f_j(z'_t) = f_j(z_t) = 0$, $t = 1, \ldots, m$ by the construction of $\mathcal{Z}$ (§IV-A2). Hence, $\alpha_j = \beta_j = 0$ $\forall j \notin C$, and a more general implication is that for every $j \in S$

$$\begin{aligned}
\beta_j &= \max_{i \in C} \alpha_i \psi_{ij}^{(m-1)}(z_1) f_j(z_{m+1}) \qquad (24) \\
&= \alpha_{i_\beta(j)} \psi_{i_\beta(j)\, j}^{(m-1)}(z_1) f_j(z_{m+1}) \text{ for some } i_\beta(j) \in C; \\
\gamma_j &= \max_{i \in C} \beta_i \psi_{ij}^{(R-1)}(z_{m+1}) f_j(e_0) \qquad (25) \\
&= \beta_{i_\gamma(j)} \psi_{i_\gamma(j)\, j}^{(R-1)}(z_{m+1}) f_j(e_0) \text{ for some } i_\gamma(j) \in C.
\end{aligned}$$

Similarly, we will use the following representation of η_j, $j \in S$, in terms of γ_i for some $i \in S$ which generally depends on j:

$$\begin{aligned}
\eta_j &= \max_{i \in S} \gamma_i \psi_{ij}^{(kL-1)}(e_0) f_j(e_{kL}) \qquad (26) \\
&= \gamma_{i_\eta(j)} \psi_{i_\eta(j)\, j}^{(kL-1)}(e_0) f_j(e_{kL}) \text{ for some } i_\eta(j) \in S.
\end{aligned}$$

8) Bounds on β: Recall (§IV-A3) that $b_0 \in C$. We show that for every $j \in S$

$$\beta_j < q^{-1} \Big(\frac{\Delta}{\delta} \Big)^m \beta_{b_0}. \tag{27}$$

Fix $j \in S$ and consider $\alpha_{i_\beta(j)}$ from (24). Let $h_{1:m-1}$ be a path that realizes $\psi_{i_\beta(j)\, j}^{(m-1)}(z_1)$. Then $\beta_j = \alpha_{i_\beta(j)} p_{i_\beta(j)\, h_1} f_{h_1}(z_2) p_{h_1\, h_2} f_{h_2}(z_3) \cdots \ p_{h_{m-1}\, j} f_j(z_{m+1}) < \alpha_{i_\beta(j)} \Delta^m$. (Recall that Δ was introduced in §IV-A2.) Let $c_{1:m-1}$ be a maximum probability path in the sense of (17) from $i_\beta(j)$ to b_0. Thus,

$$\begin{aligned}
\beta_{b_0} \quad &\geq \alpha_{i_\beta(j)} \psi_{i_\beta(j)\, b_0}^{(m-1)}(z_1) f_{b_0}(z_{m+1}) \\
&\geq \alpha_{i_\beta(j)} p_{i_\beta(j)\, c_1} f_{c_1}(z_2) p_{c_1\, c_2} f_{c_2}(z_3) \cdots \\
&\qquad \cdots p_{c_{m-1}\, b_0} f_{b_0}(z_{m+1}) \geq \alpha_{i_\beta(j)} q \delta^m.
\end{aligned}$$

(Again, recall that $\delta > 0$ was introduced in §IV-A2.) Since $q > 0$ (18), we thus obtain:

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

as required.

9) Likelihood ratio bounds: We next prove the following claims

$$\begin{aligned}
\psi_{i1}^{(L-1)}(e_{NL}) &\leq \psi_{11}^{(L-1)}(e_{NL}) \\
&\qquad \forall i \in S \quad \forall N = 0, \ldots, 2k-1, \qquad (28) \\
\frac{\psi_{ij}^{(L-1)}(e_{NL}) f_j(e_{(N+1)L})}{\psi_{11}^{(L-1)}(e_{NL}) f_1(e_{(N+1)L})} &< 1 - \epsilon \\
&\qquad \forall i, j \in S, j \neq 1, \forall N : 0 \leq N \leq 2k-1, \qquad (29) \\
\psi_{ij}^{(R-1)}(z_{m+1}) f_j(e_0) &\leq A^R \psi_{b_0\, 1}^{(R-1)}(z_{m+1}) f_1(e_0) \\
&\qquad \forall i, j \in S, \qquad (30) \\
\frac{\psi_{ij}^{(m+P-1)}(e_{2kL})}{\psi_{1j}^{(m+P-1)}(e_{2kL})} &\leq q^{-1} \Big(\frac{\Delta}{\delta} \Big)^{m-1} \\
&\qquad \forall i \in S, \forall j \in C. \qquad (31)
\end{aligned}$$

If $L = 1$, then (28) becomes $p_{i\,1} \leq p_{1\,1}$ 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 (§IV-A3). If $L = 1$, then also (29) becomes

$$\frac{p_{ij} f_j(e_{N+1})}{p_{1\,1} f_1(e_{N+1})} < 1 - \epsilon \quad \forall i, j \in S, j \neq 1,$$

and thus, since $e_{N+1} \in \mathcal{X}_1$, $0 \leq N < 2k$ in this case, (29) is true by the definition of $\mathcal{X}_1$ (§IV-A1) (and the fact that $p_1^* = p_{1\,1}$). Let us next prove (28) and (29) for the case $L > 1$. Fix i, $j \in S$ and $j \neq 1$, and consider any $N \in \{0, 1, \ldots, 2k-1\}$. Note that the definitions of the s-path (21) and the fact that $e_{NL+n} \in \mathcal{X}_{s_n}$ for $1 \leq n < L$ imply that for the given observations $e_{NL+1:(N+1)L-1}$, the path $s_{1:L-1}$ realizes the maximum in $\psi_{11}^{(L-1)}(e_{NL})$, i.e.

$$\begin{aligned}
\psi_{11}^{(L-1)}(e_{NL}) = &p_{1\, s_1} f_{s_1}(e_{NL+1}) p_{s_1\, s_2} \cdots \qquad (32) \\
&\cdots p_{s_{L-2}\, s_{L-1}} f_{s_{L-1}}(e_{(N+1)L-1}) p_{s_{L-1}\, 1}.
\end{aligned}$$

(Indeed, $p_{1\, s_1} f_{s_1}(e_{NL+1}) p_{s_1\, s_2} \cdots$

$$\cdots p_{s_{L-2}\, s_{L-1}} f_{s_{L-1}}(e_{(N+1)L-1}) p_{s_{L-1}\, 1} = p_{s_1}^* f_{s_1}(e_{NL+1}) p_{s_2}^* \cdots p_{s_{L-1}}^* f_{s_{L-1}}(e_{(N+1)L-1}) p_1^*,$$

and for $n = 1, 2, \ldots, L-1$, $p_{s_n}^* f_{s_n}(e_{NL+n}) \geq p_{ch} f_h(e_{NL+n})$ for any c, $h \in S$.) Suppose now that $h_{1:L-1}$ realizes

$\psi_{ij}^{(L-1)}(e_{NL})$, i.e.

$$\psi_{ij}^{(L-1)}(e_{NL}) = p_{i\,h_1} f_{h_1}(e_{NL+1}) p_{h_1\,h_2} \cdots \cdots p_{h_{L-2}\,h_{L-1}} f_{h_{L-1}}(e_{(N+1)L-1}) p_{h_{L-1}\,j}. \tag{33}$$

Hence, with h_0 and h_L standing for i and j, respectively (and $s_0 = s_L = 1$), the left-hand side of (29) becomes

$$\Big(\frac{p_{h_0\,h_1} f_{h_1}(e_{NL+1})}{p_{s_0\,s_1} f_{s_1}(e_{NL+1})}\Big)\Big(\frac{p_{h_1\,h_2} f_{h_2}(e_{NL+2})}{p_{s_1\,s_2} f_{s_2}(e_{NL+2})}\Big)\cdots \Big(\frac{p_{h_{L-2}\,h_{L-1}} f_{h_{L-1}}(e_{(N+1)L-1})}{p_{s_{L-2}\,s_{L-1}} f_{s_{L-1}}(e_{(N+1)L-1})}\Big)\Big(\frac{p_{h_{L-1}\,h_L} f_{h_L}(e_{(N+1)L})}{p_{s_{L-1}\,s_L} f_1(e_{(N+1)L})}\Big). \tag{34}$$

Note that for any $n \in \{1, 2, \ldots, L\}$ such that $h_n \neq s_n$,

$$\frac{p_{h_{n-1}\,h_n} f_{h_n}(e_{NL+n})}{p_{s_{n-1}\,s_n} f_{s_n}(e_{NL+n})} < 1 - \epsilon, \text{ since } e_{NL+n} \in \mathcal{X}_{s_n}. \tag{35}$$

For all other $n \in \{1, 2, \ldots, L\}$, $h_n = s_n$ and therefore, the left-hand side of (35) becomes $\frac{p_{h_{n-1}\,h_n}}{p_{s_{n-1}\,s_n}} = \frac{p_{h_{n-1}\,s_n}}{p^*_{s_n}} \leq 1$ (by the property (16)). Since the last term of the product (34) above does satisfy (35) ($j \neq 1$), (29) is thus proved. Suppose next that $h_{1:L-1}$ instead realizes $\psi_{i1}^{(L-1)}(e_{NL})$. With $s_0 = 1$ and $h_0 = i$, similarly to the previous arguments, we have

$$\frac{\psi_{i\,1}^{(L-1)}(e_{NL})}{\psi_{1\,1}^{(L-1)}(e_{NL})} = \prod_{n=1}^{L-1}\Big(\frac{p_{h_{n-1}\,h_n} f_{h_n}(e_{NL+n})}{p_{s_{n-1}\,s_n} f_{s_n}(e_{NL+n})}\Big)\frac{p_{h_{L-1}\,1}}{p_{s_{L-1}\,1}} \leq 1,$$

implying (28).

Let us now prove (30). To that end, note that for all states $i, j, h \in S$ such that $p_{jh} > 0$, it follows from the definitions (9) and (19) that

$$\frac{p_{ih}}{p_{jh}} \leq \frac{p^*_h}{p_{jh}} \leq A. \tag{36}$$

If $R = 1$, then (30) becomes

$$p_{ij} f_j(e_0) \leq A p_{b_0\,1} f_1(e_0).$$

By the definition of $\mathcal{X}_1$ (recall that $e_0 \in \mathcal{X}_1$), we have that for every $i, j \in S$ $p_{ij} f_j(e_0) \leq p^*_1 f_1(e_0)$. Using (36) with $h = 1$ and $j = b_0$, we get that $p^*_1 f_1(e_0) \leq A p_{b_0\,1} f_1(e_0)$ ($p_{b_0\,1} > 0$ by the construction of $\mathbf{b}$ §IV-A3). Putting these all together, we obtain

$$p_{ij} f_j(e_0) < p^*_1 f_1(e_0) \leq A p_{b_0\,1} f_1(e_0), \text{ as required.}$$

Consider the case $R > 1$. Let $h_{1:R-1}$ be a path that achieves $\psi_{ij}^{(R-1)}(z_{m+1})$, i.e. $\psi_{ij}^{(R-1)}(z_{m+1}) =$

$$p_{i\,h_1} f_{h_1}(e'_1) p_{h_1\,h_2} f_{h_2}(e'_2) \cdots p_{h_{R-2}\,h_{R-1}} f_{h_{R-1}}(e'_{R-1}) p_{h_{R-1}j}.$$

By the definition of the $\mathcal{X}_i$ (§IV-A1) and the facts that $e'_r \in \mathcal{X}_{b_r}$, $r = 1, 2, \ldots, R-1$, and $e_0 \in \mathcal{X}_1$, we have that

$$\psi_{ij}^{(R-1)}(z_{m+1}) f_j(e_0) \leq p^*_{b_1} f_{b_1}(e'_1) p^*_{b_2} f_{b_2}(e'_2) \cdots p^*_{b_{R-1}} f_{b_{R-1}}(e'_{R-1}) p^*_1 f_1(e_0). \tag{37}$$

Now, by the construction of $\mathbf{b}$ (§IV-A3), $p_{b_{r-1}\,b_r} > 0$ for $r = 1, \ldots, R$, ($b_R = 1$). Thus, the argument behind (36) also applies here to bound the right-hand side of (37) from above by

$$A p_{b_0\,b_1} f_{b_1}(e'_1) A p_{b_1\,b_2} f_{b_2}(e'_2) \cdots A p_{b_{R-2}\,b_{R-1}} f_{b_{R-1}}(e'_{R-1}) A p_{b_{R-1}\,1} f_1(e_0) = A^R \psi_{b_0\,1}^{(R-1)}(z_{m+1}) f_1(e_0), \text{ as required.}$$

Let us now prove (31). If $m = 1$ then (31) becomes

$$\psi_{ij}^{(P)}(e_{2kL}) \leq \psi_{1j}^{(P)}(e_{2kL}) q^{-1} \quad \forall i \in S,\ \forall j \in C. \tag{38}$$

If $P = 0$, then (38) reduces to $p_{ij} \leq p_{1j} q^{-1}$ which is true, because in this case the state $h_1 = h_T = 1$ belongs to C (§IV-A3) and $p_{1j} q^{-1} \geq 1$ ((17), (18) with $m = 1$). To see why (38) is also true with $P \geq 1$, note that by the same argument as used for proving (28) and (29), we now get that $\forall i,\ j \in S$

$$\psi_{1\,a_P}^{(P-1)}(e_{2kL}) f_{a_P}(e''_P) \geq \psi_{ij}^{(P-1)}(e_{2kL}) f_j(e''_P). \tag{39}$$

Also, since $a_P \in C$ (§IV-A3), we have that $p_{a_P\,j} q^{-1} \geq 1$ ((17), (18) with $m = 1$). Thus, for any $i \in S$ and for any $j \in C$, $\psi_{ij}^{(P)}(e_{2kL}) =$

$$\begin{aligned}
&\overset{\text{by (6)}}{=} \max_{c \in S} \psi_{i\,c}^{(P-1)}(e_{2kL}) f_c(e''_P) p_{c\,j} \\
&\overset{\text{by (39)}}{\leq} \psi_{1\,a_P}^{(P-1)}(e_{2kL}) f_{a_P}(e''_P) \max_{c \in S} p_{c\,j} \\
&\leq \psi_{1\,a_P}^{(P-1)}(e_{2kL}) f_{a_P}(e''_P) \\
&\leq \psi_{1\,a_P}^{(P-1)}(e_{2kL}) f_{a_P}(e''_P) p_{a_P\,j} q^{-1} \overset{\text{by (6)}}{\leq} \psi_{1\,j}^{(P)}(e_{2kL}) q^{-1}.
\end{aligned}$$

Now, assume that $m > 1$ and let i and j be any states in S and C, respectively. Also, let h be any state in S and let $h_{1:m-1}$ be a path realizing $\psi_{hj}^{(m-1)}(e''_P)$. Thus, $\psi_{hj}^{(m-1)}(e''_P) =$

$$\begin{aligned}
&= p_{h\,h_1} f_{h_1}(z'_1) p_{h_1\,h_2} f_{h_2}(z'_2) \cdots f_{h_{m-1}}(z'_{m-1}) p_{h_{m-1}\,j} \\
&< \Delta^{m-1}.
\end{aligned} \tag{40}$$

(This is true since $z'_n \in \mathcal{Z}$ for $n = 1, 2, \ldots, m-1$ (§IV-A2) and thus, for $\psi_{hj}^{(m-1)}(e''_P)$ to be positive it is necessary that $h_n \in C$, $n = 1, \ldots, m-1$, implying also that $f_{h_n}(z'_n) < \Delta$.) Now, let $h_{1:m-1}$ instead realize $\psi_{a_P\,j}^{(m-1)}(e''_P)$, which is clearly positive, with $h_n \in C$, $n = 1, \ldots, m-1$ ($z'_n \in \mathcal{Z}$ for $n = 1, 2, \ldots, m-1$), and $a_P, j \in C$ (recall the positivity assumption on $\mathbb{Q}^m$, §IV-A4). Also, let $c_{1:m-1} \in C^{m-1}$ be a path realizing $q^*_{a_P\,j}$ in the sense of (17). We then have that $\psi_{a_P\,j}^{(m-1)}(e''_P) = p_{a_P\,h_1} f_{h_1}(z'_1) p_{h_1\,h_2} f_{h_2}(z'_2) \cdots f_{h_{m-1}}(z'_{m-1}) p_{h_{m-1}\,j} \geq$

$$\geq q^*_{a_P\,j} f_{c_1}(z'_1) f_{c_2}(z'_2) \cdots f_{c_{m-1}}(z'_{m-1}) > q\delta^{m-1}, \tag{41}$$

where the last inequality follows from the definition of q (18) and condition (i) of §IV-A2. Combining the bounds of (40) and (41) ($q > 0$, (18)), we obtain for any $h \in S$:

$$\psi_{hj}^{(m-1)}(e''_P) < \psi_{a_P\,j}^{(m-1)}(e''_P) \left(\frac{\Delta}{\delta}\right)^{m-1} \Big/ q. \tag{42}$$

Finally, we have that $\psi_{ij}^{(P+m-1)}(e_{2kL}) =$

$$\overset{\text{by (6)}}{=} \max_{h\in S} \psi_{ih}^{(P-1)}(e_{2kL}) f_h(e''_P) \psi_{hj}^{(m-1)}(e''_P)$$

$$\overset{\text{by (39), (42)}}{<} \psi_{1\,a_P}^{(P-1)}(e_{2kL}) f_{a_P}(e''_P) \psi_{a_P\,j}^{(m-1)}(e''_P) \left(\frac{\Delta}{\delta}\right)^{m-1} \Big/ q$$

$$\overset{\text{by (6)}}{\leq} \psi_{1\,j}^{(P+m-1)}(e_{2kL}) \left(\frac{\Delta}{\delta}\right)^{m-1} \Big/ q \quad \text{as required by (31).}$$

10) $\gamma_j \leq const \times \gamma_1$*:* Combining (25), (27), and (30), we see that for every state $j \in S$,

$$\begin{aligned}\gamma_j &\overset{\text{by (25)}}{=} \beta_{i_\gamma(j)} \psi_{i_\gamma(j)\,j}^{(R-1)}(z_{m+1}) f_j(e_0) \\ &\overset{\text{by (30)}}{\leq} \beta_{i_\gamma(j)} \psi_{b_0\,1}^{(R-1)}(z_{m+1}) f_1(e_0) A^R \\ &\overset{\text{by (27)}}{\leq} q^{-1}\Big(\frac{\Delta}{\delta}\Big)^m A^R \beta_{b_0} \psi_{b_0\,1}^{(R-1)}(z_{m+1}) f_1(e_0) \\ &\leq W \max_{i\in S} \beta_i \psi_{i\,1}^{(R-1)}(z_{m+1}) f_1(e_0) \overset{\text{by (25)}}{=} W\gamma_1,\end{aligned}$$

where

$$W \overset{\text{def}}{=} q^{-1}\Big(\frac{\Delta}{\delta}\Big)^m A^R. \tag{43}$$

Hence

$$\gamma_j \leq W\gamma_1 \quad \forall j \in S. \tag{44}$$

11) Further bounds on likelihoods: Let $N \geq 0$ and $n > 0$ be integers such that $N + n \leq 2k$ but arbitrary otherwise. Expanding $\psi_{1\,1}^{(nL-1)}(e_{NL})$ recursively according to (6), we obtain

$$\begin{aligned}\psi_{1\,1}^{(nL-1)}(e_{NL}) = \max_{i_{1:n-1}\in S^{n-1}} &\psi_{1\,i_1}^{(L-1)}(e_{NL}) f_{i_1}(e_{(N+1)L}) \times \\ \times \psi_{i_1\,i_2}^{(L-1)}(e_{(N+1)L}) f_{i_2}(e_{(N+2)L}) &\cdots \psi_{i_{n-2}\,i_{n-1}}^{(L-1)}(e_{(N+n-2)L}) \times \\ \times \quad f_{i_{n-1}}(e_{(l+n-1)L}) &\psi_{i_{n-1}\,1}^{(L-1)}(e_{(N+n-1)L}).\end{aligned} \tag{45}$$

Since it follows from (29) that for any $i_1 \in S$, $\psi_{1\,i_1}^{(L-1)}(e_{NL}) f_{i_1}(e_{(N+1)L}) \leq \psi_{1\,1}^{(L-1)}(e_{NL}) f_1(e_{(N+1)L})$, as well as

$$\begin{aligned}&\psi_{i_{r-1}\,i_r}^{(L-1)}(e_{(N+r-1)L}) f_{i_r}(e_{(N+r)L}) \overset{\text{by (29)}}{\leq} \\ &\qquad \psi_{1\,1}^{(L-1)}(e_{(N+r-1)L}) f_1(e_{(N+r)L}),\ r = 2,\ldots,n-1,\end{aligned}$$

and since for any $i_{n-1} \in S$

$$\psi_{i_{n-1}\,1}^{(L-1)}(e_{(N+n-1)L}) \overset{\text{by (28)}}{\leq} \psi_{1\,1}^{(L-1)}(e_{(N+n-1)L}),$$

maximization (45) above is achieved as stated in (46) below:

$$\begin{aligned}&\psi_{1\,1}^{(nL-1)}(e_{NL}) = \\ &\psi_{1\,1}^{(L-1)}(e_{NL}) f_1(e_{(N+1)L}) \psi_{1\,1}^{(L-1)}(e_{(N+1)L}) f_1(e_{(N+2)L}) \cdots \\ &\cdots \psi_{1\,1}^{(L-1)}(e_{(N+n-2)L}) f_1(e_{(N+n-1)L}) \psi_{1\,1}^{(L-1)}(e_{(N+n-1)L}).\end{aligned} \tag{46}$$

Now, we replace state 1 by generic states $i, j \in S$ on the both ends of the paths in (45) and repeat the above arguments. Thus, also using (46), we arrive at the bound (47) below:

$$\begin{aligned}&\psi_{ij}^{(nL-1)}(e_{NL}) f_j(e_{(N+n)L}) \leq \\ &\prod_{t=N+1}^{N+n} \psi_{1\,1}^{(L-1)}(e_{(t-1)L}) f_1(e_{tL}) \overset{\text{by (46)}}{=} \\ &\psi_{1\,1}^{(nL-1)}(e_{NL}) f_1(e_{(N+n)L}) \quad \forall i,j \in S.\end{aligned} \tag{47}$$

With $n = N = k$ in particular, (47) states that $\forall i, j \in S$

$$\psi_{ij}^{(kL-1)}(e_0) f_j(e_{kL}) \leq \psi_{1\,1}^{(kL-1)}(e_0) f_1(e_{kL}). \tag{48}$$

12) $\eta_j \leq const \times \eta_1$*:* In order to see that

$$\eta_j \leq W\eta_1 \quad \forall j \in S, \tag{49}$$

$$\begin{aligned}\text{note: } \eta_j &\overset{(26)}{=} \max_{i\in S} \gamma_i \psi_{ij}^{(kL-1)}(e_0) f_j(e_{kL}) \\ &\overset{\text{by (48)}}{\leq} \max_{i\in S} \gamma_i \psi_{1\,1}^{(kL-1)}(e_0) f_1(e_{kL}) \overset{\text{by (44)}}{\leq} \\ &\overset{\text{by (44)}}{\leq} W\gamma_1 \psi_{1\,1}^{(kL-1)}(e_0) f_1(e_{kL}) \overset{\text{by (26)}}{\leq} W\eta_1.\end{aligned}$$

13) A representation of $\eta_1 (= \phi_1(e_{kL}))$*:* Recall that k, the number of cycles in the s-path, was chosen sufficiently large for (20) to hold (in particular, $k > 1$). We now prove that there exists $\kappa \in \{1,\ldots,k-1\}$ such that

$$\eta_1 = \phi_1(e_{\kappa L}) \psi_{1\,1}^{((k-\kappa)L-1)}(e_{\kappa L}) f_1(e_{kL}). \tag{50}$$

The relation (50) states that a maximum-likelihood path from time 1 (observation x_1) to time $u-\rho$ (observation e_{kL}) goes through state 1 at time $u-\rho-(k-\kappa)L = u-m-P-2kL+\kappa L$, that is when $e_{\kappa L}$ is observed.

To see this, suppose no such κ existed. Then, applying (6) to (26) and recalling that $\phi_1(e_{\kappa L})$ was introduced in (23), we would have

$$\begin{aligned}\eta_1 = \gamma_{j_\eta(1)} &\psi_{j_\eta(1)\,j_1}^{(L-1)}(e_0) f_{j_1}(e_L) \psi_{j_1\,j_2}^{(L-1)}(e_L) \times \\ &\times f_{j_2}(e_{2L}) \psi_{j_2\,j_3}^{(L-1)}(e_{2L}) \cdots \psi_{j_{k-1}\,1}^{(L-1)}(e_{(k-1)L}) f_1(e_{kL})\end{aligned}$$

for some $j_1 \neq 1, \ldots, j_{k-1} \neq 1$. Furthermore, this would imply $\eta_1 <$

$$\begin{aligned}&\overset{\text{by (29), (28)}}{<} \gamma_{j_\eta(1)}(1-\epsilon)^{k-1} \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}) \\ &\overset{\text{by (20)}}{<} \gamma_{j_\eta(1)} q^2 \left(\frac{\delta}{\Delta}\right)^{2m} A^{-R} \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}) \\ &\overset{\text{by (44)}}{\leq} \gamma_1 W q^2 \left(\frac{\delta}{\Delta}\right)^{2m} A^{-R} \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}) \\ &\overset{\text{by (43)}}{=} \gamma_1 q \left(\frac{\delta}{\Delta}\right)^{m} \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}) \\ &< \gamma_1 \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}).\end{aligned} \tag{51}$$

(The last inequality follows from $q \leq 1$ (18) and $\delta < \Delta$, §IV-A2.) On the other hand, by the definition (26) (and $k-1$-fold application of (6)), $\eta_1 \geq \gamma_1 \prod_{n=1}^{k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL})$, which evidently contradicts (51) above. Therefore, κ satisfying (50) and $1 \leq \kappa < k$, does exist.

14) An implication of (46) *and* (50) *for* $\phi_1(e_{NL})$*:* Clearly, the arguments of the previous step (§IV-A13) are valid if k is replaced by any $N \in \{k, \dots, 2k\}$. Hence the following generalization of (50): For some $\kappa(N) < N$

$$\phi_1(e_{NL}) = \phi_1(e_{\kappa(N)L})\psi_{1\,1}^{((N-\kappa(N))L-1)}(e_{\kappa(N)L})f_1(e_{NL}). \tag{52}$$

We now apply the existence assertion of the previous step (§IV-A13) to (52) recursively, starting with $\kappa^{(0)} \stackrel{\text{def}}{=} N$ and obtaining (the existence of) $\kappa^{(1)} \stackrel{\text{def}}{=} \kappa(N) < N$. If $\kappa^{(1)} \leq k$, we stop, otherwise we substitute $\kappa^{(1)}$ for N in (52), and obtain $\kappa^{(2)} \stackrel{\text{def}}{=} \kappa(\kappa^{(1)}) < \kappa^{(1)}$, and so on, until $\kappa^{(r)} \leq k$ for some $r > 0$. Thus, $\phi_1(e_{NL}) =$

$$= \phi_1(e_{\kappa^{(r)}L})\psi_{1\,1}^{((\kappa^{(r-1)}-\kappa^{(r)})L-1)}(e_{\kappa^{(r)}L})f_1(e_{\kappa^{(r-1)}L})\cdots \psi_{1\,1}^{((N-\kappa^{(1)})L-1)}(e_{\kappa^{(1)}L})f_1(e_{NL}). \tag{53}$$

Applying (46) to the appropriate factors of the right-hand side of (53) above, we obtain:

$$\begin{aligned}\phi_1(e_{NL}) = \phi_1(e_{\kappa^{(r)}L})\psi_{1\,1}^{(L-1)}(e_{\kappa^{(r)}L})f_1(e_{(\kappa^{(r)}+1)L})\cdots \\ \psi_{1\,1}^{(L-1)}(e_{(k-1)L})f_1(e_{kL})\cdots\psi_{1\,1}^{(L-1)}(e_{kL})f_1(e_{(k+1)L})\cdots \\ \psi_{1\,1}^{(L-1)}(e_{(\kappa^{(r-1)}-1)L})f_1(e_{\kappa^{(r-1)}L})\cdots \\ \psi_{1\,1}^{(L-1)}(e_{(\kappa^{(1)}-1)L})f_1(e_{\kappa^{(1)}L})\cdots \\ \psi_{1\,1}^{(L-1)}(e_{(N-1)L})f_1(e_{NL}).\end{aligned} \tag{54}$$

Also, according to (46),

$$\phi_1(e_{\kappa^{(r)}L})\psi_{1\,1}^{(L-1)}(e_{\kappa^{(r)}L})f_1(e_{(\kappa^{(r)}+1)L})\cdots \psi_{1\,1}^{(L-1)}(e_{(k-1)L}) = \phi_1(e_{\kappa^{(r)}L})\psi_{1\,1}^{((k-\kappa^{(r)})L-1)}(e_{\kappa^{(r)}L}).$$

At the same time,

$$\phi_1(e_{\kappa^{(r)}L})\psi_{1\,1}^{((k-\kappa^{(r)})L-1)}(e_{\kappa^{(r)}L})f_1(e_{kL}) \stackrel{\text{by (6)}}{\leq} \eta_1. \tag{55}$$

However, we cannot have the strict inequality in (55) above since that, by virtue of (54), would contradict maximality of $\phi_1(e_{NL})$. We have thus arrived at $\phi_1(e_{NL}) = \eta_1\psi_{1\,1}^{(L-1)}(e_{kL})f_1(e_{(k+1)L})\cdots$

$$\cdots\psi_{1\,1}^{(L-1)}(e_{(N-1)L})f_1(e_{NL}). \tag{56}$$

In summary, for any $N \in \{k, \dots, 2k\}$, there exists a realization of $\phi_1(e_{NL})$ that goes through state 1 as the e_{nL}, $n = k, \dots, N$, are observed.

15) $x_{u-\rho}$ (=e_{kL}) is a 1-node of order ρ: In §IV-A16, we will prove that for any $i \in S$ and any $j \in C$,

$$\eta_i\psi_{ij}^{(kL+m+P-1)}(e_{kL}) \leq \eta_1\psi_{1j}^{(kL+m+P-1)}(e_{kL}), \tag{57}$$

which implies that $x_{u-\rho}$ is a 1-node of order $\rho = kL+m+P$. Indeed, let $h \in S$ be arbitrary. Recall (§IV-A6) that we have been considering $x_{1:\infty} \in \mathcal{X}^\infty$ to be any sequence that contains the subsequence $z_{1:M}$ (22), i.e. $x_{u-M+1:u} = z_{1:M}$ for some $u \geq M$. Recall also that it is this latter subsequence $z_{1:M}$ which we are proving to be a 1-barrier of order $\rho = kL + m + P$. Since $f_j(z'_m) = 0$ for every $j \in S \setminus C$, any maximum likelihood path to state h at time $u+1$ must go through a state in C at time u (observation $x_u = z'_m$.) Formally,

$$\begin{aligned}\eta_i\psi_{ih}^{(kL+m+P)}(e_{kL}) = \\ &= \max_{j\in S}\eta_i\psi_{ij}^{(kL+m+P-1)}(e_{kL})f_j(z'_m)p_{jh} \\ &= \max_{j\in C}\eta_i\psi_{ij}^{(kL+m+P-1)}(e_{kL})f_j(z'_m)p_{jh} \\ &\stackrel{\text{by (57)}}{\leq} \max_{j\in C}\eta_1\psi_{1j}^{(kL+m+P-1)}(e_{kL})f_j(z'_m)p_{jh} \\ &\stackrel{\text{by (6)}}{=} \eta_1\psi_{1h}^{(kL+m+P)}(e_{kL}).\end{aligned}$$

Therefore, by Definition 2.1 e_{kL} is a 1-node of order $kL + m + P$.

16) Proof of (57)*:* Let us write $\nu(i,j)$ for $\psi_{ij}^{(kL-1)}(e_{kL})f_j(e_{2kL})$, $i, j \in S$. Let $i \in S$ and $j \in C$ be arbitrary and let state $j^* \in S$ be such that

$$\psi_{ij}^{(kL+m+P-1)}(e_{kL}) = \nu(i,j^*)\psi_{j^*j}^{(m+P-1)}(e_{2kL}). \tag{58}$$

We consider the following two cases separately:

1. There exists a path realizing $\psi_{ij^*}^{(kL-1)}(e_{kL})$ and going through state 1 at the time of observing e_{NL} for some $N \in \{k, \dots, 2k\}$. That is, $\psi_{ij^*}^{(kL-1)}(e_{kL}) =$

$$\psi_{i\,1}^{((N-k)L-1)}(e_{kL})f_1(e_{NL})\psi_{1\,j^*}^{((2k-N)L-1)}(e_{NL}). \tag{59}$$

Equation (59) above together with the fundamental recursion (6) yields the following:

$$\begin{aligned}\eta_i\psi_{ij^*}^{(kL-1)}(e_{kL}) = \\ &\stackrel{\text{by (59)}}{=} \eta_i\psi_{i\,1}^{((N-k)L-1)}(e_{kL})f_1(e_{NL})\psi_{1\,j^*}^{((2k-N)L-1)}(e_{NL}) \\ &\stackrel{\text{by (23), (6)}}{\leq} \phi_1(e_{NL})\psi_{1\,j^*}^{((2k-N)L-1)}(e_{NL}).\end{aligned} \tag{60}$$

At the same time, the right hand-side of (60) can be expressed as follows:

$$\begin{aligned}\phi_1(e_{NL})\psi_{1\,j^*}^{((2k-N)L-1)}(e_{NL}) \\ &\stackrel{\text{by (56)}}{=} \eta_1\psi_{1\,1}^{((N-k)L-1)}(e_{kL})f_1(e_{NL})\psi_{1\,j^*}^{((2k-N)L-1)} \\ &\stackrel{\text{by (46)}}{=} \eta_1\psi_{1\,j^*}^{(kL-1)}(e_{kL}).\end{aligned} \tag{61}$$

Therefore, if there exists $N \in \{k, \dots, 2k\}$ such that (59) holds, we have by virtue of (60) and (61) that $\eta_i\psi_{i\,j^*}^{(kL-1)}(e_{kL}) \leq \eta_1\psi_{1\,j^*}^{(kL-1)}(e_{kL})$, and, multiplying both sides by $f_{j^*}(e_{2kL})$, that

$$\eta_i\nu(i,j^*) \leq \eta_1\nu(1,j^*). \tag{62}$$

Hence, $\eta_i\psi_{ij}^{(kL+m+P-1)}(e_{kL}) =$

$$\begin{aligned}&\stackrel{\text{by (58)}}{=} \eta_i\nu(i,j^*)\psi_{j^*j}^{(m+P-1)}(e_{2kL}) \\ &\stackrel{\text{by (62)}}{\leq} \eta_1\nu(1,j^*)\psi_{j^*j}^{(m+P-1)}(e_{2kL}) \\ &\stackrel{\text{by (6)}}{\leq} \eta_1\psi_{1\,j}^{(kL+m+P-1)}(e_{kL})\end{aligned}$$

and (57) holds.

2. Assume now that no path exists to satisfy (59). Argue as for (51) to obtain that

$$\nu(i,j^*) < (1-\epsilon)^{k-1} \prod_{n=k+1}^{2k} \psi_{1\,1}^{(L-1)}(e_{(n-1)L}) f_1(e_{nL}). \quad (63)$$

By (46), the (partial likelihood) product in the right-hand side of (63) equals $\nu(1,1)$. Thus,

$$\begin{aligned}
&\eta_i \nu(i,j^*) \psi_{j^*j}^{(m+P-1)}(e_{2kL}) < \\
&\overset{\text{by (63)}}{<} \eta_i (1-\epsilon)^{k-1} \nu(1,1) \psi_{j^*j}^{(m+P-1)}(e_{2kL}) \\
&\overset{\text{by (20)}}{<} \eta_i q^2 \left(\frac{\delta}{\Delta}\right)^{2m} A^{-R} \nu(1,1) \psi_{j^*j}^{(m+P-1)}(e_{2kL}) \\
&\overset{\text{by (43), (49)}}{\leq} \eta_1 q \left(\frac{\delta}{\Delta}\right)^{m} \nu(1,1) \psi_{j^*j}^{(m+P-1)}(e_{2kL}) \\
&\overset{\text{by (31)}}{\leq} \eta_1 \left(\frac{\delta}{\Delta}\right) \nu(1,1) \psi_{1\,j}^{(m+P-1)}(e_{2kL}) \\
&< \eta_1 \nu(1,1) \psi_{1\,j}^{(m+P-1)}(e_{2kL}) \\
&\overset{\text{by (6)}}{\leq} \eta_1 \psi_{1\,j}^{(kL+m+P-1)}(e_{kL}),
\end{aligned}$$

which, by the way the state j^* is defined in (58), yields (57). (It was also used above that $\Delta > \delta > 0$, cf. §IV-A2.)

17) Completion of the s-path to $y_{1:M}$ and conclusion: Recall from §IV-A3 that $b_0 \in C$. Since all the entries of $\mathbb{Q}^m$ are positive, there exists a path $c_{0:m-1} \in C^m$ such that $p_{c_i\, c_{i+1}} > 0$, $i = 0, 1, \ldots, m-2$, and $p_{c_{m-1} b_0} > 0$. Similarly, there must exist a path $o_{1:m} \in C^m$ such that $p_{o_i\, o_{i+1}} > 0$, $i = 1, \ldots, m-1$, and $p_{a_P\, o_1} > 0$ (recall that $a_P \in C$). Hence, by these, and the constructions of §IV-A5, all of the transitions of the sequence $y_{1:M}$ given in (64) below occur with positive probabilities.

$$y_{1:M} \overset{\text{def}}{=} (c_{0:m-1}, b_{0:R}, s_{1:2kL}, a_{1:P}, o_{1:m}). \quad (64)$$

Clearly, the actual probability of observing $y_{1:M}$ is positive, as required. By the constructions of §§IV-A1-IV-A3, the conditional probability of B (22), given $Y_{1:M} = y_{1:M}$, is evidently positive, as required. Finally, since the sequence $z_{1:M} = (z_{1:m+1}, e'_{1:R-1}, e_{0:2kL}, e''_{1:P}, z'_{1:m})$ (22) was chosen from B arbitrarily (cf. §IV-A6) and has been shown to be a barrier, this completes the proof of the Lemma.

B. Illustration of the construction

Most of the above proof trivializes in the case of $K = 2$ where strong (and generally short) barriers of sensible probability can be easily constructed following the adapted proof in [15].

Here, we illustrate the above construction in a more challenging setting which in particular has a large proportion of forbidden transitions and continuous emissions. Once again, we do not attempt here to optimize the probability $\mathbf{P}(X_{1:M} \in B)$ of barriers or their length M, or order ρ, but simply illustrate the steps of the construction.

Example 4.1: Thus, let

$$\mathbb{P} = \begin{pmatrix} 1/4 & 1/4 & 0 & 1/2 \\ 1/4 & 1/4 & 1/2 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \end{pmatrix}$$

and let $f_1(x) = \exp(-x)_{x \geq 0}$, $f_2(x) = 2\exp(-2x)_{x \geq 0}$, and $f_3(x) = \exp(x)_{x \leq 0}$, $f_4(x) = 2\exp(2x)_{x \leq 0}$. Note that $C_1 \overset{\text{def}}{=} \{1, 2\}$ is a cluster which (unlike cluster $C_2 \overset{\text{def}}{=} \{3, 4\}$) satisfies the assumptions of the Lemma.

Take ϵ, δ, and Δ to be 0.1, 0.9, and 2, respectively. Thus obtain

$$\mathcal{X}_i = \begin{cases} (\log(20/9), \infty) & \text{if } i = 1 \\ (0, \log(1.8)) & \text{if } i = 2 \\ (-\infty, -\log(20/9)) & \text{if } i = 3 \\ (-\log(1.8), 0) & \text{if } i = 4 \end{cases},$$

and $\hat{\mathcal{Z}} = (0, \log(10/9))$ and $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_2 = \hat{\mathcal{Z}}$, i.e. $s = 2$, and so we have to take h_1 to be 2. Note then that $h_2 = 4$ and $p_4^* = p_{4\,4}$, hence choose h_3 to be also 4. Since $h_2 = h_3$, $U = 3$, $T = 2$, $P = L = 1$, $a_1 \equiv a_P = 2$ and $s_1 \equiv s_L = 4$ (we do not rename 4 into 1 here). We also obtain here $\mathbf{b} = (1, 4)$, i.e. $b_0 = 1 \in C_1$ and $R = 1$.

To determine k, half the number of the s-cycles, note that since $m = 1$, we have $q^*_{ij} = q_{ij}$ for all $i, j \in C_1$. Hence, $q = 1/4$. Also, $A = \max\{2, 2, 1, 1\} = 2$. Thus, k must be (at least) 50. Therefore, the s-path is given by

$$(1, \underbrace{4, \ldots, 4}_{101 \text{ times}}, 2),$$

and we can take $y_{1:M}$ to be

$$(1, 1, \underbrace{4, \ldots, 4}_{101 \text{ times}}, 2, 2),$$

where evidently $M = 105$, and the type and order of the obtained barriers/nodes are $l = 4$ and $\rho = 52$, respectively. Also,

$$B = \mathcal{Z} \times \mathcal{Z} \times \underbrace{\mathcal{X}_4 \times \cdots \times \mathcal{X}_4}_{101 \text{ times}} \times \mathcal{X}_2 \times \mathcal{Z}.$$

Since $\mathcal{X}_4$ and $\mathcal{Z}$ are disjoint, the barriers in B are already separated.

Note that the stationary distribution π here is uniform, and that there are only eight (out of the total of 4^{105}) paths of length M and of positive probability, which can "emit B"! Since the probability of each of those paths is 2^{-108}, the probability $\mathbf{P}(X_{1:M} \in B)$ of encountering a barrier from B is extremely low, i.e. less than 2^{-105}. Certainly, if one wanted to achieve sensible values for $\mathbf{P}(X_{1:M} \in B)$, a critical point for optimization would be bound (20).

C. Proof of Lemma 3.2

Proof: We use the notation of the previous proof in §IV-A and consider the following two distinct situations: In the first one (§IV-C1), all the barriers from B as constructed in the proof of Lemma 3.1 are already separated. In the other, complimentary situation, a simple extension will immediately ensure separation (§IV-C2).

1) All $z_{1:M} \in B$ are already separated: Recall the definition of $\mathcal{Z}$ from §IV-A2. Consider the two cases in the definition separately. First, suppose $\mathcal{Z} = \hat{\mathcal{Z}} \backslash (\cup_{i \in S} \mathcal{X}_i)$, in which case $\mathcal{Z}$ and $\mathcal{X}_i$ are disjoint for every $i \in S$. This implies that every barrier (22) is already separated. Indeed, for any r, $1 \leq r \leq \rho$, and for any $z_{1:M} \in B$, the fact that $z_{M-\max(m,r)}$, for example, is not in $\mathcal{Z}$, makes it impossible for $(e'_{1:r}, z_{1:M-r})$ to be in B for any $e'_{1:r} \in \mathcal{X}^r$. Consider now the case when $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$ for some $s \in C$, and $s = a_P$. Then

$$B = \mathcal{X}_s^{m+1} \times \mathcal{X}_{b_1} \times \cdots \times \mathcal{X}_{b_{R-1}} \times \mathcal{X}_1 \times \mathcal{X}_{s_1} \times \cdots \mathcal{X}_{s_{2kL-1}} \times \mathcal{X}_1 \times \mathcal{X}_{a_1} \times \cdots \times \mathcal{X}_{a_{P-1}} \times \mathcal{X}_s^{m+1}. \quad (65)$$

Let $z_{1:M} \in B$ be arbitrary. Assume first $L > 1$. By construction (§IV-A3), the states $s_1, \ldots, s_L$ are all distinct. We now show that $(e'_{1:r}, z_{1:M-r}) \notin B$ for any $e'_{1:r} \in \mathcal{X}^r$ when $1 \leq r \leq \rho$. Note that $y_{1:M}$ was chosen such that in the sequence

$$y_{m+2:m+R+2kL+P+1} = (b_{1:R-1}, 1, s_{1:2kL-1}, 1, a_{1:P-1}, s)$$

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

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

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

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

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

Next consider the case of $L = 1$ but $s \neq 1$ (that is, $P > 0$). Then $B = \mathcal{X}_s^{m+1} \times \mathcal{X}_{b_1} \times \cdots$

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

Since $s \neq 1$, then also $b_r \neq 1$, $r = 1, \ldots, R-1$ (by minimality of R), and $a_r \neq 1$, $r = 1, \ldots, P-1$. To see that $z_{1:M}$ is separated in this case, simply note that $z_{M-max(r,m+1)} \notin \mathcal{X}_s$ for any admissible r.

2) Barriers $z_{1:M} \in B$ need not be separated: Finally, we consider the case when $L = 1$ and $s = 1$, where $s \in C$ is such that $\mathcal{Z} = \hat{\mathcal{Z}} \cap \mathcal{X}_s$. Note that in this case, $P = 0$, $1 \in C$, and $p_{1\,1} > 0$, which in turn implies that $R = 1$, and

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

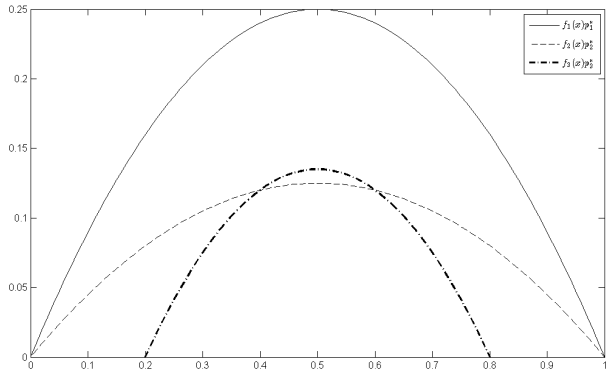

Fig. 1. Violation of condition (10) for states $j = 2$ and $j = 3$ ($K = 3$). To help interpret this situation, suppose Y is an i.i.d. mixture model, hence $p_i^* = \pi_i$, $i = 1, 2, 3$, and the Viterbi alignment $v = (1, 1, \ldots)$ does not "see" the hidden states 2 and 3 at all.

Clearly, the barriers from B need not be, and in fact are not separated. It is, however, easy to extend them to achieve separation. Indeed, let $y_0 \neq 1$ be such that $p_{y_0\,1} > 0$ and redefine $B \stackrel{\text{def}}{=} \mathcal{X}_{y_0} \times B$. Evidently, any shift of any $z_{1:M+1} \in B$ by r ($1 \leq r \leq \rho$) positions to the right makes it impossible for z_1 to be simultaneously in $\mathcal{X}_{y_0}$ and in $\mathcal{X}_1$ (since the latter sets are disjoint, §IV-A1). ■

V. Conclusion

We conclude by discussing briefly the assumptions under which Lemmas 3.1 and 3.2 are presently proved.

Condition (10) simply requires that state $j \in S$ be "detectable". (See Figure 1 for an example of possible violation of this condition.) Namely, there must be a subset of the emission space such that $\left\{x \in \mathcal{X} : f_j(x) p_j^* > \max_{i \in S, i \neq j} f_i(x) p_i^*\right\}$ is of positive λ-measure. Presently, we require that this condition holds for every state $j \in S$. If the emission space $\mathcal{X}$ is finite and has fewer than K symbols, then certainly (10) will be violated for at least $K - |\mathcal{X}|$ states (where $|\mathcal{X}|$ is the size of $\mathcal{X}$). While it is not difficult to accommodate formally for such violations (see discussion below), it might actually be more meaningful in practice to redefine the model by either discarding some of the offending states or aggregating them in a suitable manner. After all, some such states may simply never appear in the Viterbi alignment. In short, we are not aware of practical situations where this requirement would cause an obstacle. Moreover, for many models (e.g. $K = 2$) it is actually sufficient for proving the existence of barriers that (10) holds for *at least one* state $j \in S$. Clearly, provided that the emission distributions P_i, $i \in S$, are all distinct, (10) does indeed hold for at least one state $j \in S$ for any (general) setting of the transition probabilities when $K = 2$ ($K \geq 2$). In [8], a stronger version of (10) [8, equation (3.6)] is discussed as it appears in the hypotheses of Theorems 1 and 3 of [9] and [10], respectively. Specifically, under the assumption that [8, equation (3.6)] holds for at least one state in S, Theorem 1 of [9] establishes the existence of infinite Viterbi alignments. It is shown in [8] that under the same condition the claims

of Lemmas 3.1 and 3.2, i.e. the occurrence of barriers, follow immediately. At the same time, [8] also demonstrates how that assumption (despite being made for a single state) can be too restrictive in practice. Hence the need for weaker conditions, such as those based on (10).

One can certainly relax the requirement that (10) holds for all $j \in S$ without introducing any additional assumptions (such as strict positivity of the transition probabilities or $K = 2$). Specifically, note that in the above proofs, (10) is used only along the s-path (cf. §IV-A5), which in turn depends on the cluster C. Thus, one can immediately generalize the present results by requiring that (10) hold at least for the relevant subset of states.

Of course, if we consider the broader problem of parameter estimation in HMMs, the parameter space might need to be suitably restricted to insure (10) does not reach the boundary (i.e. equality in place of the inequalities) for the relevant states. Note that restricting (emission or transition) parameters is done routinely in that context due to possible unboundedness of the likelihood, lack of identifiability, or exclusion of prescribed transitions.

The condition on C in Lemma 3.1 might seem even more technical if not redundant. We next give an example of an HMM where this cluster condition is not met and no node (barrier) can occur. (Curiously, this example appears to be also useful in other contexts [6].) Then, we will modify the example to enforce the cluster condition and consequently gain barriers.

Example 5.1: Let $K = 4$ and consider an ergodic Markov chain with transition matrix

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

Let the emission distributions be such that (10) is satisfied (say, for entire S) and $G_1 = G_2$ and $G_3 = G_4$ and $G_1 \cap G_3 = \emptyset$. To be concrete, take the emission distributions from Example 4.1 in §IV-B above. Hence, there are two disjoint clusters, $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$. The matrices $\mathbb{Q}_i$ corresponding to C_i, $i = 1, 2$ are

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

Evidently, the cluster assumption of Lemma 3.1 is not satisfied. Note also that the Viterbi alignment cannot change (in one step) its state to the opposite one within the same cluster, i.e. transitions $1 \leftrightarrow 2$ or $3 \leftrightarrow 4$ do not occur. Since the supports $G_{1,2}$ and $G_{3,4}$ are disjoint, any observation exposes the corresponding cluster. In effect, any sequence of observations $x_{1:T}$ is partitioned by the alignment $v_{1:T}$ into blocks $x_{1:t_1}$, $x_{t_1+1:t_2}$, …, $x_{t_N+1,T}$ (for some $N \leq T$) where the alignment inside each block stays constant, e.g. $v_1 = v_2 = \ldots = v_{t_1}$, but no two neighboring blocks can be emitted from the same cluster, e.g. if $v_1 = v_2 = \ldots = v_{t_1} = 1$ then it must be that $v_{t_1+1} = v_{t_1+2} = \ldots = v_{t_2} \in C_2$. It can then be shown that in this case no x_t can be a node (of any order) (cf. Example 3.11 in [19]).

Let us modify the HMM in Example 5.1 to ensure the assumptions of Lemma 3.1.

Example 5.2: Let ϵ be such that $0 < \epsilon < \frac{1}{2}$ and let us replace $\mathbb{P}$ by the following transition matrix

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

Let the emission distributions be as discussed in the previous example. In this case, the cluster C_1 satisfies the assumption of Lemma 3.1. As previously, every observation exposes its cluster. Lemma 3.1 now applies to guarantee barriers and nodes. To be more concrete, take again the emission distributions from Example 4.1 of §IV-B, and let $\epsilon = 1/4$, i.e. $\mathbb{P}$ also as in Example 4.1. It can then be verified that if $x_{1:3} = (1, 1, 1)$ 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.

Another way to modify the HMM in Example 5.1 to enforce the assumptions of Lemma 3.1 is to change the emission probabilities. Namely, 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 (10) holds (for all $j \in S$). Now, $S = \{1, \ldots, 4\}$ is the only cluster. Since all entries of the matrix $\mathbb{P}^2$ are positive, the conditions of Lemma 3.1 are now satisfied and barriers can now be constructed.

Thus, the cluster condition is essential. We need to clarify that this is all that we meant in [8] where instead of "essential", "necessary" was used loosely. Indeed, to see that the present formulation of the cluster condition can be relaxed, consider the following simple example with $K = 4$ and $\mathcal{X} = \{a, b, c, d\}$. Similarly to Examples 5.1 and 5.2, assume that $G_1 = G_2$ is disjoint from $G_3 = G_4$, say, $G_1 = G_2 = \{a, b\}$, and $G_3 = G_4 = \{c, d\}$. Let

$$\mathbb{P} = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 1/3 & 1/3 & 1/3 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/3 & 1/3 & 0 & 1/3 \end{pmatrix}.$$

so that we still have two clusters, $C_1 = \{1, 2\}$ and $C_2 = \{3, 4\}$, which do not satisfy the cluster condition in its present form. It is not difficult to specify the emission probabilities in ways that would ensure the existence of barriers. For example, let $f_1(a) = f_2(b) = f_3(c) = f_4(d) = 3/4$, and note that (c, a) is a 1-barrier of order $\rho = 0$, i.e. $x_2 = a$ is a 1-node (of order 0) for any realization containing $x_{1:2} = (c, a)$ ($y_{1:2}$ can be taken to be either $(3, 1)$ or $(4, 1)$). We actually conjecture that with some effort, the present proofs can be modified to prove the same results under the following relaxed cluster condition: There must exist $m \geq 1$ and state $i \in C$ such that any $j \in C$ can be reached from i via a positive probability path of length m that is also in C.

ACKNOWLEDGMENT

The first author has been supported by the Estonian Science Foundation Grant 7553, which has also supported the second author's visits to the University of Tartu. The authors thank

Eurandom (The Netherlands) for initiating and stimulating their research on hidden Markov models, of which this work has been an integral part. The authors are especially thankful to Dr. A. Caliebe for valuable discussions and for emphasizing the significance of the topic of path estimation in HMMs, which have encouraged this work. The authors are also grateful to anonymous reviewers as well as to the associate editor for their thorough review of this work, additional references, and comments and other suggestions to improve this manuscript.

REFERENCES

[1] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002, special issue on Shannon theory: perspective, trends, and applications.

[2] V. Genon-Catalot, T. Jeantheau, and C. Larédo, "Stochastic volatility models as hidden Markov models and statistical applications," *Bernoulli*, vol. 6, no. 6, pp. 1051–1079, December 2000.

[3] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process. Appl.*, vol. 40, no. 1, pp. 127–143, March 1992.

[4] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, February 1989.

[5] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Inform. Theory*, vol. 13, no. 2, pp. 260–269, April 1967.

[6] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, ser. Springer Series in Statistics. New York: Springer, 2005, with Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

[7] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, April 1976.

[8] J. Lember and A. Koloydenko, "The Adjusted Viterbi training for hidden Markov models," *Bernoulli*, vol. 14, no. 1, pp. 180–206, February 2008.

[9] A. Caliebe and U. Rösler, "Convergence of the maximum a posteriori path estimator in hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 48, no. 7, pp. 1750–1758, July 2002.

[10] A. Caliebe, "Properties of the maximum a posteriori path estimator in hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 52, no. 1, pp. 41–51, January 2006.

[11] J. Hayes, T. Cover, and J. Riera, "Optimal sequence detection and optimal symbol-by-symbol detection: Similar algorithms," *Communications, IEEE Transactions on*, vol. 30, no. 1, pp. 152–157, January 1982. [Online]. Available: http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1095391

[12] J. A. Kogan, "Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm," in *Image models (and their speech model cousins) (Minneapolis, MN, 1993/1994)*, ser. IMA Vol. Math. Appl. New York: Springer, 1996, vol. 80, pp. 115–130.

[13] J. Lember and A. Koloydenko, "Adjusted Viterbi training: A proof of concept," *Probab. Eng. Inf. Sci.*, vol. 21, no. 3, pp. 451–475, July 2007.

[14] A. Koloydenko, M. Käärik, and J. Lember, "On adjusted Viterbi training," *Acta Appl. Math.*, vol. 96, no. 1-3, pp. 309–326, May 2007.

[15] A. Koloydenko and J. Lember, "Infinite Viterbi alignments in the two-state hidden Markov models," *Acta Comment. Univ. Tartu. Math.*, vol. 12, pp. 109–124, December 2008, proc. 8th Tartu Conf. Multivariate Statist. June 2007.

[16] R. Durbin, S. Eddy, K. A., and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[17] S. Lin and D. J. Costello Jr., *Error Control Coding: Fundamental and Applications*, ser. Computer Applications in Electrical Engineering, F. F. Kuo, Ed. Englewood Cliffs, New Jersey 07632: Prentice-Hall, Inc., 1983.

[18] P. J. Bickel, Y. Ritov, and T. Rydén, "Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models," *Ann. Statist.*, vol. 26, no. 4, pp. 1614–1635, August 1998.

[19] J. Lember and A. Koloydenko, "Adjusted Viterbi training for hidden Markov models," School of Mathematical Sciences, Nottingham University, arXiv:0709.2317v1 [math.ST], Tech. Rep. 07-01, January 2007, posted on institutional website in April 2006.

Jüri Lember was born in 1968 in Tallinn, Estonia. He received the diploma and M.Sc. degrees in mathematical statistics in 1992 and 1994, respectively, from the University of Tartu, Estonia. He received the Ph.D. degree in mathematics from the University of Tartu, Estonia, in 1999.

He completed his compulsory military service in 1987–1989, and was a Postdoctoral Research Fellow in the Institute of Mathematical Statistics, University of Tartu, in 1999–2000. He held a Postdoctoral Research position in Eurandom, The Netherlands, in 2001–2003. Since 2003, he has been a Lecture and a Senior Researcher in the Institute of Mathematical Statistics, University of Tartu. His scientific interests include probability theory, theoretical statistics, information theory, hidden Markov models and speech recognition.

Dr. Lember has been a member of the Estonian statistical society as well as Estonian mathematical society since 2003. He has been awarded Estonian Science foundation grants for periods of 2004–2007 and 2008–2011.

Alexey Koloydenko received the B.S. degrees in physics and mathematics (with information systems minor) in 1994 from the Voronezh University, Russian Federation and Norwich University, USA, respectively. He received in 1996 the M.S.(tech.) degree in physics and radio-electronics from the Voronezh State University, Russian Federation, and the M.S. degree in mathematics and statistics from the University of Massachusetts at Amherst, USA. He received the Ph.D. degree in mathematics and statistics from the University of Massachusetts at Amherst, USA, in 2000.

He held Postdoctoral Research and Teaching positions with the Department of Mathematics and Statistics of the University of Massachusetts at Amherst, Statistics and Computer Science Departments of the University of Chicago, and Eurandom, The Netherlands, in 2000, 2001–2002, and 2002–2005, respectively. He was a Lecturer in Statistics at the University of Nottingham, UK, in 2005-2008, and has been a lecturer in Probability Theory and Statistics at Royal Holloway, University of London. His research interests include statistical processing and analysis of images, diffusion weighted MRI, algebraic aspects of probability theory and statistics, and hidden Markov models.

Dr. Koloydenko was a member of the Pattern Analysis, Statistical Modelling and Computational Learning European network (PASCAL) in 2004–2008, and he has been a member of the British Machine Vision Association and Society for Pattern Recognition since 2009.