Available online at www.sciencedirect.com

SciVerse ScienceDirect

Stochastic Processes and their Applications 122 (2012) 3312–3341

stochastic processes and their applications

www.elsevier.com/locate/spa

Asymptotic risks of Viterbi segmentation

K. Kuljus[a,*], J. Lember[b]

[a] *Swedish University of Agricultural Sciences, Centre of Biostochastics, 901 83 Umeå, Sweden*
[b] *Tartu University, J. Liivi 2 - 507, Tartu 50408, Estonia*

Received 26 January 2012; received in revised form 16 May 2012; accepted 16 May 2012
Available online 5 June 2012

Abstract

We consider the maximum likelihood (Viterbi) alignment of a hidden Markov model (HMM). In an HMM, the underlying Markov chain is usually hidden and the Viterbi alignment is often used as the estimate of it. This approach will be referred to as the Viterbi segmentation. The goodness of the Viterbi segmentation can be measured by several risks. In this paper, we prove the existence of asymptotic risks. Being independent of data, the asymptotic risks can be considered as the characteristics of the model that illustrate the long-run behavior of the Viterbi segmentation.

Keywords: Hidden Markov model; Viterbi alignment; Segmentation; Asymptotic risk

1. Introduction

The present paper deals with asymptotics of the Viterbi segmentation. Before we can present the main results, we introduce the segmentation problem and different risks for measuring goodness of segmentations.

1.1. Notation

Let $Y = \{Y_t\}_{t=-\infty}^{\infty}$ be a double-sided stationary Markov chain with states $S = \{1, \ldots, |S|\}$ and irreducible aperiodic transition matrix $\left(p_{i,j}\right)$. Let $X = \{X_t\}_{t=-\infty}^{\infty}$ be a double-sided process such that: (1) given $\{Y_t\}$, the random variables $\{X_t\}$ are conditionally independent; (2) the

* Corresponding author.
E-mail address: kristi.kuljus@slu.se (K. Kuljus).

0304-4149/$ - see front matter
doi:10.1016/j.spa.2012.05.014

distribution of X_j depends on $\{Y_t\}$ only through Y_j. The process X is sometimes called a *hidden Markov process* (HMP) and the pair (Y, X) is referred to as a *hidden Markov model* (HMM). The name is motivated by the assumption that the process Y, which is sometimes called the *regime*, is non-observable. The distributions $P_s := \mathbf{P}(X_1 \in \cdot | Y_1 = s)$ are called *emission distributions*. We shall assume that the emission distributions are defined on a measurable space $(\mathcal{X}, \mathcal{B})$, where $\mathcal{X}$ is usually $\mathbb{R}^d$ and $\mathcal{B}$ is the Borel σ-algebra. Without loss of generality we shall assume that the measures P_s have densities f_s with respect to some reference measure μ. Our notation differs from the one used in the HMM literature, where usually X stands for the regime and Y for the observations. Since our study is mainly motivated by statistical learning, we would like to be consistent with the notation used there, and keep X for observations and Y for latent variables.

Given a set $\mathcal{A}$ and integers m and $n, m < n$, we shall denote any $(n - m + 1)$-dimensional vector with all the components in $\mathcal{A}$ by $a_m^n := (a_m, \ldots, a_n)$. When $m = 1$, it will be often dropped from the notation and we write $a^n \in \mathcal{A}^n$.

HMMs are widely used in various fields of applications, including speech recognition [29,14], bioinformatics [19,8], language processing [28], image analysis [27] and many others. For general overview about HMMs, we refer to [5,10].

1.2. Segmentation and risks

The *segmentation or decoding* problem consists in estimating the unobserved realization of the underlying Markov chain $Y_1, \ldots, Y_n$ given n observations $x^n = (x_1, \ldots, x_n)$ from a hidden Markov model. Formally, we are looking for a mapping $g : \mathcal{X}^n \to S^n$ called a *classifier*, that maps every sequence of observations into a state sequence. In [25,26], a general approach of segmentation problem in the framework of statistical pattern theory was introduced. Let us here give a brief overview of the main concepts.

For finding the best classifier g, it is natural to set to every state sequence $s^n \in S^n$ into correspondence a measure of goodness of s^n, referred to as the *risk of* s^n. Let us denote the risk of s^n for a given x^n by $R(s^n|x^n)$. The solution of the segmentation problem is then a state sequence with minimum risk. In pattern recognition theory, a risk is specified via a *loss function* $L : S^n \times S^n \to [0, \infty]$, where $L(y^n, s^n)$ measures the loss when the actual state sequence is y^n and the estimated sequence is s^n. For any state sequence $s^n \in S^n$ the risk is then

$$R(s^n|x^n) := E[L(Y^n, s^n)|X^n = x^n] = \sum_{y^n \in S^n} L(y^n, s^n)\mathbf{P}(Y^n = y^n|X^n = x^n).$$

One common loss function is the so-called *symmetrical or zero–one loss* L_∞ defined as

$$L_\infty(y^n, s^n) = \begin{cases} 1, & \text{if } y^n \neq s^n; \\ 0, & \text{if } y^n = s^n. \end{cases}$$

We shall denote the corresponding risk by R_∞. With this loss, $R_\infty(s^n|x^n) = \mathbf{P}(Y^n \neq s^n|X^n = x^n)$, thus the minimizer of $R_\infty(\cdot|x^n)$ is a sequence with maximum posterior probability, called the *Viterbi alignment*. The name is inherited from the dynamic programming algorithm (Viterbi algorithm) used for finding it. Let v stand for the Viterbi alignment, i.e. $v(x^n) = \arg\max_{s^n} p(s^n|x^n)$, where $p(s^n|x^n) = \mathbf{P}(Y^n = s^n|X^n = x^n)$. Obviously, the Viterbi alignment is not necessarily unique. The Viterbi alignment minimizes also the following risk:

$$\bar{R}_\infty(s^n|x^n) := -\frac{1}{n} \ln p(s^n|x^n). \tag{1}$$

The log-likelihood based risk (1) is often preferable to use since, as we shall see later, it allows various generalizations.

Another common classifier is based on the pointwise loss function

$$L_1(y^n, s^n) = \frac{1}{n}\sum_{t=1}^{n} l(y_t, s_t), \tag{2}$$

where $l(y_t, s_t) \geq 0$ is the loss of classifying the t-th symbol y_t as s_t. Let us denote the corresponding risk by $R_1(s^n|x^n)$. Then

$$R_1(s^n|x^n) = \frac{1}{n}\sum_{t=1}^{n} R_1^t(s_t|x^n),$$

where $R_1^t(s|x^n) := \sum_{y\in S} l(y, s)p_t(y|x^n)$ and $p_t(y|x^n) := \mathbf{P}(Y_t = y|X^n = x^n)$. Most frequently, $l(s, s') = I_{\{s\neq s'\}}$, and then $R_1(s^n|x^n)$ just counts the expected number of misclassified symbols given that the data are x^n and the sequence s^n is used for segmentation. For that l,

$$R_1(s^n|x^n) = 1 - \frac{1}{n}\sum_{t=1}^{n} p_t(s_t|x^n). \tag{3}$$

The minimizer of (3) over all possible state sequences is called the *pointwise maximum a posteriori* (PMAP) alignment. In statistics, especially spatial statistics and image analysis, the PMAP-classifier is also known as *marginal posterior mode* [33] or *maximum posterior marginals* [31] estimator. In applications, the terms *optimal symbol-by-symbol detection* [13], *symbol-by-symbol MAP estimation* [30] and *MAP-state estimation* [2] have been used. The Viterbi and the PMAP-classifier – the so-called standard classifiers – are by far the two most popular classifiers used in practice. However, despite the fact that the PMAP-classifier is optimal in the sense of maximizing the expected number of correctly estimated states, a PMAP-path might at the same time have very low or even zero probability (see Fig. 1). That explains partially why the Viterbi classifier is often preferred over the PMAP-classifier in practice. Since the PMAP-classifier can result in paths with zero probability, one could constrain the PMAP-decoder to admissible paths, i.e. paths with positive posterior probability. In other words, the R_1-risk can simply be minimized over the admissible paths:

$$\min_{s^n: p(s^n|x^n)>0} R_1(s^n|x^n) \Leftrightarrow \max_{s^n: p(s^n|x^n)>0} \sum_{t=1}^{n} p_t(s_t|x^n). \tag{4}$$

We shall also consider the risk

$$\bar{R}_1(s^n|x^n) := -\frac{1}{n}\sum_{t=1}^{n} \ln p_t(s_t|x^n).$$

The risks R_1 and $\bar{R}_1$ are closely related. Minimizing (3) over all possible state sequences is clearly equivalent to minimizing $\bar{R}_1$, but this is not necessarily so for restricted minimization: the solution of (4) is not necessarily the solution of the following problem:

$$\min_{s^n: p(s^n|x^n)>0} \bar{R}_1(s^n|x^n) \Leftrightarrow \max_{s^n: p(s^n|x^n)>0} \sum_{t=1}^{n} \ln p_t(s_t|x^n). \tag{5}$$

The solution of (5) is sometimes called as the *posterior Viterbi decoding* (PVD) [11]. Both problems – (4) and (5) – can be solved by dynamical programming algorithm very close to the one of Viterbi [25].

Although the constrained problems (4) and (5) both result in paths with positive probability, this probability can still be very small. This suggests to consider instead of (4) the following more general penalized optimization problem:

$$\min_{s^n}\big[R_1(s^n|x^n) + Ch(s^n)\big], \tag{6}$$

where C is a positive constant and $h(s^n)$ is some penalty term. For example, $h(s^n)$ can be $-\frac{1}{n}\ln p(s^n)$ and then (6) with $C = 1$ is equivalent to

$$\max_{s^n}\left[\sum_{t=1}^{n} p_t(s^n|x^n) + \ln p(s^n)\right],$$

i.e. the goal is to minimize the expected number of errors and maximize the a priori path probability simultaneously. When $h(s^n) = I_{\{p(s^n)=0\}}$ and C is sufficiently large, then (6) becomes (4). Going a step further, the penalty $h(s^n)$ in (6) could depend on the observation x^n as well. In particular, replacing $-\frac{1}{n}\ln p(s^n)$ with $\bar{R}_\infty(s^n|x^n)$ would give the following minimization problem:

$$\min_{s^n} R_{C+1}(s^n|x^n), \quad \text{where } R_{C+1}(s^n|x^n) := R_1(s^n|x^n) + C\bar{R}_\infty(s^n|x^n). \tag{7}$$

Clearly, the R_1-risk in (6) and (7) could be replaced by the $\bar{R}_1$-risk, hence the $\bar{R}_1$ counterpart of (7) is

$$\min_{s^n} \bar{R}_{C+1}(s^n|x^n), \quad \text{where } \bar{R}_{C+1}(s^n|x^n) := \bar{R}_1(s^n|x^n) + C\bar{R}_\infty(s^n|x^n). \tag{8}$$

The solutions of (7) and (8) are in general not the same, but both problems naturally interpolate between the two standard alignments. For C big enough the minimizer of both risks is the Viterbi alignment. The weight of the $\bar{R}_\infty(s^n|x^n)$-risk decreases when C decreases, and for $C = 0$ the minimizer of both is the PMAP-alignment. It is important to note that for any $C > 0$, both problems guarantee admissible solutions. For a detailed discussion of the obtained alignment class, see [25], where it is also shown that problems (7) and (8) can be solved by a dynamic programming algorithm that is similar to the Viterbi algorithm and easy to implement.

The risk $\bar{R}_{C+1}(s^n|x^n)$ has a nice interpretation. Namely, as a remedy against vanishing probabilities, Rabiner [29] mentions in his celebrated tutorial about HMMs maximization of the expected number of correctly decoded (overlapping) blocks of length two or three rather than single states. He proposes to minimize the following risk:

$$R_k(s^n|x^n) := 1 - \frac{1}{n-k+1}\sum_{t=1}^{n-k+1} p_t(s_t^{t+k-1}|x^n), \tag{9}$$

where $k \in \mathbb{N}$ is the block length ($k = 1$ corresponds to the R_1-risk in (3)) and $p_t(s_t^{t+k-1}|x^n) := \mathbf{P}(Y_t^{t+k-1} = s_t^{t+k-1}|X^n = x^n)$. The risk in (9) derives from the loss function

$$L_k(y^n, s^n) := \frac{1}{n-k+1}\sum_{t=1}^{n-k+1} I_{\{s_t^{t+k-1} \neq y_t^{t+k-1}\}}. \tag{10}$$

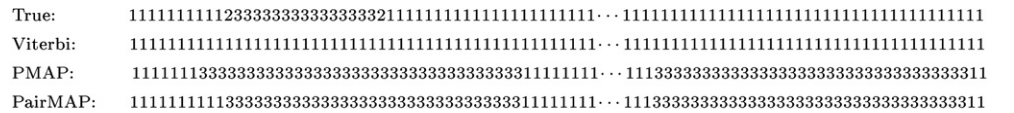

Fig. 1. Decoding performance of different classifiers.

Intuitively, one might expect that the probability of a path obtained in such a way increases as k increases. However, this is not true in general. Moreover, a path obtained in such a way is not guaranteed to be admissible [25, Appendix A]. When, however, in definition (9) the sum is replaced by the product (or equivalently, $p_t(s_t^{t+k-1}|x^n)$ is replaced by $\ln p_t(s_t^{t+k-1}|x^n)$), then, as shown in [25], the solution of the modified problem is basically the same as the minimizer of $\bar{R}_k(s^n|x^n)$. That gives another nice interpretation and justification of the risk $\bar{R}_{C+1}(s^n|x^n)$.

Example. To illustrate the introduced classifiers, we consider the following practical example taken from [26]. A long sequence has been simulated from an HMM with transition probabilities $p_{i,j}$ and stationary initial distribution π given by

$$(p_{i,j}) = \begin{pmatrix} 0.99 & 0.01 & 0 \\ 0.3 & 0.3 & 0.4 \\ 0 & 0.02 & 0.98 \end{pmatrix}, \qquad \pi = \begin{pmatrix} 0.5882 \\ 0.0196 \\ 0.3922 \end{pmatrix},$$

and the emission distributions given by

$$p_1 = (0.3, 0.2, 0.2, 0.3), \qquad p_2 = (0.1, 0.3, 0.3, 0.3),$$
$$p_3 = (1/6, 1/6, 1/6, 1/2).$$

Note that the direct transitions between the first and third state are not allowed. The sequence is then decoded with the following classifiers:

(a) Viterbi,
(b) PMAP,
(c) PairMAP: the minimizer of $R_2(s^n|x^n)$ as in (9),
(d) PVD: the solution of (5),
(e) Constrained PMAP (CnstrPMAP): the solution of (4),
(f) HybridK2: the minimizer of $\bar{R}_2(s^n|x^n)$ as in (8).

Fig. 1 gives two fragments of the ground truth and decoded outputs, the complete output can be found in [18]. The example illustrates the typical tendency of the Viterbi classifier to get stuck in a state of sizable probability and therefore systematically misclassify less frequent states. In its extreme form, this behavior would predict a constant sequence of *Heads* even when the *Heads* side of the coin is only slightly heavier than the *Tails* one. The inaccuracy problem can be alleviated by exploiting the PMAP-classifier. However, as can be noted from the forbidden transitions between states 1 and 3 along the PMAP-path in Fig. 1, the PMAP-path is inadmissible in the present example.

The constrained PMAP-path is admissible, i.e. the forbidden zero-probability transitions are not present any more. Otherwise, as expected, it is very similar to the PMAP-path. Also the PVD-path is admissible, but it differs slightly from the constrained PMAP-path.

The PairMAP-classifier looks similar to the PMAP and is also inadmissible. Based on these fragments, all the four paths – PMAP, PairMAP, PVD and Constrained PMAP – are rather close to each other and remarkably differ from the true realization as well as the Viterbi path. From Fig. 1 it can be seen that HybridK2 seems to perform best: it is admissible, in comparison with the four previous ones it is remarkably more accurate and, as desired, it really looks to have the properties of both the Viterbi and the PMAP-classifier.

1.3. Asymptotic risks, the main results and organization of the paper

Given a risk R, the quantity $R(g, x^n) := R(g(x^n)|x^n)$ measures the goodness of a classifier g, when it is applied to the observations x^n. When g is optimal in the sense of risk, then $R(g, x^n) = \min_{s^n} R(s^n|x^n) =: R(x^n)$. We are interested in the convergence of random variables $R(g, X^n)$. Here for any n, the classifier g is supposed to be obtained using the same *decoding or classifying method*, like the Viterbi, the PMAP, the Rabiner k-block or something else. If the limits – *asymptotic risks* – exists, then they are constants that all depend on the model and characterize the goodness of the segmentation method. If, for example, R_1 is the limit of $R_1(g, X^n)$ and R_1^* is the limit of $R_1(X^n)$, then the difference $R_1 - R_1^*$ shows how well the decoding method g performs the segmentation in the long run in the sense of R_1-risk in comparison to the best possible classifier in R_1-sense. Using the method g instead of the best possible classifier for R_1, which is the PMAP-classifier, causes for large n about $(R_1 - R_1^*)n$ additional classification errors. Thus, every asymptotic risk is a characterization of a decoding method, and comparing several asymptotic risks of different methods gives us insight into how different these methods are. If all asymptotic risks are very close to each other for two different methods, then in the long run these methods perform similarly, even if the corresponding alignments can be visually very different. On the other hand, if for two methods, say the Viterbi and the PMAP, most of the asymptotic risks differ largely, then one should carefully examine which segmentation method to choose. In practice, a decoding method is often expected to perform well in many aspects, for example making possibly less errors and at the same time having a large likelihood. This means that for both candidate methods several risks should be measured — not just the risk the method is optimal for. As explained above, seeking a compromise leads often to the hybrid classifiers as in (7) or (8).

Of course, for given data, the main objects of interest would be several non-asymptotic risks $R(g, x^n)$ of different classifiers. However, when one is interested in the decoding methods rather than in the best alignments for the data, then the data effect should be eliminated. Since asymptotic risks depend solely on the model and the decoding method, they are the right object to look at in this case. Given the model, one could in principle find them theoretically, but the convergence results allow to estimate the asymptotic risks also by independent simulations.

Another motivation for studying asymptotic risks comes from the large deviation inequalities. For a given loss function L and classifier g, the quantity of actual interest is typically the *actual loss* $L(Y^n, g(X^n))$, which for hidden Y^n cannot be measured directly. Let $R(g, X^n) = E[L(Y^n, g(X^n))|X^n]$ be the corresponding risk and $R(g)$ be the corresponding asymptotic risk. The existence of $R(g)$ usually implies that the actual loss converges to the asymptotic risk as well:

$$L(Y^n, g(X^n)) \to R(g) \quad \text{a.s.} \tag{11}$$

In fact, convergence (11) implies the convergence of risks, see Lemma 3.1. Suppose now that besides (11) the following large deviation inequality holds:

$$\mathbf{P}\big(|L(Y^n, g(X^n)) - R(g)| > \epsilon\big) \le \exp[-I_g(\epsilon)n], \quad n > N(\epsilon), \tag{12}$$

where $I_g(\epsilon) > 0$ is a rate function. The inequalities like (12) are most useful in statistical learning since they provide confidence intervals (generalization bounds) for the unknown risk. However, observe that when in supervised learning the actual loss is known and the large deviation inequalities like (12) are used to estimate the unknown $R(g)$, then in our unsupervised case the theoretical asymptotic risk $R(g)$ can be used to find confidence bounds for the unknown actual loss $L(Y^n, g(X^n))$. For the Viterbi classifier and the R_1-risk, the large deviation bound (12) is proved in [12] using the large deviation result for regenerative processes. We conjecture that the bounds like (12) hold for other classifiers and loss functions as well. Of course, inequality (12) presumes the existence of $R(g)$, thus proving the existence of $R(g)$ is the first step towards a more general result.

The present paper deals mostly with convergence of risks of the Viterbi classifier. The main reasons are the following. First, as explained above, the Viterbi classifier is probably the most popular classifier in practice. This means that it is important to study all risks of the Viterbi alignment, because every risk is a characterization of this popular classifier. The risk $R_1(v, X^n)$ (as in (3)) shows the expected number of incorrectly estimated states, $R_k(v, X^n)$ (as in (9)) shows the expected number of incorrectly estimated k-tuples, $\bar{R}_1(v, X^n)$ gives the logarithmic counterpart of the R_1-risk etc. As explained above, the risks are especially useful when comparing the Viterbi with other alignments. Thus, to prove the convergence of risks for various decoding methods, the Viterbi classifier is a natural choice to start with. The second reason is a consequence of the first one — due to its popularity, the Viterbi classifier is also the most studied one. In a series of papers [4,3, 24,23,17,6] the existence and regenerativity of the Viterbi process has been proved. The Viterbi process will be defined in Section 2.2. Right now let us just mention that it is an S-valued stochastic process that is in a sense a limit of the random vectors $v(X^n)$ as n grows. The analysis in the present paper as well as in the above-mentioned paper [12] is based on the results of [24,23,17], where the Viterbi process is constructed piecewise. The piecewise construction entails several important properties of the process including regenerativity. These properties are crucial for proving the main convergence results.

As explained above, it is very informative to compare asymptotic risks of a decoding method to the ones of other methods. Therefore, the convergence of risks of other classifiers besides the Viterbi is equally important. In particular, one could study convergence of different risks for the corresponding optimal classifiers, i.e. the convergence of $R(X^n)$ for various risks. The convergence $R_1(X^n) \to R_1^*$, where R_1^* is a constant, was proved in [21,20]. For the sake of completeness, we recall the result in Section 2.3 (Theorem 2.5). Since the minimizer of the R_1-risk is not the Viterbi classifier, the proof of the existence of R_1^* is based on the exponential forgetting and it differs from the rest of the proofs that (since they are all about the Viterbi classifier) are more or less based on the ergodic properties of the Viterbi process. This implies that although we have established a unified risk-based framework for several classifiers, there is no universal method known yet to prove the convergence of general risks, and every optimal classifier needs a special treatment. In particular, the convergence of $R_{C+1}(X^n)$ or $\bar{R}_{C+1}(X^n)$ (recall (7) and (8)) as well as of many other risks has not yet been proved, although it is reasonable to conjecture that these convergences hold. To prove this conjecture might be rather challenging, because the alignment process corresponding to $R_{C+1}(X^n)$ or $\bar{R}_{C+1}(X^n)$, if it exists, might not have such good properties as the Viterbi process. Therefore, as a first step towards more general

results, in this paper we prove the convergence of the Viterbi risks, leaving the other convergences for the future research.

The paper is organized as follows. In Section 2, some preliminary results are introduced. Since we are going to use coupling of regenerative processes, in Section 2.1 some basics of regenerative processes as well as some coupling results adapted from the book by Thorisson [32] are recalled. Section 2.2 presents the main results concerning the construction of the Viterbi process. The piecewise construction under general assumptions is rather technical (see [24,17]). However, when it is performed, the regenerativity of the Viterbi process as well as the ergodicity of the double-sided Viterbi process easily follow. Although the results in Section 2.2 are mostly of preliminary nature, the main convergence result, Theorem 2.3, is interesting in its own rights and it can be used in many other purposes. The use of Theorem 2.3 is well demonstrated in Section 3, where it implies almost immediately the convergence of $R_k(v, X^n)$. In Section 2.3, the necessary results about exponential smoothing are recalled.

The main results of this paper are the convergences of $R_k(v, X^n)$, $\bar{R}_1(v, X^n)$ and $\bar{R}_\infty(v, X^n) = \bar{R}_\infty(X^n)$ to constant limits almost surely (Theorems 3.1, 4.1 and 5.1, respectively). Obviously, from these convergences also the convergence of $R_{C+1}(v, X^n)$ and $\bar{R}_{C+1}(v, X^n)$ follows. These results without the proofs are also presented in Chapter 3 of [26]. Section 3 deals with the convergence of the R_k-risk. We prove that the actual loss $L_k(Y^n, v(X^n))$ as well as $R_k(v, X^n)$ converge to a constant R_k almost surely. In Section 4, the convergence of the $\bar{R}_1$-risk for the Viterbi and PMAP-classifier is proved (Theorem 4.1 and Corollary 4.2, respectively). The proof of Theorem 4.1 is the most technical one, involving the ergodic properties of the Viterbi process as well as the exponential forgetting. The last main result, the convergence of $\bar{R}_\infty(X^n)$, is proved in Section 5.

2. Preliminary results

2.1. Regenerativity

We are following the coupling approach developed by Thorisson in [32]. One of the main instruments we are going to use is that any regenerative process can be successfully coupled with a stationary and ergodic regenerative process (Theorem 2.1). With a successful coupling, a general pathwise limit theorem for the Viterbi alignment (Theorem 2.3) can be proved. This is the main preliminary result and it can be used for many other purposes besides proving the convergence of risks.

Let $Z = \{Z_t\}_{t=1}^\infty$ in $(\Omega, \mathcal{F}, \mathbf{P})$ be a $\mathcal{Z} := \mathbb{R}^d$-valued classical regenerative process with respect to the renewal process $S = \{S_t\}_{t=0}^\infty$ (see, e.g., Chapter 10 in [32]). Following the notation in [32], we shall denote the regenerative process by (Z, S). Let $T_1 := S_1 - S_0$. The regenerative process (Z, S) is *positive recurrent* if $ET_1 < \infty$ and *aperiodic* if T_1 is aperiodic, i.e. $\mathbf{P}(T_1 \in a\mathbb{N}) < 1$ for every $a > 1$. A pair (Z', S') is a *version* of the regenerative process (Z, S) if it is also regenerative and $\theta_{S_0}(Z, S) \stackrel{D}{=} \theta_{S'_0}(Z', S')$, where θ_t is a shift operator: $\theta_t(x_1, x_2, \ldots) = (x_{t+1}, x_{t+2}, \ldots)$, and $\stackrel{D}{=}$ means equal in law. The version $(Z^o, S^o) := \theta_{S_0}(Z, S)$ of (Z, S) is a *zero-delayed* regenerative process. Thus, $S_0^o = 0$ and $S_1^o = T_1$. Recall that (Z, S) is stationary if $\theta_t(Z, S)$ has the same distribution as (Z, S). If (Z, S) is positive recurrent regenerative, then there exists a stationary version (Z^*, S^*) of this process such that the distribution of the delay length S_0^* is given by

$$\mathbf{P}(S_0^* = k) = \frac{1}{ET_1}\mathbf{P}(T_1 > k), \quad k \geq 0,$$

and for every $\sigma(\mathcal{Z}^\infty)$-measurable function $g : \mathcal{Z}^\infty \to \mathbb{R}$ the following inequality holds:

$$Eg(Z_1^*, Z_2^*, \ldots) = \frac{1}{ET_1} E\left[\sum_{t=0}^{T_1-1} g(\theta_t(Z^o))\right], \tag{13}$$

see, e.g., Theorems 2.1 and 2.2 of Chapter 10 in [32] or Theorem 6.1 in [15].

The following version of Theorem 3.3 of Chapter 10 in [32] states that an aperiodic positive recurrent regenerative process can be successfully coupled with a stationary ergodic process.

Theorem 2.1. *Let (Z, S) be an aperiodic and positive recurrent regenerative process. Let (Z^*, S^*) be a stationary version of it. Then the following statements hold:*

(a) *The space $(\Omega, \mathcal{F}, \mathbf{P})$ can be extended to support a finite random time T and a copy Z' of Z^* such that (Z, Z', T) is a successful exact coupling of Z and Z^*, i.e.*

$$\theta_T Z = \theta_T Z', \quad \text{where } Z' \stackrel{D}{=} Z^*.$$

(b) *The processes Z and Z' are ergodic.*

Proof. The process Z is aperiodic, which means that T_1 is a lattice with span 1. Since (Z, S) and (Z^*, S^*) are discrete, the random variables S_0 and S_0^* are $\mathbb{Z}$-valued. So the assumptions of Theorem 3.3 of Chapter 10 in [32] are fulfilled. The claim (a) is claim (a) of that theorem, the ergodicity of Z follows from claim (d) of that theorem. Finally, the process Z', being a stationary version of Z, is also an aperiodic regenerative process with S_0' being $\mathbb{Z}$-valued. Hence it satisfies the same assumptions and is therefore also ergodic. □

Corollary 2.1. *Let (Z, S) be an aperiodic and positive recurrent regenerative process, and let (Z^*, S^*) be a stationary version of it. Let $g : \mathcal{Z}^\infty \to \mathbb{R}$ be such that $E|g(Z_1^*, Z_2^*, \ldots)| < \infty$. Then*

$$\frac{1}{n}\sum_{t=1}^{n} g(Z_t, Z_{t+1}, \ldots) \to E[g(Z_1^*, Z_2^*, \ldots)] \quad \text{a.s. and in } L_1. \tag{14}$$

Proof. Let us extend the space $(\Omega, \mathcal{F}, \mathbf{P})$ so that the statements of Theorem 2.1 hold. Then the process Z' is stationary and ergodic having the same distribution as Z^*. By Birkhoff's ergodic theorem (see p. 296 in [9]),

$$\frac{1}{n}\sum_{t=1}^{n} g(Z_t', Z_{t+1}', \ldots) \to E[g(Z_1', Z_2', \ldots)] = E[g(Z_1^*, Z_2^*, \ldots)] \quad \text{a.s. and in } L_1. \tag{15}$$

Since the original process Z can be successfully coupled with Z', it holds for almost every realization of Z and Z' that they differ at the finite beginning only. Since for a pathwise limit the beginning does not matter, we immediately get the almost sure convergence of (14). The L_1-convergence follows from applying Scheffe's lemma separately to $g^+(Z_t, Z_{t+1}, \ldots)$ and $g^-(Z_t, Z_{t+1}, \ldots)$. □

Remark. If (Z, S) is positive recurrent but not aperiodic, then Theorem 2.1 cannot be applied. However, using Theorem 2.2 of Chapter 10 in [32] and noting that aperiodicity is not used in its proof, a similar result can be obtained for shift-coupling instead of exact coupling. The process Z' can be shown to be ergodic, thus Corollary 2.1 still holds. In this paper we consider only aperiodic regenerative processes.

If $f : \mathcal{Z} \to \mathbb{R}$ is measurable, then convergence (14) together with (13) yields

$$\frac{1}{n}\sum_{t=1}^{n} f(Z_t) \to Ef(Z_1^*) = \frac{1}{ET_1} E\left[\sum_{t=1}^{T_1} f(Z_t^o)\right] = \frac{1}{ET_1} E\left[\sum_{t=S_0+1}^{S_1} f(Z_t)\right]$$
a.s. and in L_1. (16)

2.2. Infinite Viterbi alignment

2.2.1. One-sided infinite Viterbi alignment

Definition 2.1. Let for every n, $g^n : \mathcal{X}^n \to S^n$ be a classifier. We say that the sequence $\{g^n\}$ of classifiers can be *extended to infinity*, if there exists a function $g : \mathcal{X}^\infty \to S^\infty$ such that for almost every realization $x^\infty \in \mathcal{X}^\infty$ the following statement holds: for every $k \in \mathbb{N}$ there exists $m(x^\infty) \geq k$ such that for every $n \geq m$ the first k elements of $g^n(x^n)$ are the same as the first k elements of $g(x^\infty)$, i.e. $g^n(x^n)_i = g(x^\infty)_i,\ i = 1, \ldots, k$. The function g will be referred to as an *infinite alignment*.

If every observation is not classified independently, then the existence of an infinite alignment is not trivial. It often happens that adding one more observation x_{n+1} changes the alignment $g^n(x^n)$. This happens often with Viterbi or PMAP-alignments. The existence of an infinite alignment allows to study asymptotic properties of the alignment, which is usually done via the corresponding *alignment process* $\{G_t\}_{t=1}^\infty := g(X)$. Thus, the existence of an infinite alignment g means that the alignment process $g(X)$ is defined for almost every realization. We consider the existence of infinite Viterbi alignments. Under rather restrictive assumptions on HMMs, the existence of an infinite Viterbi alignment was first proved in [4]. In [24] it was proved under less restrictive assumptions. We now introduce these assumptions and the corresponding results.

Recall that f_s are the densities of $P_s = \mathbf{P}(X_1 \in \cdot | Y_1 = s)$ with respect to some reference measure μ on $(\mathcal{X}, \mathcal{B})$. For each $s \in S$, let $D_s := \{x \in \mathcal{X} : f_s(x) > 0\}$. We call a subset $C \subset S$ *a cluster* if the following conditions are satisfied:

$$\min_{j \in C} P_j(\cap_{s \in C} D_s) > 0 \quad \text{and} \quad \max_{j \notin C} P_j(\cap_{s \in C} D_s) = 0.$$

Hence, a cluster is a maximal subset of states such that $D_C = \cap_{s \in C} D_s$, the intersection of the supports of the corresponding emission distributions, 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. If $|S| = 2$, then S is the only cluster possible, because otherwise the underlying Markov chain would cease to be hidden. The existence of C implies the existence of a set $\mathcal{X}_o \subset \cap_{s \in C} D_s$ such that the conditions given in the definition below hold for some $\epsilon > 0$ and $M < \infty$. For proof, see [24].

Definition 2.2. Let $\mathcal{X}_o \subset \cap_{s \in C} D_s$ with $\mu(\mathcal{X}_o) > 0$ be a set such that $\forall x \in \mathcal{X}_o$ the following statements hold for some $\epsilon > 0$ and $M < \infty$: (i) $\min_{s \in C} f_s(x) > \epsilon$; (ii) $\max_{s \in C} f_s(x) < M$; (iii) $\max_{s \notin C} f_s(x) = 0$.

The following two assumptions on HMMs are needed for the existence of an infinite Viterbi alignment.

A1 (*cluster-assumption*) There exists a cluster $C \subset S$ such that the sub-stochastic matrix $R = (p_{i,j})_{i,j\in C}$ is primitive, i.e. there is a positive integer r such that the rth power of R is strictly positive.

A2 For each state $l \in S$,

$$P_l\left(\left\{x \in \mathcal{X} : f_l(x)p_l^* > \max_{s,s\neq l} f_s(x)p_s^*\right\}\right) > 0, \quad p_l^* = \max_j p_{j,l},\ \forall l \in S. \tag{17}$$

The cluster assumption A1 is often met in practice. It is clearly satisfied if all elements of the matrix P are positive. Since any irreducible aperiodic matrix is primitive, the assumption A1 is also satisfied if the densities f_s satisfy the following condition: for every $x \in \mathcal{X}$, $\min_{s\in S} f_s(x) > 0$, i.e. for all $s \in S$, $D_s = \mathcal{X}$. Note that A1 implies the aperiodicity of Y, but not vice versa. The assumption A2 is more technical in nature. In [17] it was shown that for a two-state HMM, (17) always holds for one state, and this is sufficient for the infinite Viterbi alignment. Hence, for the case $|S| = 2$, A2 can be relaxed. Another possibilities for relaxing A2 are discussed in [24,23]. To summarize: we believe that the cluster assumption A1 is essential for HMMs, while the assumption A2, although natural and satisfied for many models, can be relaxed. For more general discussion about these assumptions, see [24,23,21,17].

In the following, let $\tilde{V}^n = v^n(X^n)$, where v^n is a finite Viterbi alignment. Consider a set $\mathcal{X}_o$ satisfying Definition 2.2. Let U_t and W_t be stopping times defined as

$$\begin{aligned} W_t &= \min\{\tau \geq t + r + 1 : X_{\tau-r}^{\tau} \in \mathcal{X}_0^{r+1}\}, \\ U_t &= \max\{\tau \leq t - r - 1 : X_{\tau}^{\tau+r} \in \mathcal{X}_0^{r+1}\}. \end{aligned} \tag{18}$$

The results of the present paper are largely based on the following theorem, which has been proved in [24,23]. See also Lemma 2.1 in [12].

Theorem 2.2. *Let $(X, Y) = \{(X_t, Y_t)\}_{t=1}^{\infty}$ be a one-sided ergodic HMM satisfying* A1 *and* A2. *Then there exists an infinite Viterbi alignment $v : \mathcal{X}^{\infty} \to S^{\infty}$. Moreover, the finite Viterbi alignments $v^n : \mathcal{X}^n \to S^n$ can be chosen so that the following conditions are satisfied:*

R1 *the process $Z := (X, Y, V)$, where $V := \{V_t\}_{t=1}^{\infty}$ is the alignment process, is a positively recurrent aperiodic regenerative process with respect to some renewal process $\{S_k\}_{k=0}^{\infty}$;*

R2 *there exists an integer $m > 0$ such that $S_0 > m$ and*

(1) *for all $k \geq 0$ such that $S_k + m \leq n$, $\tilde{V}_t^n = V_t$ for all $t \leq S_k$,*

(2) *$S_k - S_{k-1} \geq m,\ k = 1, 2, \ldots;$*

R3 *the renewal times $\{S_k\}$ have the following property:*

(1) *if $S_k \geq t$, then $W_t \leq S_k + m$,*

(2) *if $S_k \leq t$, then $U_t > S_k - m$;*

R4 *the increments $R_k := S_{k+1} - S_k,\ k = 1, 2, \ldots,$ form and i.i.d. sequence which is furthermore independent of S_0;*

R5 *there exist $a > 0$ and $b > 0$ such that $\mathbf{P}(R_1 > t) \leq a \exp[-bt]$ for every $t \geq 0$.*

Proof. The required infinite alignment is constructed piecewise, see [24]. The regenerativity and positive recurrence is shown in Section 4 of [23]. The aperiodicity follows from the aperiodicity of Y that follows from A1. The piecewise construction guarantees R2, R3 and R4. For R5, see [12]. □

Observe that under the assumptions of Theorem 2.2, its properties hold for almost every realization $x^\infty \in \mathcal{X}^\infty$.

From now on we assume that the finite Viterbi alignments $v^n : \mathcal{X}^n \to S^n$ are chosen according to Theorem 2.2. These choices of alignments are called *consistent*. Obviously, the consistent choice becomes an issue only if the finite Viterbi alignment is not unique. In practice, the consistent choices can be obtained just by predefined tie-breaking rules. With consistent choices, the process $\tilde{Z}^n := \{(X_t, Y_t, \tilde{V}_t^n)\}_{t=1}^n$ satisfies by R2 the following property: $\tilde{Z}_t^n = Z_t$ for every $t = 1, \ldots, S_{k(n)}$, where $k(n) = \max\{k \geq 0 : S_k + m \leq n\}$.

The proof of the following theorem is based on the same argument as the proof of Theorem 3.1 of Chapter VI in [1], it is given in Appendix. Let $p \in \mathbb{N}$ and $g_p : \mathcal{Z}^p \to \mathbb{R}$ be measurable. Define for every $i = p, \ldots, n$,

$$\tilde{U}_i^n := g_p(\tilde{Z}_{i-p+1}^n, \ldots, \tilde{Z}_i^n).$$

If $i \leq S_{k(n)}$, then $\tilde{U}_i^n = U_i := g_p(Z_{i-p+1}, \ldots, Z_i)$. Finally, let

$$M_k := \max_{S_k < n \leq S_{k+1}} |\tilde{U}_{S_k+1}^n + \cdots + \tilde{U}_n^n|.$$

The random variables $M_p, M_{p+1}, \ldots$ are identically distributed, but for $p > 1$ not necessarily independent. Recall that Z^* is a stationary version of Z.

Theorem 2.3. *Let g_p be such that $EM_p < \infty$ and $E|g_p(Z_1^*, \ldots, Z_p^*)| < \infty$. Then*

$$\frac{1}{n-p+1} \sum_{i=p}^{n} \tilde{U}_i^n \to EU_p = Eg_p(Z_1^*, \ldots, Z_p^*) \quad \text{a.s. and in } L_1. \tag{19}$$

2.2.2. Double-sided infinite Viterbi alignment

Definition 2.3. Let for every $z_1, z_2 \in \mathbb{Z}$, $g_{z_1}^{z_2} : \mathcal{X}^{z_2-z_1+1} \to S^{z_2-z_1+1}$ be a classifier. We say that the set $\{g_{z_1}^{z_2}\}$ of classifiers can be *extended to infinity*, if there exists a function $g : \mathcal{X}^{\mathbb{Z}} \to S^{\mathbb{Z}}$ such that for almost every realization $x_{-\infty}^{\infty} \in \mathcal{X}^{\mathbb{Z}}$ the following statement holds: for every $k \in \mathbb{N}$ there exists $m(x_{-\infty}^{\infty}) \geq k$ such that for every $n \geq m$,

$$g_{-n}^{n}(x_{-n}^{n})_i = g(x_{-\infty}^{\infty})_i, \quad i = -k, \ldots, k.$$

The function g will be referred to as an *infinite double-sided alignment*.

Again, the existence of an infinite double-sided alignment g means that the corresponding alignment process $\{G_t\}_{t=-\infty}^{\infty} := g(X)$ is defined for almost every realization. The piecewise construction of the infinite Viterbi alignment allows the double-sided extension as well.

Theorem 2.4. *Let $(X, Y) = \{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ be a double-sided ergodic HMM satisfying* A1 *and* A2. *Then there exists an infinite Viterbi alignment $v : \mathcal{X}^{\mathbb{Z}} \to S^{\mathbb{Z}}$. Moreover, the finite Viterbi alignments $v_{z_1}^{z_2}$ can be chosen so that the following conditions are satisfied:*

RD1 *the process (X, Y, V), where $V := \{V_t\}_{t=-\infty}^{\infty}$ is the alignment process, is a positively recurrent aperiodic regenerative process with respect to some renewal process $\{S_k\}_{k=-\infty}^{\infty}$;*

RD2 *there exists a nonnegative integer $m < \infty$ such that*

(1) *for every $k \geq 0$ such that $S_k + m \leq n$, $\tilde{V}_t^n = V_t$ for all $S_0 \leq t \leq S_k$;*

(2) *$S_k - S_{k-1} \geq m$, $k \in \mathbb{Z}$;*

RD3 *the renewal times* $\{S_k\}$ *have the following property:*
(1) *if* $S_k \geq t$, *then* $W_t \leq S_k + m$,
(2) *if* $S_k \leq t$, *then* $U_t > S_k - m$;
RD4 *the increments* $R_k := S_{k+1} - S_k$, $k \in \mathbb{Z}$, *form an i.i.d. sequence;*
RD5 *there exist* $a > 0$ *and* $b > 0$ *such that* $\mathbf{P}(R_1 > t) \leq a \exp[-bt]$ *for every* $t \geq 0$;
RD6 *the mapping* v *is a stationary coding, i.e.* $v(\theta(X)) = \theta v(X)$, *where* θ *is a shift operator:* $\theta(\ldots, x_{-1}, x_0, x_1, \ldots) = (\ldots, x_0, x_1, x_2, \ldots)$.

Proof. The proof of RD1–RD5 is the same as in Theorem 2.2. Note the difference between R2 and RD2. The stationarity of v follows from the fact that the barriers in the construction of the infinite alignment are separated (Lemma 3.2 in [24]). □

In the following, the finite Viterbi alignments $v_{z_1}^{z_2}$ are chosen to be consistent. The property RD6 is important. Since X is an ergodic process, from RD6 it follows that the double-sided alignment process $V = \{V_t\}_{t=-\infty}^{\infty}$ as well as the process $\{(X_t, Y_t, V_t)\}_{t=-\infty}^{\infty}$ is an ergodic process. When Theorem 2.4 holds, it holds for almost every realization $x_{-\infty}^{\infty} \in \mathcal{X}^{\mathbb{Z}}$. Let Z^* denote the restriction of $\{(X_t, Y_t, V_t)\}_{t=-\infty}^{\infty}$ to the nonnegative integers, i.e. $Z^* = \{(X_t, Y_t, V_t)\}_{t=1}^{\infty}$. Since $(X_i, Y_i, V_i) \stackrel{D}{=} (X_j, Y_j, V_j)$ for every i and j, we have $(X_0, Y_0, V_0) \stackrel{D}{=} (X_1^*, Y_1^*, V_1^*) = Z_1^*$ and we shall often use this. Note that the one-sided Viterbi process V in R1 is not defined at time zero, therefore the random variable V_0 always implies that we are considering the double-sided and hence the stationary case.

2.3. Smoothing probabilities and convergence of $R_1(X^n)$

Let $(X, Y) = \{(X_t, Y_t)\}_{t=-\infty}^{\infty}$ be a double-sided HMM. From Levy's martingale convergence theorem it immediately follows that for every state $j \in S$ and $z, t \in \mathbb{Z}$, the limits of the smoothing probabilities $\mathbf{P}(Y_t = j|X_z^{\infty}) := \lim_n \mathbf{P}(Y_t = j|X_z^n)$ and $\mathbf{P}(Y_t = j|X_{-\infty}^{\infty}) := \lim_{z\to-\infty} \mathbf{P}(Y_t = j|X_z^{\infty})$ exist almost surely. In [21,20] it is shown that under A1 these probabilities satisfy the following exponential forgetting inequalities:

$$\|\mathbf{P}(Y_t \in \cdot|X_1^{\infty}) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\| \leq C_1 \rho^{t-1} \quad \text{a.s., } 1 \leq t, \tag{20}$$

$$\|\mathbf{P}(Y_t \in \cdot|X_1^{\infty}) - \mathbf{P}(Y_t \in \cdot|X_1^{n})\| \leq C_k' \rho^{k-t} \quad \text{a.s., } 1 \leq t \leq k \leq n, \tag{21}$$

where C_1 is a finite random variable, $\{C_k'\}$ is an almost surely finite ergodic process, and $\rho \in (0, 1)$. Here $\|\cdot\|$ stands for the total variation distance. Since the variables C_k' are obtained as a stationary coding of a stationary process $\{X_k\}$ (see [21,20]), they have the same distribution. Observe that approximation of the smoothing probabilities is usually considered in the literature under the *strong mixing condition* (Assumption 4.3.21 in [5]), which ensures exponential forgetting. Assumption A1 is more general than the strong mixing condition and also weaker than Assumption 4.3.29 in [5]. For forgetting properties, see also Section 4.3 in [5].

Let $M < \infty$ be such that $\mathbf{P}(C_k' \leq M) > 0$. Define

$$u(n) := \max\{k \leq n : C_k' \leq M\},$$

and let $u(n) = 0$, if up to n, every C_k' is larger than M. Hence, inequalities (20) and (21) imply

$$\|\mathbf{P}(Y_t \in \cdot|X_1^{n}) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\| \leq \begin{cases} C_1 \rho^{t-1} + M\rho^{u(n)-t} & \text{a.s., if } 1 \leq t \leq u(n); \\ 2, & \text{if } u(n) < t \leq n. \end{cases} \tag{22}$$

To bound the difference $n - u(n)$, note that since $\{C'_k\}$ is ergodic and $\mathbf{P}(C'_k \le M) > 0$, from Birkhoff's ergodic theorem it follows that $u(n) \to \infty$ almost surely. Let $Z_k = I_{[0,M]}(C'_k)$, i.e. $Z_k = 1$ if and only if $C'_k \le M$, otherwise $Z_k = 0$. Then $u(n) = \max\{k \le n : Z_k = 1\}$. Since $\{C'_k\}$ is ergodic, so is the 0–1 process $\{Z_k\}$, and by Birkhoff's ergodic theorem,

$$\frac{1}{n}\sum_{k=1}^{n} Z_k \to \mathbf{P}(Z_1 = 1) = \mathbf{P}(C'_k \le M) > 0 \quad \text{a.s.} \tag{23}$$

Convergence (23) holds also along the subsequence $\{u(n)\}$, therefore

$$\frac{1}{u(n)}\sum_{k=1}^{u(n)} Z_k = \frac{1}{u(n)}\sum_{k=1}^{n} Z_k \to \mathbf{P}(Z_1 = 1) \quad \text{a.s.} \tag{24}$$

Convergences (23) and (24) imply

$$\frac{n - u(n)}{\sum_{k=1}^{n} Z_k} = \frac{n - u(n)}{n}\,\frac{n}{\sum_{k=1}^{n} Z_k} \to 0 \quad \text{a.s.}$$

Hence, due to (23) it now follows that

$$\frac{n - u(n)}{n} \to 0 \quad \text{a.s.} \tag{25}$$

In what follows, we shall use the notation

$$p_t(s|x_{-\infty}^{\infty}) := \mathbf{P}(Y_t = s|X_{-\infty}^{\infty} = x_{-\infty}^{\infty}).$$

As an application of the forgetting inequalities, we show how (22) and (25) imply the convergence of $R_1(X^n)$. Recall that

$$R_1(X^n) = \frac{1}{n}\sum_{t=1}^{n} \min_{s\in S} R_1^t(s|X^n),$$
$$R_1^t(s|x^n) := E[l(Y_t, s)|X^n = x^n] = \sum_{y\in S} l(y, s)\mathbf{P}(Y_t = y|X^n = x^n).$$

The following theorem is proved in [20], we present it here for the sake of completeness only.

Theorem 2.5. *There exists a constant R_1^* such that*

$$R_1(X^n) \to R_1^* \quad \text{a.s.}$$

Proof. Define

$$R_1^t(s|X_{-\infty}^{\infty}) := E[l(Y_t, s)|X_{-\infty}^{\infty}] = \sum_{y\in S} l(y, s)\mathbf{P}(Y_t = y|X_{-\infty}^{\infty}),$$

and note that

$$|\min_s R_1^t(s|X_{-\infty}^{\infty}) - \min_s R_1^t(s|X^n)| \le A\|\mathbf{P}(Y_t \in \cdot|X^n) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\|, \tag{26}$$

where $A = \max_{y,s} l(y, s)$. The process X is ergodic, so for a constant R_1^*,

$$\frac{1}{n}\sum_{t=1}^{n} \min_s R_1^t(s|X_{-\infty}^{\infty}) \to R_1^* \quad \text{a.s. and in } L_1. \tag{27}$$

Thus, from (26), (22) and (25) it follows that

$$\begin{aligned}
&\left| R_1(X^n) - \frac{1}{n}\sum_{t=1}^{n} \min_{s\in S} R_1^t(s|X_{-\infty}^{\infty}) \right| \\
&= \left| \frac{1}{n}\sum_{t=1}^{n} \min_{s\in S} R_1^t(s|X^n) - \frac{1}{n}\sum_{t=1}^{n} \min_{s\in S} R_1^t(s|X_{-\infty}^{\infty}) \right| \\
&\le \frac{A}{n}\sum_{t=1}^{n} \|\mathbf{P}(Y_t \in \cdot|X^n) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\| \\
&\le \frac{A}{n}\sum_{t=1}^{u(n)} \|\mathbf{P}(Y_t \in \cdot|X^n) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\| + \frac{2A}{n}(n - u(n)) \\
&\le A\frac{C_1}{n}\sum_{t=1}^{u(n)} \rho^{t-1} + \frac{AM}{n}\sum_{t=1}^{u(n)} \rho^{u(n)-t} + 2A\frac{n-u(n)}{n} \\
&\le \frac{A(C_1+M)}{n}\sum_{t=0}^{\infty} \rho^t + 2A\frac{n-u(n)}{n} \to 0 \quad \text{a.s.}
\end{aligned}$$

The statement of the theorem follows from (27). □

3. Convergence of R_k-risk for Viterbi alignment

Let $k = 1, 2, \ldots$ be fixed and let $l_k : S^k \times S^k \to \mathbb{R}^+$ be a loss function acting on k-tuples. The loss function

$$L_k(y^n, s^n) := \frac{1}{n-k+1}\sum_{t=1}^{n-k+1} l_k(y_t^{t+k-1}, s_t^{t+k-1}) \tag{28}$$

generalizes simultaneously (2) (take $k = 1$) and (10) (take $l_k = I_{\{s^k \neq y^k\}}$). Observe that for $k = 1$ we previously denoted $l_1 =: l$. Let R_k be the risk corresponding to L_k.

Consider now a consistently chosen Viterbi alignment v^n. If the underlying Markov chain would not be hidden, the *actual loss of the Viterbi alignment* could be directly calculated as

$$L_k(Y^n, v^n(X^n)) = L_k(Y^n, \tilde{V}^n) = \frac{1}{n-k+1}\sum_{t=1}^{n-k+1} l_k(Y_t^{t+k-1}, \tilde{V}_t^{t+k-1}). \tag{29}$$

The conditional expectation of $L_k(Y^n, \tilde{V}^n)$ given X^n is the random variable $R_k(v, X^n) = E[L_k(Y^n, \tilde{V}^n)|X^n]$. Since S is finite and $l_k : S^k \times S^k \to \mathbb{R}^+$ is bounded, from Theorem 2.3 it follows that

$$L_k(Y^n, \tilde{V}^n) \to El_k(\{Y^*\}_1^k, \{V^*\}_1^k) =: R_k \quad \text{a.s. and in } L_1. \tag{30}$$

Moreover, for $k = 1$, from (16) we have

$$L_1(Y^n, \tilde{V}^n) \to El(Y_0, V_0) = \frac{1}{ET_1} E\left(\sum_{t=S_0+1}^{S_1} l(Y_t, V_t) \right) =: R_1 \quad \text{a.s. and in } L_1. \tag{31}$$

We shall call the constant R_k the *asymptotic Viterbi risk*. It depends only on the model (X, Y) and the loss function l_k. For $k = 1$ and $l(s, s') = I_{\{s' \neq s\}}$, the actual loss is the average number of mistakes made by the Viterbi alignment:

$$L_1(Y^n, v^n(X^n)) = \frac{1}{n} \sum_{t=1}^{n} I_{\{Y_t \neq \tilde{V}_t^n\}}, \tag{32}$$

and the corresponding asymptotic risk is the asymptotic misclassification probability $\mathbf{P}(Y_0 \neq V_0)$.

To our knowledge, the idea of considering the R_1-type limits for the Viterbi alignment has been first mentioned in [3], the convergence of the actual loss is also stated in [12]. To show the convergence of $L_k(Y^n, v^n(X^n))$, we use the following lemma (see Theorem 9.4.8 in [7]).

Lemma 3.1. *Let X_n be bounded random variables such that $X_n \to 0$ almost surely. Let $\{\mathcal{F}_n\}_{n=1}^{\infty}$ be a filtration. Then $E[X_n|\mathcal{F}_n] \to 0$ almost surely.*

The following theorem is the first main result of this paper.

Theorem 3.1. *Let $\{(X_t, Y_t)\}_{t=1}^{\infty}$ be an ergodic HMM satisfying* A1 *and* A2. *Then there exists a constant $R_k \geq 0$ such that the actual loss and the risk of the Viterbi alignment both converge to R_k almost surely and in L_1:*

$$\lim_{n\to\infty} L_k(Y^n, v^n(X^n)) = \lim_{n\to\infty} R_k(v, X^n) = R_k \quad \text{a.s. and in } L_1.$$

Moreover, the expected risk of the Viterbi alignment converges to R_k as well: $E R_k(v, X^n) \to R_k$.

Proof. The convergence of the actual loss is (30). To show that $R_k(v, X^n) \to R_k$ a.s., apply Lemma 3.1 with $X_n := L_k(Y^n, v^n(X^n)) - R_k$. Clearly, $L_k(Y^n, v^n(X^n)) - R_k$ is bounded and by (30) it goes to 0 a.s. Thus, by Lemma 3.1,

$$\begin{aligned} |E[L_k(Y^n, v^n(X^n)) - R_k|X^n]| &= |E[L_k(Y^n, v^n(X^n))|X^n] - R_k| \\ &= |R_k(v, X^n) - R_k| \to 0 \quad \text{a.s.} \end{aligned}$$

By Scheffe's theorem, the convergence in L_1 follows by the non-negativity and boundedness of $R_k(v, X^n)$. The convergence in L_1 implies the convergence of the expected risk. □

4. Convergence of $\bar{R}_1$-risk

For the convergence of the $\bar{R}_1$-risk we use Theorem 2.4. Recall that a double-sided infinite alignment v is a stationary coding. Consider the function $f : \mathcal{X}^{\mathbb{Z}} \to (-\infty, 0]$, where

$$f(x_{-\infty}^{\infty}) := \ln p_0\big(v(x_{-\infty}^{\infty})_0 | x_{-\infty}^{\infty}\big) = \ln \mathbf{P}(Y_0 = V_0 | X_{-\infty}^{\infty} = x_{-\infty}^{\infty}).$$

In the following, let $v_i(x_{-\infty}^{\infty}) := v(x_{-\infty}^{\infty})_i$ be the i-th element of the infinite alignment. Note that for every $t = 1, 2, \ldots,$

$$\begin{aligned} f\big(\theta_t(x_{-\infty}^{\infty})\big) &= \ln p_0\big(v_0(\theta_t(x_{-\infty}^{\infty}))|\theta_t(x_{-\infty}^{\infty})\big) = \ln p_t\big(v_0(\theta_t(x_{-\infty}^{\infty}))|x_{-\infty}^{\infty}\big) \\ &= \ln p_t\big(v_t(x_{-\infty}^{\infty})|x_{-\infty}^{\infty}\big) = \ln \mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty} = x_{-\infty}^{\infty}). \end{aligned}$$

Thus, by Birkhoff's ergodic theorem there exists a constant $\bar{R}_1$ such that

$$-\frac{1}{n}\sum_{t=1}^{n}\ln \mathbf{P}(Y_t = V_t|X_{-\infty}^{\infty}) \to -E\big(\ln \mathbf{P}(Y_0 = V_0|X_{-\infty}^{\infty})\big) =: \bar{R}_1 \quad \text{a.s. and in } L_1, \tag{33}$$

provided the expectation is finite.

The main idea for proving the convergence of $\bar{R}_1(v, X^n)$ is the following. Consider without loss of generality a double-sided HMM $\{(X_t, Y_t)\}_{t=-\infty}^{\infty}$. Then by RD2, $\tilde{V}_t^n = V_t$ for every $S_0 \le t \le S_{k(n)}$, where $k(n) = \max\{k \ge 0 : S_k + m \le n\}$ and $\{S_k\}_{k\ge 0}$ is the renewal process as in Theorem 2.4. Thus,

$$\begin{aligned}-\frac{1}{n}\sum_{t=1}^{n}\ln \mathbf{P}(Y_t = \tilde{V}_t^n|X^n) &= -\frac{1}{n}\sum_{t=1}^{S_0-1}\ln \mathbf{P}(Y_t = \tilde{V}_t^n|X^n) - \frac{1}{n}\sum_{t=S_0}^{S_{k(n)}}\ln \mathbf{P}(Y_t = V_t|X^n)\\ &\quad -\frac{1}{n}\sum_{t=S_{k(n)}+1}^{n}\ln \mathbf{P}(Y_t = \tilde{V}_t^n|X^n).\end{aligned} \tag{34}$$

The first term in the partition above converges to zero almost surely. We will prove that the second term converges to $\bar{R}_1$ almost surely and that the third term converges to zero almost surely. To prove the convergence of the second term, we need some auxiliary results.

Let C be the cluster as in A1 and let $\mathcal{X}_o$ be the corresponding set.

Proposition 4.1. *Let $x_{-\infty}^{\infty} \in \mathcal{X}^{\mathbb{Z}}$ be such that $v(x_{-\infty}^{\infty})$ is defined. Suppose that for some $u, w \in \mathbb{N}$, $x_{-u}^{-u+r} \in \mathcal{X}_o^{r+1}$, $x_{w-r}^{w} \in \mathcal{X}_o^{r+1}$ and for every $s \in S$, $\lim_n p_0(s|x_{-n}^{n}) = p_0(s|x_{-\infty}^{\infty})$. Let $v_0 = v_0(x_{-\infty}^{\infty})$. Then there exist constants $c > 0$ and $0 < B < \infty$ that are independent of data such that*

$$p_0\big(v_0|x_{-\infty}^{\infty}\big) \ge c\exp[-B(u+w)]. \tag{35}$$

The proof of Proposition 4.1 is given in Appendix. It reveals that the proposition holds also for a finite sequence of observations x^n. Moreover, the following corollary holds.

Corollary 4.1. *Let $x^n \in \mathcal{X}^n$ be such that for some $u < n - r$, $x_u^{u+r} \in \mathcal{X}_o^{r+1}$. Let $\tilde{v}_t = v_t^n(x^n)$. Then there exist $c > 0$ and $0 < D < \infty$ that are independent of data such that for every $t, u < t \le n$,*

$$p_t(\tilde{v}_t|x^n) \ge c\exp[-D(n-u)]. \tag{36}$$

The proof of Corollary 4.1 follows the one of Proposition 4.1 and is sketched in Appendix.

Lemma 4.1. *There exists $\alpha > 0$ such that for every $t \in \mathbb{Z}$,*

$$E\left(\frac{1}{\mathbf{P}(Y_t = V_t|X_{-\infty}^{\infty})}\right)^{\alpha} < \infty. \tag{37}$$

Proof. Let W_0 and U_0 be the stopping times defined in (18). Because for every $s \in S$, $\lim_n \mathbf{P}(Y_0 = s|X_{-n}^{n}) = \mathbf{P}(Y_0 = s|X_{-\infty}^{\infty})$ almost surely, from (35) it follows that

$$\mathbf{P}(Y_0 = V_0|X_{-\infty}^{\infty}) \ge c\exp[-B(W_0 - U_0)] \quad \text{a.s.} \tag{38}$$

From RD4 and RD5 it follows that for some positive constants a and b and for every $k = 1, 2, \ldots,$

$$\mathbf{P}(W_0 > k) \le a \exp[-bk].$$

This inequality implies that for $\alpha > 0$ small enough, $E(e^{\alpha W_0}) < \infty$. Analogously, for sufficiently small $\alpha > 0$, $E\left(e^{\alpha(-U_0)}\right) < \infty$. Thus, by the Cauchy–Schwarz inequality it holds that for sufficiently small α,

$$E\left(e^{\alpha(W_0-U_0)}\right) = E\left(e^{\alpha W_0} e^{\alpha(-U_0)}\right) \le \left(E\left(e^{2\alpha W_0}\right) E\left(e^{2\alpha(-U_0)}\right)\right)^{\frac{1}{2}} < \infty. \tag{39}$$

Inequalities (38) and (39) imply (37) for $t = 0$. By the stationarity of (X, Y), (37) holds for arbitrary t. □

Recall inequalities (20)–(21). Unfortunately these bounds do not immediately hold for the logarithms. The following lemma uses the inequality $|\ln a - \ln b| \le \frac{1}{\min\{a,b\}}|a - b|$, provided that $a, b > 0$. Recall M and $u(n)$ from (22).

Lemma 4.2. *Suppose that for an $\alpha > 0$,*

$$E\left(\frac{1}{\mathbf{P}(Y_0 = V_0 | X_{-\infty}^{\infty})}\right)^{\alpha} < \infty. \tag{40}$$

Then

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{S_{k(n)}} \ln \mathbf{P}(Y_t = V_t | X^n) = \bar{R}_1 \quad a.s. \tag{41}$$

Proof. Without loss of generality, we assume $\alpha \in (0, 1]$. Let $\xi_t := \mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty})$, $\eta_t^n := \mathbf{P}(Y_t = V_t | X^n)$, $\eta_t := \mathbf{P}(Y_t = V_t | X_1^{\infty})$ and let $\beta = \frac{1}{\alpha}$. Take $m(n) = u(n) - (\ln n)^2$. Split the sum in (41) as

$$-\frac{1}{n}\sum_{t=1}^{S_{k(n)}} \ln \eta_t^n = -\frac{1}{n}\sum_{t=1}^{m(n)} \ln \eta_t^n - \frac{1}{n}\sum_{t=m(n)+1}^{S_{k(n)}} \ln \eta_t^n = \mathrm{Term_I} + \mathrm{Term_{II}}.$$

For $m(n) > S_{k(n)}$, $\mathrm{Term_{II}} = 0$. We will prove that $\mathrm{Term_I}$ converges to $\bar{R}_1$ and $\mathrm{Term_{II}}$ converges to zero almost surely.

$\mathrm{Term_I}$. Recall that $\{\xi_t\}$ is a stationary ergodic process. Hence, by assumption,

$$\sum_{t=1}^{\infty} \mathbf{P}\left(\xi_t \le \frac{1}{t^{\beta}}\right) = \sum_{t=1}^{\infty} \mathbf{P}(\xi_t^{-\alpha} \ge t) \le E(\xi_t^{-\alpha}) + 1 < \infty.$$

Thus, the sequence $\xi_t, t = 1, 2, \ldots,$ satisfies $\mathbf{P}(\xi_t > \frac{1}{t^{\beta}} \text{ ev}) = 1$. From (20) it follows that $\mathbf{P}(\eta_t > \frac{1}{2t^{\beta}} \text{ ev}) = 1$. Thus, almost surely $|\ln \eta_t - \ln \xi_t| \le C_1 2t^{\beta} \rho^{t-1}$ eventually. The assumption in (40) ensures that $E|\ln \xi_0| < \infty$. Since $-\frac{1}{n}\sum_{t=1}^{n} \ln \xi_t \to \bar{R}_1$ almost surely, we now have

$$-\frac{1}{n}\sum_{t=1}^{n} \ln \eta_t \to \bar{R}_1 \quad \text{a.s.} \tag{42}$$

Let (random) T be so big that $\xi_t, \eta_t > \frac{1}{2t^\beta}$ when $t \geq T$. By (21), for every $t \leq u(n)$, $|\eta_t^n - \eta_t| \leq M\rho^{u(n)-t}$ holds almost surely. Observe that for n large enough,

$$M\rho^{(\ln n)^2} \leq \frac{1}{4t^\beta}.$$

Hence, for n large enough and t such that $T < t \leq u(n) - (\ln n)^2$, we have $|\eta_t^n - \eta_t| \leq M\rho^{u(n)-t} \leq \frac{1}{4t^\beta}$, implying that $\min\{\eta_t, \eta_t^n\} \geq \frac{1}{4t^\beta}$, and $|\ln \eta_t^n - \ln \eta_t| \leq 4t^\beta M\rho^{u(n)-t}$. Thus, as $n \to \infty$,

$$\begin{aligned}\left|\frac{1}{n}\sum_{t=1}^{m(n)} \ln \eta_t^n - \frac{1}{n}\sum_{t=1}^{m(n)} \ln \eta_t\right| &\leq \frac{1}{n}\sum_{t=1}^{T} |\ln \eta_t^n - \ln \eta_t| + \frac{1}{n}\sum_{t=T+1}^{m(n)} 4t^\beta M\rho^{u(n)-t} \\ &\leq \frac{1}{n}\sum_{t=1}^{T} |\ln \eta_t^n - \ln \eta_t| + \frac{4Mn}{n} n^\beta \rho^{(\ln n)^2} \to 0 \quad \text{a.s.}\end{aligned}$$

Recall that $u(n)/n \to 1$ almost surely. Hence, $m(n)/n \to 1$ almost surely, and it follows from (42) that $-\frac{1}{n}\sum_{t=1}^{m(n)} \ln \eta_t^n \to \bar{R}_1$ almost surely.

Term$_{\mathrm{II}}$. It remains to prove that

$$-\frac{1}{n}\sum_{t=m(n)+1}^{S_{k(n)}} \ln \eta_t^n \to 0 \quad \text{a.s.} \tag{43}$$

By Proposition 4.1, $\mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty}) \geq c \exp[-B(W_t - U_t)]$, where U_t and W_t are the stopping times defined in (18). Thus,

$$\eta_t^n = \mathbf{P}(Y_t = V_t | X^n) = E\left[\mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty}) | X^n\right] \geq cE\left[\exp[-B(W_t - U_t)] | X^n\right]. \tag{44}$$

Fix t and n, $t < n$. Define the events

$$E_{t,n}(k) := \{S_k < t \leq S_{k+1} \leq S_{k(n)}\}, \quad k = 1, 2, \ldots.$$

For any k, $E_{t,n}(k) \in \sigma(X^n)$. When $S_1 \leq t \leq S_{k(n)}$, i.e. the event $E_{t,n}(k)$ holds for some k, then according to RD3, $U_t > 0$ and $W_t \leq S_{k(n)} + m \leq n$. Therefore, for any constant $B > 0$ and for any k, the random variable $\exp[-B(W_t - U_t)] I_{E_{t,n}(k)}$ is $\sigma(X^n)$-measurable. Together with (44) this implies that

$$\begin{aligned}\eta_t^n I_{E_{t,n}(k)} &\geq cE\left[\exp[-B(W_t - U_t)] | X^n\right] I_{E_{t,n}(k)} \\ &= cE\left[\exp[-B(W_t - U_t)] I_{E_{t,n}(k)} | X^n\right] = c\exp[-B(W_t - U_t)] I_{E_{t,n}(k)}.\end{aligned}$$

If $S_k \leq t \leq S_{k+1}$, then by RD3 and RD4, $W_t - U_t \leq R_k + 2m$. Thus,

$$\eta_t^n I_{E_{t,n}(k)} \geq c\exp[-B(R_k + 2m)] I_{E_{t,n}(k)},$$

and we get the following bound:

$$-\ln \eta_t^n I_{E_{t,n}(k)} \leq \big(-\ln c + B(R_k + 2m)\big) I_{E_{t,n}(k)} \leq -\ln c + B(R_k + 2m). \tag{45}$$

Let $r(n) := \max\{k : S_k \leq m(n)\}$. Note that

$$\sum_{t=S_{r(n)}+1}^{S_{k(n)}} -\ln \eta_t^n = \sum_{k=r(n)}^{k(n)-1} \sum_{t>S_k}^{S_{k+1}} -\ln \eta_t^n = \sum_{k=r(n)}^{k(n)-1} \sum_{t>S_k}^{S_{k+1}} \big(-\ln \eta_t^n\big) I_{E_{t,n}(k)}. \tag{46}$$

Hence, from (45) and (46) it follows that

$$\sum_{t=m(n)+1}^{S_{k(n)}} -\ln \eta_t^n \le \sum_{t=S_{r(n)}+1}^{S_{k(n)}} -\ln \eta_t^n \le \sum_{j=r(n)+1}^{k(n)} R_j(-\ln c + B(R_j+2m))$$
$$\le \sum_{j=r(n)+1}^{k(n)} \left(BR_j^2 + AR_j\right),$$

where A and B are finite positive constants. We know that $u(n)/n \to 1$ almost surely. From this it follows that $r(n) \to \infty$ almost surely, and since R_j are i.i.d. with finite expectation, we get $R_{r(n)}/n \to 0$ almost surely. Thus,

$$\frac{S_{r(n)}}{n} = \frac{S_{r(n)+1} - R_{r(n)}}{n} \ge \frac{m(n) - R_{r(n)}}{n} = \frac{u(n) - (\ln n)^2 - R_{r(n)}}{n} \to 1 \quad \text{a.s.}$$

Clearly, also $S_{k(n)}/n \to 1$ almost surely. Since $r(n) \to \infty$ almost surely and R_j are i.i.d. random variables with all finite moments (RD4), we obtain for $s = 1, 2$ that

$$\frac{1}{r(n)} \sum_{j=1}^{r(n)} R_j^s \to ER_1^s < \infty \quad \text{a.s.}$$

Thus,

$$\frac{r(n)}{n} = \frac{r(n)}{S_{r(n)}} \frac{S_{r(n)}}{n} = \frac{r(n)}{\sum_{j=1}^{r(n)-1} R_j + S_1} \frac{S_{r(n)}}{n} \to \frac{1}{ER_1} \quad \text{a.s.},$$

and

$$\frac{1}{n} \sum_{j=1}^{r(n)} R_j^s = \frac{\sum_{j=1}^{r(n)} R_j^s}{r(n)} \frac{r(n)}{n} \to \frac{ER_1^s}{ER_1} \quad \text{a.s.}$$

By similar argument,

$$\frac{1}{n} \sum_{j=1}^{k(n)} R_j^s = \frac{\sum_{j=1}^{k(n)} R_j^s}{k(n)} \frac{k(n)}{n} \to \frac{ER_1^s}{ER_1} \quad \text{a.s.}$$

These two convergences imply that

$$\frac{1}{n} \sum_{j=r(n)}^{k(n)} R_j^s = \frac{1}{n} \sum_{j=1}^{k(n)} R_j^s - \frac{1}{n} \sum_{j=1}^{r(n)-1} R_j^s \to 0 \quad \text{a.s.},$$

which proves that

$$-\frac{1}{n} \sum_{t=m(n)+1}^{S_{k(n)}} \ln \eta_t^n \to 0 \quad \text{a.s.} \quad \square$$

We are now ready to prove the convergence of $\bar{R}_1(v, X^n)$.

Theorem 4.1. *Let* $\{(X_t, Y_t)\}_{t=1}^{\infty}$ *be an ergodic HMM satisfying* A1 *and* A2. *Then there exists a constant* $\bar{R}_1$ *such that*

$$\lim_{n\to\infty} \bar{R}_1(v, X^n) = \lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) = \bar{R}_1 \quad \text{a.s. and in } L_1.$$

Proof. Consider the partition in (34). By Lemma 4.2, the second term in (34) converges to $\bar{R}_1$ almost surely. Thus, it suffices to prove that

$$\frac{1}{n}\sum_{t=S_{k(n)}+1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) \to 0 \quad \text{a.s.} \tag{47}$$

The construction of S_k implies that for every $t \geq S_k$, $U_t > S_k - m$ (see R3). Hence, for every t such that $S_{k(n)} < t \leq n$, by (36),

$$|\ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n)| \leq D(n - U_t) + |\ln c| < D(S_{k(n)+1} - S_{k(n)} + 2m) + |\ln c|.$$

Therefore,

$$\begin{aligned}\left|\sum_{t=S_{k(n)}+1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n)\right| &\leq D(S_{k(n)+1} - S_{k(n)} + 2m)^2 \\ &\quad + (S_{k(n)+1} - S_{k(n)} + m)|\ln c| \\ &= D(R_{k(n)} + 2m)^2 + (R_{k(n)} + m)|\ln c|.\end{aligned}$$

For every $k > 0$, let

$$M_k = D(R_k + 2m)^2 + (R_k + m)|\ln c|.$$

The random variables M_k are i.i.d., and because renewal times have all finite moments, $EM_k < \infty$. Since the random variables M_k, $k \geq 1$, are identically distributed, it holds for every $\epsilon > 0$ that

$$\sum_{k=1}^{\infty} \mathbf{P}\left(\frac{M_k}{k} > \epsilon\right) = \sum_{k=1}^{\infty} \mathbf{P}\left(\frac{M_1}{\epsilon} > k\right) \leq \frac{EM_1}{\epsilon} < \infty.$$

Thus, by the Borel–Cantelli lemma $\frac{M_k}{k} \to 0$ almost surely as $k \to \infty$. Since $k(n) \to \infty$ as $n \to \infty$, we obtain

$$\frac{1}{n}\left|\sum_{t=S_{k(n)}+1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n)\right| \leq \frac{M_{k(n)}}{n} \leq \frac{M_{k(n)}}{k(n)} \to 0 \quad \text{a.s.} \quad \square$$

Remark. Note that the approach of the present section can be easily applied to prove that $R_1(v, X^n) \to R_1$ almost surely. Indeed, the counterpart of (33) is

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty}) \to E\big(\mathbf{P}(Y_0 = V_0 | X_{-\infty}^{\infty})\big) =: 1 - R_1 \quad \text{a.s. and in } L_1.$$

Inequalities (22) and convergence (25) imply

$$\begin{aligned}&\left|\frac{1}{n}\sum_{t=1}^{n}\big(\mathbf{P}(Y_t = V_t|X^n) - \mathbf{P}(Y_t = V_t|X_{-\infty}^{\infty})\big)\right| \\ &\quad\le \frac{1}{n}\sum_{t=1}^{u(n)}(C_1\rho^{t-1} + M\rho^{u(n)-t}) + \frac{2(n-u(n))}{n} \\ &\quad\le \frac{C_1+M}{n}\sum_{t=0}^{\infty}\rho^t + \frac{2(n-u(n))}{n} \to 0 \quad \text{a.s.},\end{aligned}$$

so that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathbf{P}(Y_t = V_t|X^n) = 1 - R_1 \quad \text{a.s.}$$

Since the probabilities are bounded, the convergence

$$R_1(v, X^n) = 1 - \frac{1}{n}\sum_{t=1}^{n}\mathbf{P}(Y_t = \tilde{V}_t^n|X^n) \to R_1 \quad \text{a.s.}$$

now easily follows.

From the remark above it is clear that the difficulties with the $\bar{R}_1$-risk are due to unboundedness of $\ln \mathbf{P}(Y_t = \tilde{V}_t^n|X^n)$, since in principle $\mathbf{P}(Y_t = \tilde{V}_t^n|X^n)$ can be arbitrarily small. However, the latter is not so when instead of the Viterbi alignment the PMAP-alignment is used. Indeed, by Birkhoff's theorem,

$$-\frac{1}{n}\sum_{t=1}^{n}\max_{s\in S}\ln \mathbf{P}(Y_t = s|X_{-\infty}^{\infty}) \to \bar{R}_1^* \quad \text{a.s. and in } L_1, \tag{48}$$

where $\bar{R}_1^*$ is a constant. The approximation of logarithms in the PMAP-alignment is considerably easier, since $\max_s \mathbf{P}(Y_t = s|X^n) \ge |S|^{-1}$, and therefore,

$$\begin{aligned}&|\max_s \ln \mathbf{P}(Y_t = s|X^n) - \max_s \ln \mathbf{P}(Y_t = s|X_{-\infty}^{\infty})| \\ &\quad\le |S|\|\mathbf{P}(Y_t \in \cdot|X^n) - \mathbf{P}(Y_t \in \cdot|X_{-\infty}^{\infty})\|.\end{aligned}$$

Therefore, inequality (22) implies

$$\begin{aligned}&|\max_s \ln \mathbf{P}(Y_t = s|X^n) - \max_s \ln \mathbf{P}(Y_t = s|X_{-\infty}^{\infty})| \\ &\quad\le \begin{cases}|S|(C_1\rho^{t-1} + M\rho^{u(n)-t}), & 1 \le t \le u(n); \\ 2|S|, & u(n) < t \le n.\end{cases}\end{aligned}$$

Thus, convergence (48) together with (25) gives

$$\bar{R}_1(X^n) = -\frac{1}{n}\sum_{t=1}^{n}\max_{s\in S}\ln \mathbf{P}(Y_t = s|X^n) \to \bar{R}_1^* \quad \text{a.s. and in } L_1.$$

Hence, the following corollary holds.

Corollary 4.2. *There exists a constant $\bar{R}_1^*$ such that $\bar{R}_1(X^n) \to \bar{R}_1^*$ a.s. and in L_1.*

5. Convergence of $\bar{R}_\infty$-risk

Recall that $\bar{R}_\infty(X^n) = -\frac{1}{n}\ln \mathbf{P}(Y^n = \tilde{V}^n|X^n)$ and $\tilde{V}^n = v^n(X^n)$. Let $p(x^n)$ be the likelihood of x^n and $p(x^n|s^n)$ denote the conditional likelihood of observing x^n given that $\{Y^n = s^n\}$. Note that $\ln p(x^n|s^n)$ can be expressed as

$$\ln p(x^n|s^n) = \sum_{t=1}^n \ln f_{s_t}(x_t) = \sum_{t=1}^n \ln f_1(x_t)I_1(s_t) + \cdots + \sum_{t=1}^n \ln f_{|S|}(x_t)I_{|S|}(s_t). \tag{49}$$

Let $\mathbf{P}(Y^n = v^n(X^n)) = \mathbf{P}(Y^n = s^n)_{|s^n:=v^n(X^n)}$. To prove the convergence of $\bar{R}_\infty(X^n)$, write $\mathbf{P}(Y^n = v^n(X^n)|X^n)$ as

$$\mathbf{P}(Y^n = v^n(X^n)|X^n) = \frac{p(X^n|v^n(X^n))\mathbf{P}(Y^n = v^n(X^n))}{p(X^n)}.$$

Then

$$\bar{R}_\infty(X^n) = -\frac{1}{n}\Big(\ln p(X^n|v^n(X^n)) + \ln \mathbf{P}(Y^n = v^n(X^n)) - \ln p(X^n)\Big). \tag{50}$$

Before stating the theorem about the convergence of $\bar{R}_\infty(X^n)$, we introduce the conditional measure $Q_s := \mathbf{P}(X_0 \in \cdot|V_0 = s), s \in S$. As it follows from Theorem 2.3, the measure Q_s is the almost sure limit of the empirical measure corresponding to the Viterbi alignment state s, i.e. for every Borel set A,

$$\frac{\sum_{t=1}^n I_{A\times s}(X_t, \tilde{V}_t^n)}{\sum_{t=1}^n I_s(\tilde{V}_t^n)} \to Q_s(A) \quad \text{a.s.}$$

This convergence is the basis of the adjusted Viterbi training introduced in [22,23]. Note that for every Q_s-integrable g,

$$E\big(g(X_0)I_s(V_0)\big) = E\big(g(X_0)|V_0 = s\big)\mathbf{P}(V_0 = s) = m_s \int g(x)Q_s(dx), \tag{51}$$

where $m_s := \mathbf{P}(V_0 = s)$. Recall that $Z^* = \{(X_t, Y_t, V_t)\}_{t=1}^\infty$ is a restriction of $Z = \{(X_t, Y_t, V_t)\}_{t=-\infty}^\infty$.

Theorem 5.1. *Let for every $s \in S$ the logarithm of the conditional density f_s be P_s-integrable. Then*

$$\bar{R}_\infty(X^n) \to -\sum_{s\in S} m_s \int \ln f_s(x)Q_s(dx) - E[\ln p_{V_1^*,V_2^*}] - H_X =: \bar{R}_\infty$$

a.s. and in L_1,

where H_X is the entropy rate of X and $p_{i,j} = \mathbf{P}(Y_2 = j|Y_1 = i)$.

Proof. Consider (50). To prove the convergence of the first term of the right-hand side, apply (49) to the Viterbi alignment. In [16] it was shown that if $\ln f_s$ is P_s-integrable, then $\ln f_s$ is also Q_s-integrable for every s. Then by Theorem 2.3 and (51), for every state $s \in S$,

$$\frac{1}{n}\sum_{t=1}^{n}\ln f_s(X_t)I_s(\tilde{V}_t^n) \to E\big(\ln f_s(X_0)I_s(V_0)\big) = m_s\int \ln f_s(x)Q_s(dx)$$
a.s. and in L_1.

This together with (49) gives

$$\frac{1}{n}\ln p(X^n|Y^n = v^n(X^n)) \to \sum_{s\in S} m_s\int \ln f_s(x)Q_s(dx) \quad \text{a.s. and in } L_1.$$

For the second term use the Markov property

$$\ln \mathbf{P}(Y^n = v^n(X^n)) = \ln \mathbf{P}(Y^n = \tilde{V}^n) = \ln \pi_{\tilde{V}_1^n} + \ln p_{\tilde{V}_1^n,\tilde{V}_2^n} + \cdots + \ln p_{\tilde{V}_{n-1}^n,\tilde{V}_n^n},$$

where $\pi_s = \mathbf{P}(Y_1 = s)$. Since $\tilde{V}^n$ is a path with positive likelihood, $p_{\tilde{V}_t^n,\tilde{V}_{t+1}^n} > 0$ almost surely for every t. Because the number of states is finite, there exists a constant $M > 0$ such that for every i, $-\ln p_{\tilde{V}_i^n,\tilde{V}_{i+1}^n} < M$ almost surely. Hence the assumptions of Theorem 2.3 hold and, with $p_{\tilde{V}_0^n,\tilde{V}_1^n} = \pi_{\tilde{V}_1^n}$, we get

$$\frac{1}{n}\ln \mathbf{P}(Y^n = \tilde{V}^n) = \frac{1}{n}\sum_{t=0}^{n-1}\ln p_{\tilde{V}_t^n,\tilde{V}_{t+1}^n} \to E[\ln p_{V_1^*,V_2^*}] \quad \text{a.s. and in } L_1,$$

where $E[\ln p_{V_1^*,V_2^*}] = \sum_{i,j\in S}\ln p_{i,j}\mathbf{P}(V_1^* = i, V_2^* = j)$. Finally, the Shannon–McMillan–Breiman theorem implies the convergence of the third term of the right-hand side in (50):

$$\frac{1}{n}\ln p(X^n) \to -H_X \quad \text{a.s. and in } L_1. \quad \square$$

Remark. Note that $-E[\ln p_{Y_1,Y_2}]$ is the entropy rate of Y. By the same argument,

$$\frac{1}{n}\ln \mathbf{P}(Y^n|X^n) \to \sum_{s\in S}\pi_s\int \ln f_s(x)P_s(dx) - H_Y + H_X =: -\bar{R}_\infty^Y \quad \text{a.s. and in } L_1,$$

where H_Y is the entropy rate of Y. The convergence in L_1 implies

$$-\frac{1}{n}E[\ln \mathbf{P}(Y^n|X^n)] \to \bar{R}_\infty^Y,$$

where the expectation is taken over X^n and Y^n. Since $-E[\ln \mathbf{P}(Y^n|X^n)] = H(Y^n|X^n)$ (the conditional entropy of Y^n given X^n), the limit $\bar{R}_\infty^Y$ could be interpreted as the conditional entropy rate of Y given X, it is not the entropy rate of Y. Clearly, $\bar{R}_\infty \le \bar{R}_\infty^Y$, and the difference of those two numbers shows how much the Viterbi alignment "overestimates" the likelihood. This means that the smaller the constant $\bar{R}_\infty$ is compared to $\bar{R}_\infty^Y$, the larger is the conditional likelihood of the Viterbi alignment compared to $\mathbf{P}(Y^n|X^n)$ for large n.

Acknowledgments

The authors are grateful to the anonymous referees for their valuable comments and remarks. The second author was supported by Estonian science foundation grant no 9288 and targeted financing project SF0180015s12.

Appendix. Proofs of Theorem 2.3, Proposition 4.1 and Corollary 4.1

A.1. Proof of Theorem 2.3

Proof. Partition the sum in (19) as

$$\frac{1}{n-p+1}\sum_{i=p}^{n}\tilde{U}_i^n = \frac{1}{n-p+1}\left(\sum_{i=p}^{S_{k(n)}}U_i + \sum_{i=S_{k(n)}+1}^{n}\tilde{U}_i^n\right).$$

Since $S_{k(n)} \to \infty$ almost surely, from (14) we know that

$$\frac{1}{S_{k(n)}}\sum_{i=p}^{S_{k(n)}}U_i \to Eg_p(Z_1^*,\dots,Z_p^*) \quad \text{a.s. and in } L_1. \tag{A.1}$$

Since $ET_1 < \infty$ and $n \geq p$, by SLLN and the elementary renewal theorem

$$\frac{S_{k(n)}}{n-p+1} = \frac{S_{k(n)}}{k(n)}\frac{k(n)}{n-p+1} \to 1 \quad \text{a.s. and in } L_1.$$

Combining this with (A.1) and taking into account that the sequence $\{\frac{S_{k(n)}}{n-p+1}\}$ is bounded, we obtain that

$$\frac{1}{n-p+1}\sum_{i=p}^{S_{k(n)}}U_i \to Eg_p(Z_1^*,\dots,Z_p^*) \quad \text{a.s. and in } L_1.$$

Note that

$$\left|\frac{1}{n-p+1}\sum_{i=S_{k(n)}+1}^{n}\tilde{U}_i^n\right| \leq \frac{M_{k(n)}}{S_{k(n)}+1-p} \leq \frac{M_{k(n)}}{k(n)-p+1}.$$

Since the random variables $M_k, k \geq p$, are identically distributed, it holds for every $\epsilon > 0$ that

$$\sum_{k=p}^{\infty}\mathbf{P}\left(\frac{M_k}{k} > \epsilon\right) = \sum_{k=p}^{\infty}\mathbf{P}\left(\frac{M_p}{\epsilon} > k\right) \leq \frac{EM_p}{\epsilon} < \infty.$$

Thus, by the Borel–Cantelli lemma $\frac{M_k}{k} \to 0$ almost surely as $k \to \infty$. Clearly, $E\left[\frac{M_k}{k}\right] \to 0$, so by Scheffe's theorem $\frac{M_k}{k} \to 0$ in L_1 as well. □

A.2. Preliminaries for proving Proposition 4.1 and Corollary 4.1

Let us start with some notation. Recall that $p_{i,j} = \mathbf{P}(Y_2 = j|Y_1 = i)$ and $\pi_i = \mathbf{P}(Y_1 = i)$. For every sequence of observations $x_k^l = (x_k,\dots,x_l) \in \mathcal{X}^{l-k+1}$, for every sequence of states $y_k^l = (y_k,\dots,y_l) \in S^{l-k+1}$ and states $i, j \in S$, we denote by $p(x_k^l, y_k^l, j|i)$ the following conditional likelihood:

$$p(x_k^l, y_k^l, j|i) := p_{i,y_k}\prod_{u=k}^{l-1}p_{y_u,y_{u+1}}p_{y_l,j}\prod_{u=k}^{l}f_{y_u}(x_u).$$

Similarly,

$$p(x_k^l, y_k^l|i) := \sum_j p(x_k^l, y_k^l, j|i), \qquad p(x_k^l, y_k^l) := \sum_i p(x_k^l, y_k^l|i)\pi_i.$$

We also define

$$\alpha(x_k^l, s) := \sum_{y_k^l \in S^{k-l+1}: y_l = s} p(x_k^l, y_k^l), \qquad \beta(x_k^l|i) = \sum_{y_k^l \in S^{k-l+1}} p(x_k^l, y_k^l|i).$$

The last two notations are standard in the HMM literature, see e.g. [10,5]. Let

$$\beta(x_k^l, s|i) = \sum_{y_k^l \in S^{k-l+1}: y_l = s} p(x_k^l, y_k^l|i), \qquad \alpha(s, x_k^l) := \sum_{y_k^l \in S^{k-l+1}: y_k = s} p(x_k^l, y_k^l).$$

Finally, let

$$\sigma(x_k^l, j|i) := \max_{y_k^l} p(x_k^l, y_k^l, j|i), \qquad \sigma(x_k^l|i) := \max_{y_k^l} p(x_k^l, y_k^l|i).$$

Let C be the cluster as in A1. Thus, there is an $r \geq 1$ such that the matrix R^r has positive entries. Let $\mathcal{X}_o$ be the corresponding set. Suppose $z^r \in \mathcal{X}_o^r$ and $y^r \in C^r$. By the definition of $\mathcal{X}_o$, it holds that

$$\epsilon^r \leq \left(\prod_{u=1}^r f_{y_u}(z_u)\right) \leq M^r.$$

By the cluster assumption, $0 < \min_{i,j \in C} R^r(i, j) \leq \left(p_{i,y_1} p_{y_1,y_2} \dots p_{y_{r-1},j}\right) \leq 1$, provided $i, j \in C$. Hence there exist constants $0 < a < A < \infty$, not depending on the observations, such that

$$a < p(x^r, y^r|i) < A \quad \text{and} \quad a < p(x^{r-1}, y^{r-1}, j|i) < A, \quad j \in C. \tag{A.2}$$

Suppose now x^m, $m > r$, is a sequence of observations such that the first r elements belong to the set $\mathcal{X}_o$, i.e. $x^r \in \mathcal{X}_o^r$. Then for every i, $p(x^m, y^m|i) > 0$ only if $y^r \in C^r$, implying that

$$\sigma(x^m, j|i) = \max_{s \in C} \max_{y^r \in C^r: y_r = s} p(x^r, y^r|i)\sigma(x_{r+1}^m, j|s).$$

Let now $i_1, i_2 \in C$. Then for some states $s_1, s_2 \in C$,

$$\begin{aligned}
\sigma(x^m, j|i_1) &= \max_{y^r \in C^r: y_r = s_1} p(x^r, y^r|i_1)\sigma(x_{r+1}^m, j|s_1), \\
\sigma(x^m, j|i_2) &= \max_{y^r \in C^r: y_r = s_2} p(x^r, y^r|i_2)\sigma(x_{r+1}^m, j|s_2) \\
&\geq \max_{y^r \in C^r: y_r = s_1} p(x^r, y^r|i_2)\sigma(x_{r+1}^m, j|s_1).
\end{aligned}$$

Hence, inequalities (A.2) imply that for every state j,

$$\frac{\sigma(x^m, j|i_1)}{\sigma(x^m, j|i_2)} \leq \frac{\max_{y^r \in C^r: y_r = s_1} p(x^r, y^r|i_1)}{\max_{y^r \in C^r: y_r = s_1} p(x^r, y^r|i_2)} \leq \frac{A}{a}. \tag{A.3}$$

Similarly, if x^m is such that the last r elements belong to $\mathcal{X}_o$, i.e. $x_{m-r+1}^m \in \mathcal{X}_o^r$, then for arbitrary states $j_1, j_2 \in C$ there exist $s_1, s_2 \in C$ such that

$$\begin{aligned}
\sigma(x^m, j_1|i) &= \max_{y^{m-r+1}: y_{m-r+1}=s_1} p(x^{m-r+1}, y^{m-r+1}|i)\sigma(x_{m-r+2}^m, j_1|s_1),\\
\sigma(x^m, j_2|i) &= \max_{y^{m-r+1}: y_{m-r+1}=s_2} p(x^{m-r+1}, y^{m-r+1}|i)\sigma(x_{m-r+2}^m, j_2|s_2)\\
&\geq \max_{y^{m-r+1}: y_{m-r+1}=s_1} p(x^{m-r+1}, y^{m-r+1}|i)\sigma(x_{m-r+2}^m, j_2|s_1).
\end{aligned}$$

So from (A.2) it follows that

$$\frac{\sigma(x^m, j_1|i)}{\sigma(x^m, j_2|i)} \leq \frac{\sigma(x_{m-r+2}^m, j_1|s_1)}{\sigma(x_{m-r+2}^m, j_2|s_1)} \leq \frac{A}{a}. \tag{A.4}$$

Proof of Proposition 4.1. Let $x_{-\infty}^{\infty}$ be a sequence of observations and let x_{-n}^n be its subword. For every state $i \in S$, we are interested in probability $p_0(i|x_{-n}^n) := \mathbf{P}(Y_0 = i|X_{-n}^n = x_{-n}^n)$. Note that

$$p_0(i|x_{-n}^n)p(x_{-n}^n) = \sum_{y_{-n}^n: y_0=i} p(x_{-n}^n, y_{-n}^n) =: \gamma_0(x_{-n}^n, i).$$

Observe that for every $u, w \in \{1, \ldots, n-1\}$ and for an arbitrary state, let it be 1,

$$\begin{aligned}
\gamma_0(x_{-n}^n, 1) &= \sum_{s_1\in S}\sum_{s_2\in S}\sum_{s_3\in S}\sum_{s_4\in S} \alpha(x_{-n}^{-u}, s_1)\beta(x_{-u+1}^{-1}, s_2|s_1)\\
&\quad \times p_{s_2,1} f_1(x_0)\beta(x_1^{w-1}, s_3|1)p_{s_3,s_4}\alpha(s_4, x_w^n)\\
&\geq \sum_{s_1\in S}\sum_{s_4\in S} \alpha(x_{-n}^{-u}, s_1)\sigma(x_{-u+1}^{-1}, 1|s_1) f_1(x_0)\sigma(x_1^{w-1}, s_4|1)\alpha(s_4, x_w^n)\\
&\geq p(x_{-n}^{-u})\big(\min_{s\in S}\sigma(x_{-u+1}^{-1}, 1|s)\big) f_1(x_0)\big(\min_{s\in S}\sigma(x_1^{w-1}, s|1)\big)p(x_w^n).
\end{aligned}$$

Without loss of generality assume $v_0(x_{-\infty}^{\infty}) = 1$. Let $v_{-u}(x_{-\infty}^{\infty}) = a$ and $v_w(x_{-\infty}^{\infty}) = b$. By Bellman's optimality principle, for every $i_o \in S$

$$\sigma(x_{-u+1}^{-1}, 1|a) f_1(x_0)\sigma(x_1^{w-1}, b|1) \geq \sigma(x_{-u+1}^{-1}, i_o|a) f_{i_o}(x_0)\sigma(x_1^{w-1}, b|i_o),$$

implying that for every state i_o,

$$f_1(x_0) \geq \frac{\sigma(x_{-u+1}^{-1}, i_o|a)}{\sigma(x_{-u+1}^{-1}, 1|a)} f_{i_o}(x_0)\frac{\sigma(x_1^{w-1}, b|i_o)}{\sigma(x_1^{w-1}, b|1)}.$$

Thus,

$$\begin{aligned}
\gamma_0(x_{-n}^n, 1) &\geq p(x_{-n}^{-u})\frac{\big(\min_{s\in S}\sigma(x_{-u+1}^{-1}, 1|s)\big)}{\sigma(x_{-u+1}^{-1}, 1|a)}\sigma(x_{-u+1}^{-1}, i_o|a) f_{i_o}(x_0)\sigma(x_1^{w-1}, b|i_o)\\
&\quad \times \frac{\big(\min_{s\in S}\sigma(x_1^{w-1}, s|1)\big)}{\sigma(x_1^{w-1}, b|1)}p(x_w^n).
\end{aligned} \tag{A.5}$$

Note that for every x_k^m,

$$\sum_s \beta(x_k^m, s|i)p_{s,j} = \sum_{y_k^m} p(x_k^m, y_k^m, j|i) \le |S|^{m-k+1}\sigma(x_k^m, j|i).$$

Therefore, for every $i_o \in S$,

$$\begin{aligned}
\gamma_0(x_{-n}^n, i_o) &= \sum_{s_1\in S}\sum_{s_2\in S}\sum_{s_3\in S}\sum_{s_4\in S} \alpha(x_{-n}^{-u}, s_1)\beta(x_{-u+1}^{-1}, s_2|s_1) \\
&\quad \times p_{s_2,i_o} f_{i_o}(x_0)\beta(x_1^{w-1}, s_3|i_o)p_{s_3,s_4}\alpha(s_4, x_w^n) \\
&\le \sum_{s_1\in S}\sum_{s_4\in S} \alpha(x_{-n}^{-u}, s_1)|S|^{u-1}\sigma(x_{-u+1}^{-1}, i_o|s_1) \\
&\quad \times f_{i_o}(x_0)|S|^{w-1}\sigma(x_1^{w-1}, s_4|i_o)\alpha(s_4, x_w^n) \\
&\le p(x_{-n}^{-u})|S|^{u-1}\big(\max_{s\in S}\sigma(x_{-u+1}^{-1}, i_o|s)\big) f_{i_o}(x_0)|S|^{w-1} \\
&\quad \times \big(\max_{s\in S}\sigma(x_1^{w-1}, s|i_o)\big)p(x_w^n).
\end{aligned}$$

Let x_{-n}^n be such that $x_{-u}^{-u+r} \in \mathcal{X}_o^{r+1}$ and $x_{w-r}^w \in \mathcal{X}_o^{r+1}$. Then $\alpha(x_{-n}^{-u}, s_1) = 0$ if $s_1 \notin C$, since $x_{-u} \in \mathcal{X}_o$. Analogously, $\alpha(s_4, x_w^n) = 0$ if $s_4 \notin C$. Thus, in this case the inequality above becomes

$$\begin{aligned}
\gamma_0(x_{-n}^n, i_o) &\le p(x_{-n}^{-u})|S|^{u-1}\big(\max_{s\in C}\sigma(x_{-u+1}^{-1}, i_o|s)\big) f_{i_o}(x_0)|S|^{w-1} \\
&\quad \times \big(\max_{s\in C}\sigma(x_1^{w-1}, s|i_o)\big)p(x_w^n).
\end{aligned}$$

The same holds for (A.5), implying that

$$\begin{aligned}
\frac{\gamma_0(x_{-n}^n, 1)}{\gamma_0(x_{-n}^n, i_o)} &\ge \frac{\min_{s\in C}\sigma(x_{-u+1}^{-1}, 1|s)}{\sigma(x_{-u+1}^{-1}, 1|a)} \frac{\sigma(x_{-u+1}^{-1}, i_o|a)}{\max_{s\in C}\sigma(x_{-u+1}^{-1}, i_o|s)} \\
&\quad \times \frac{\sigma(x_1^{w-1}, b|i_0)}{\max_{s\in C}\sigma(x_1^{w-1}, s|i_o)} \frac{\min_{s\in C}\sigma(x_1^{w-1}, s|1)}{\sigma(x_1^{w-1}, b|1)}|S|^{2-(u+w)}.
\end{aligned}$$

Inequalities (A.3) and (A.4) imply that the ratios above are bounded below by $\frac{a}{A}$ that does not depend on the observations. Thus, there exist constants c_1 and $0 < B < \infty$ not depending on the data such that for every state i_o,

$$\frac{p_0(1|x_{-n}^n)}{p_0(i_o|x_{-n}^n)} = \frac{\gamma_0(x_{-n}^n, 1)}{\gamma_0(x_{-n}^n, i_o)} \ge c_1 \exp[-B(u+w)]. \tag{A.6}$$

Since $\sum_{i\in S} p_0(i|x_{-n}^n) = 1$, there exists i_o such that $p_0(i_o|x_{-n}^n) \ge |S|^{-1}$. Thus, by (A.6),

$$p_0(1|x_{-n}^n) \ge \frac{c_1}{|S|}\exp[-B(u+w)].$$

Because $p_0(1|x_{-n}^n) \to p_0(1|x_{-\infty}^{\infty})$, inequality (35) follows by taking $c = \frac{c_1}{|S|}$. □

Proof of Corollary 4.1. The proof is analogous to the proof of Proposition 4.1. Using the same notations we obtain that for every t, $u < t < n$,

$$\gamma_t(x^n, \tilde{v}_t) \ge p(x^u)\big(\min_{s\in C}\sigma(x_{u+1}^{t-1}, \tilde{v}_t|s)\big) f_{\tilde{v}_t}(x_t)\sigma(x_{t+1}^n|\tilde{v}_t).$$

For every $i_o \in S$,

$$\gamma_t(x^n, i_o) \le p(x^u)\big(\max_{s\in C} \sigma(x_{u+1}^{t-1}, i_o|s)\big) f_{i_o}(x_t)\sigma(x_{t+1}^n|i_o)|S|^{n-u-1}.$$

Let $v_u(x^n) = b$. By Bellman's optimality principle,

$$f_{\tilde{v}_t}(x_t) \ge \frac{\sigma(x_{u+1}^{t-1}, i_o|b)}{\sigma(x_{u+1}^{t-1}, \tilde{v}_t|b)} f_{i_o}(x_t) \frac{\sigma(x_{t+1}^n|i_o)}{\sigma(x_{t+1}^n|\tilde{v}_t)}.$$

Thus,

$$\frac{p_t(\tilde{v}_t|x^n)}{p_t(i_o|x^n)} = \frac{\gamma_t(x^n, \tilde{v}_t)}{\gamma_t(x^n, i_o)} \ge \frac{\min_{s\in C} \sigma(x_{u+1}^{t-1}, \tilde{v}_t|s)}{\sigma(x_{u+1}^{t-1}, \tilde{v}_t|b)} \frac{\sigma(x_{u+1}^{t-1}, i_o|b)}{\max_{s\in C} \sigma(x_{u+1}^{t-1}, i_o|s)} |S|^{-(n-u-1)}.$$

Because the ratios above are bounded below by $\frac{a}{A}$ and $p_t(i_o|x^n) \ge |S|^{-1}$ for some $i_o \in S$, the statement of the corollary follows with $D = \ln|S|$. □

References

[1] S. Asmussen, Applied Probability and Queues, Springer, New York, 2003.
[2] G.D. Brushe, R.E. Mahony, J.B. Moore, A soft output hybrid algorithm for ML/MAP sequence estimation, IEEE Trans. Inform. Theory 44 (7) (1998) 3129–3134.
[3] A. Caliebe, Properties of the maximum a posteriori path estimator in hidden Markov models, IEEE Trans. Inform. Theory 52 (1) (2006) 41–51.
[4] A. Caliebe, U. Rösler, Convergence of the maximum a posteriori path estimator in hidden Markov models, IEEE Trans. Inform. Theory 48 (7) (2002) 1750–1758.
[5] O. Cappé, E. Moulines, T. Rydén, Inference in Hidden Markov Models, Springer, New York, 2005.
[6] P. Chigansky, Y. Ritov, On the Viterbi process with continuous state space, Bernoulli 17 (2) (2011) 609–627.
[7] K.L. Chung, A Course in Probability Theory, Academic Press, New York, 1974.
[8] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, 1998.
[9] R. Durrett, Probability: Theory and Examples, Brooks/Cole, Pacific Grove, 1991.
[10] Y. Ephraim, N. Merhav, Hidden Markov processes, IEEE Trans. Inform. Theory 48 (6) (2002) 1518–1569.
[11] P. Fariselli, P.L. Martelli, R. Casadio, A new decoding algorithm for hidden Markov models improves the prediction of the topology of all-beta membrane proteins, BMC Bioinformatics 6 (Suppl. 4) (2005) S12.
[12] A.P. Ghosh, E. Kleiman, A. Roitershtein, Large deviation bounds for functionals of Viterbi paths, IEEE Trans. Inform. Theory 57 (6) (2011) 3932–3937.
[13] J. Hayes, T. Cover, J. Riera, Optimal sequence detection and optimal symbol-by-symbol detection: similar algorithms, IEEE Trans. Commun. 30 (1) (1982) 152–157.
[14] F. Jelinek, Statistical Methods for Speech Recognition, The MIT Press, Cambridge, MA, USA, 1997.
[15] V.V. Kalashnikov, Topics on Regenerative Processes, CRC Press, Boca Raton, 1994.
[16] A. Koloydenko, M. Käärik, J. Lember, On adjusted Viterbi training, Acta Appl. Math. 96 (1–3) (2007) 309–326.
[17] A. Koloydenko, J. Lember, Infinite Viterbi alignments in the two state hidden Markov models, Acta Comment. Univ. Tartu. Math. 12 (2008) 109–124.
[18] A. Koloydenko, J. Lember, Hidden path inference, Tech. Rep., Mathematics Department, Royal Holloway, University of London, 2010. http://personal.rhul.ac.uk/utah/113/pfinds/index.html.
[19] T. Koski, Hidden Markov Models for Bioinformatics, Kluwer Academic Publishers, Dordrecht, 2001.
[20] J. Lember, A correction on approximation of smoothing probabilities for hidden Markov models, Statist. Probab. Lett. 81 (9) (2011) 1463–1464.
[21] J. Lember, On approximation of smoothing probabilities for hidden Markov models, Statist. Probab. Lett. 81 (2) (2011) 310–316.
[22] J. Lember, A. Koloydenko, Adjusted Viterbi training: a proof of concept, Probab. Eng. Inf. Sci. 21 (3) (2007) 451–475.

[23] J. Lember, A. Koloydenko, The adjusted Viterbi training for hidden Markov models, Bernoulli 14 (1) (2008) 180–206.
[24] J. Lember, A. Koloydenko, A constructive proof of the existence of Viterbi processes, IEEE Trans. Inform. Theory 56 (4) (2010) 2017–2033.
[25] J. Lember, A. Koloydenko, A generalized risk-based approach to segmentation based on hidden Markov models, 2010. arXiv:1007.3622.
[26] J. Lember, K. Kuljus, A. Koloydenko, Theory of segmentation, in: P. Dymarsky (Ed.), Hidden Markov Models, Theory and Applications, InTech, 2011, pp. 51–84.
[27] J. Li, R.M. Gray, R.A. Olshen, Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models, IEEE Trans. Inform. Theory 46 (5) (2000) 1826–1841.
[28] F.J. Och, H. Ney, Improved statistical alignment models, in: Proc. 38th Ann. Meet. Assoc. Comput. Linguist., 2000, pp. 440–447.
[29] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[30] P. Robertson, E. Villebrun, P. Hoeher, A comparison of optimal and sub-optimal MAP decoding algorithms operating in the log domain, in: Communications, 1995. ICC'95 Seattle, 'Gateway to Globalization', 1995, IEEE International Conference on, vol. 2, 1995, pp. 1009–1013. http://dx.doi.org/10.1109/ICC.1995.524253.
[31] H. Rue, New loss functions in Bayesian imaging, J. Amer. Statist. Assoc. 90 (431) (1995) 900–908.
[32] H. Thorisson, Coupling, Stationarity, and Regeneration, Springer, New York, 2000.
[33] G. Winkler, Image Analysis, Random Fields and Markov Chain Monte Carlo Methods, Springer, Berlin, 2003.