Asymptotic risks of Viterbi segmentation

Kristi Kuljus and Jüri Lember

Research Report
Centre of Biostochastics

Swedish University of
Agricultural Sciences

Report 2010:02
ISSN 1651-8543

Asymptotic risks of Viterbi segmentation

Kristi Kuljus

Centre of Biostochastics
Swedish University of Agricultural Sciences, 901 83 Umeå, Sweden

Jüri Lember[1]

Tartu University, J. Liivi 2 - 507, Tartu 50408, Estonia

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.

[1]E-mail address to the correspondence author: jyril@ut.ee

1 Introduction

1.1 Notation

Let $Y = \{Y_t\}_{t=-\infty}^{\infty}$ be a double-sided stationary MC with states $S = \{1, \dots, |S|\}$ and irreducible aperiodic transition matrix $\big(P(i,j)\big)$. 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 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 (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 the observations and Y for the latent variables.

Given a set $\mathcal{A}$ and integers m and n, $m < n$, we shall denote any $m - n + 1$-dimensional vector with all the components in $\mathcal{A}$ by $a_m^n := (a_m, \dots, a_n)$. 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 [22, 10], bioinformatics [15, 7], language processing [21], image analysis [20] and many others. For general overview about HMMs, we refer to [4] and [8].

1.2 Segmentation

The present paper deals with the asymptotics of the Viterbi segmentation. The *segmentation* problem consists of estimating the unobserved realization of the underlying Markov chain $Y_1, \dots, Y_n$ given n observations $x^n = x_1, \dots, x_n$ 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 (see [13] for details). For finding the best 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 the minimum risk. In the framework of pattern recognition theory the risk is specified via a *loss function*

$$L : S^n \times S^n \to [0, \infty],$$

where $L(a^n, b^n)$ measures the loss when the actual state sequence is a^n and the prognose is b^n. For any state sequence $s^n \in S^n$ the risk is then

$$R(s^n|x^n) := E[L(Y^n, s^n)|X^n = x^n] = \sum_{a^n \in S^n} L(a^n, s^n)\mathbf{P}(Y^n = a^n|X^n = x^n). \quad (1.1)$$

The most popular loss function is the so-called *symmetric loss* L_∞ defined as

$$L_\infty(a^n, b^n) = \begin{cases} 1, & \text{if } a^n \neq b^n; \\ 0, & \text{if } a^n = b^n. \end{cases} \quad (1.2)$$

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:

$$\bar{R}_\infty(s^n|x^n) := -\frac{1}{n} \ln p(s^n|x^n). \quad (1.3)$$

The log-likelihood based risk (1.3) is often preferable to use since it allows various generalizations, see (1.6).

Another popular classifier is based on the point-wise loss function

$$L_1(a^n, b^n) = \frac{1}{n} \sum_{t=1}^{n} l(a_t, b_t), \quad (1.4)$$

where $l(a_t, b_t) \geq 0$ is the loss of classifying the t-th symbol a_t as b_t. Typically, for every state s, $l(s, s) = 0$. Let us denote the corresponding risk by $R_1(s^n|x^n)$. It is not hard to see that

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

where $R_1^t(s|x^n) := \sum_{a \in S} l(a, s)p_t(a|x^n)$ and $p_t(a|x^n) := \mathbf{P}(Y_t = a|X^n = x^n)$. Most frequently $l(s, s') = I_{\{s \neq s'\}}$ (symmetric l), 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,

$$R_1(s^n|x^n) = 1 - \frac{1}{n} \sum_{t=1}^{n} p_t(s_t|x^n). \quad (1.5)$$

The minimizer of (1.5) over all the possible state sequences is called the *pointwise maximum a posteriori* (PMAP) alignment. The Viterbi and the PMAP-classifier – the so-called standard classifiers – are by far the two most popular classifiers used in practice.

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 (1.5) over all possible state sequences is clearly equivalent to minimizing $\bar{R}_1$, but this is not necessarily so for restricted minimization. The importance of $\bar{R}_1$ and $\bar{R}_\infty$ becomes apparent in [13], where the following penalized $\bar{R}_1$-risk is considered:

$$\bar{R}_C(s^n|x^n) := \bar{R}_1(s^n|x^n) + C\bar{R}_\infty(s^n|x^n). \tag{1.6}$$

Here $C \geq 0$ is a given regularization constant. The risk $\bar{R}_C$ naturally interpolates between the two standard alignments: for $C = 0$ the minimizer of (1.6) is the PMAP-alignment, and it is not hard to see that for C big enough the minimizer of (1.6) is the Viterbi alignment. Obviously, the likelihood of the minimizer of (1.6) increases with C as well as the $\bar{R}_1$-risk. Similar properties hold of course for the minimizer of the risk $R_1(s^n|x^n)+C\bar{R}_\infty(s^n|x^n)$, but the risk $\bar{R}_C$ has a nice easily understandable additional interpretation. Recall that the PMAP-alignment maximizes the expected number of correctly estimated sates. Unfortunately, it might have zero likelihood and as already mentioned by Rabiner in his seminal tutorial [22], a possible solution to that problem might be the alignment that maximizes the expected number of correctly estimated pairs or triplets of adjacent states rather than the expected number of correct states. In [13] it was shown that for integer C, minimizing the risk $\bar{R}_C$ is closely related to maximizing the expected number of correctly estimated tuples of $C + 1$ adjacent states.

In [13] it was also shown that the minimization of $\bar{R}_C(s^n|x^n)$ as well as of $R_1(s^n|x^n) + C\bar{R}_\infty(s^n|x^n)$ can be carried out by a dynamic programming algorithm that is similar to the Viterbi algorithm and easy to implement.

1.3 Asymptotic risks and the organization of the paper

Given a classifier g, the quantity $R(g, x^n) := R(g(x^n)|x^n)$ measures the goodness of it when 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 random variables $R(g, X^n)$. In this paper we shall show that under fairly general assumptions on an HMM, the random variables $R_1(v, X^n)$, $\bar{R}_1(v, X^n)$ as well as $\bar{R}_\infty(X^n) := \bar{R}_\infty(v, X^n)$ all converge to constant limits almost surely (Theorems 5, 6, 7, respectively). The convergence of $\bar{R}_\infty(v, X^n)$ and $\bar{R}_1(v, X^n)$ obviously implies the convergence of $\bar{R}_C(v, X^n)$. In [16] it was shown that under the same assumptions $R_1(X^n) = \min_{s^n} R_1(s^n|X^n)$ converges to a constant limit, here we prove this for $\bar{R}_1(X^n) = \min_{s^n} \bar{R}_1(s^n|X^n)$ (Corollary 3).

The limits – *asymptotic risks* – are constants that all depend on the model and characterize the goodness of the segmentation based on the Viterbi alignment. If, for example, R_1 is the limit of $R_1(v, X^n)$ and R_1^* is the limit of $R_1(X^n)$, then the difference $R_1 - R_1^*$ shows how well the Viterbi alignment performs the segmentation in the sense

of R_1-risk in the long run in comparison to the best possible alignment. If R_1 is defined as in (1.5), then for n big enough the Viterbi alignment makes approximatively nR_1 classification errors, while the best alignment in this case – the PMAP-alignment – makes approximatively nR^* errors. Since the model is known, the asymptotic risks could in principle be found theoretically, but the convergence theorems show that they could also be found by simulations.

Of course, when measuring the goodness of a segmentation with the R_1-risk, the quantity of actual interest is the so-called empirical (or true) risk

$$R_1(g, Y^n, X^n) := \frac{1}{n} \sum_{t=1}^{n} l(Y_t, g_t(X^n)),$$

where $g_t(X^n)$ is the t-th element of the n-dimensional vector $g(X^n)$. Since Y^n is hidden, the empirical risk $R_1(g, Y^n, X^n)$ cannot be found, but Theorem 5 implies that for the Viterbi alignment the empirical risk converges to R_1 almost surely. Assuming that the asymptotic risk R_1 has been found (by independent simulations, for example), one would now be interested in a large deviation type upper bound to $\mathbf{P}(R_1(v, Y^n, X^n) - R_1 > \epsilon)$. In [9] it has been shown that under the same assumptions as in the present paper, the following large deviation principle holds:

$$\lim_n \frac{1}{n} \ln \mathbf{P}(R_1(v, Y^n, X^n) > \epsilon + R_1) = -I(R_1 + \epsilon), \tag{1.7}$$

where I is a rate function and ϵ is small enough. The authors of [9] do not state the exact bound to the probability $\mathbf{P}(R_1(v, Y^n, X^n) - R_1 > \epsilon)$, but it could be derived from the proof. We would like to draw the reader's attention to the differences with supervised learning. In supervised learning (pattern recognition) the model is unknown, but the variables Y^n are observable, thus the empirical risk $R_1(g, Y^n, X^n)$ for any classifier could be calculated. The main object of interest then is the unknown asymptotic risk and the large deviation inequalities are used to estimate the unknown asymptotic risk by the known empirical risk. In our setting the data Y^n are hidden, but the model, and therefore the asymptotic risk, is known, so that the true risk can be used to estimate the unknown empirical risk.

The present paper deals mostly with convergence of the risks of Viterbi alignments. These results are all largely based on the regenerativity of the Viterbi process. The Viterbi process $\{V_t\}_{t=1}^{\infty}$ is an S-valued stochastic process that is in a sense the limit of the random vectors $v(X^n)$ as n grows. The existence of the Viterbi process is crucial and not obvious; our analysis is based on the results in [19, 18, 14], where the Viterbi process is constructed piecewisely. The piecewise construction under general assumptions is rather technical (see [19, 14]). However, when it is performed, the regenerativity of the Viterbi process as well as the ergodicity of the double-sided Viterbi process easily follow. The corresponding results and the construction of the Viterbi process are introduced in Subsection 2.2. The references to necessary results from the theory of regenerative processes are given in Subsection 2.1. We are following the coupling approach developed by Thorisson in [23]. 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 1). With a successful coupling, a general pathwise limit theorem for the Viterbi alignment (Theorem 3) can be proven. This is the main preliminary result and it can be used for many other purposes besides proving the convergence of risks.

Section 3 deals with the convergence of the R_1-risk. Section 4 deals with the convergence of the $\bar{R}_1$-risk and in Section 5, the convergence of the log-likelihood ($\bar{R}_\infty$-risk) is proven.

Since the regenerativity of the PMAP-process (the analogue of the Viterbi process for the PMAP-alignment) is not proven, the regenerativity-based methods cannot be used for the long-run analysis of PMAP-alignments. However, as shown in [16], the convergence of $R_1(X^n)$ (the R_1-risk of the PMAP-alignment) can be proven with a completely different method based on the exponential forgetting of smoothing probabilities. The exponential forgetting inequalities are introduced in Subsection 2.3, in Section 4 we show that they also imply the convergence of $\bar{R}_\infty(X^n)$ (the $\bar{R}_1$-risk of the PMAP-alignment).

From the discussion above it follows that there is no universal method known yet to prove the convergence of general risks and every optimal alignment needs a special treatment. For example, the convergence of $\bar{R}_C(X^n)$ (as well as of several other more general risks introduced in [13]) has not yet been proven, although it is reasonable to conjecture that it holds. Moreover, we conjecture that the dynamic programming algorithm for finding the minimizer of $\bar{R}_C$-risk together with the exponential smoothing could be used to find the $\bar{R}_C$-optimal alignment process piecewisely. If this is true, then the alignment process is regenerative and the results and methods in the present paper can be applied to many other optimal alignments.

2 Preliminary results

2.1 Regenerativity

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 [23]). Following the notation in [23], 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 a regenerative process 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, $S_0^o = T_1$. 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:

$$Eg(Z_1^*, Z_2^*, \ldots) = \frac{1}{ET_1}E\big[\sum_{t=0}^{T_1-1} g(\theta_t(Z^o))\big], \tag{2.1}$$

see, e.g. Theorem 2.1 and 2.2 of Chapter 10 in [23] or Theorem 6.1 in [11].

Recall that a stochastic process $Z = \{Z_t\}_{t=1}^\infty$ is *mixing in the sense of ergodic theory* if for every $A, B \in \sigma(\mathcal{Z}^\infty)$ (cylindrical σ-algebra) the following holds:

$$\lim_{t\to\infty} |\mathbf{P}(\theta_t Z \in A, Z \in B) - \mathbf{P}(\theta_t Z \in A)\mathbf{P}(Z \in B)| = 0, \tag{2.2}$$

see, e.g. [6]. Recall also that a sub-σ-algebra of $\mathcal{F}$ is called *trivial* if its elements have probability 1 or 0. In the following we consider two σ-algebras: the tail-σ-algebra $\mathcal{T} := \cap_{t=1}^\infty \theta_t^{-1}(\sigma(\mathcal{Z}^\infty))$ and the σ-algebra of shift-invariant sets $\mathcal{I} := \{A \in \sigma(\mathcal{Z}^\infty) : \theta_t^{-1}A = A\}$. A stationary $\mathcal{I}$-trivial process is ergodic. Since $\mathcal{I} \subseteq \mathcal{T}$ (see Section 5.1 in [23]), a stationary $\mathcal{T}$-trivial process (sometimes also called regular) is also ergodic. The following version of Theorem 3.3 of Chapter 10 in [23] states that an aperiodic positive recurrent regenerative process can be successfully coupled with a stationary ergodic process.

Theorem 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.*

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

b) *The processes Z and Z' are $\mathcal{T}$-trivial.*

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 [23] are fulfilled. The claim **a)** is claim **a)** of that theorem, the $\mathcal{T}$-triviality of Z is 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 $\mathcal{T}$-trivial. □

Corollary 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*

$$\frac{1}{n}\sum_{t=1}^n g(Z_t, Z_{t+1}, \ldots) \to E[g(Z_1^*, Z_2^*, \ldots)] \quad \textit{a.s. and in } L_1. \tag{2.3}$$

Proof. Let us extend the space $(\Omega, \mathcal{F}, \mathbf{P})$ so that the statements of Theorem 1 hold. Then the process Z' is stationary and ergodic having the same distribution as Z^*. By Birkhoff's ergodic theorem then,

$$\frac{1}{n}\sum_{t=1}^{n} g(Z'_t, Z'_{t+1}, \ldots) \to E[g(Z'_1, Z'_2, \ldots)] = E[g(Z^*_1, Z^*_2, \ldots)] \quad \text{a.s. and in } L_1. \tag{2.4}$$

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 (2.3). The L_1-convergence follows from applying Scheffe's lemma to $g^+(Z_t, Z_{t+1}, \ldots)$ and $g^-(Z_t, Z_{t+1}, \ldots)$ separately. □

Remark: If (Z, S) is positive recurrent but not aperiodic, then Theorem 1 cannot be applied. However, using Theorem 2.2 of [23] 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 $\mathcal{I}$-trivial and hence ergodic, thus Corollary 1 still holds. In this paper we consider only aperiodic regenerative processes.

If $f : \mathcal{Z} \to \mathbb{R}$ is measurable, then the convergence (2.3) together with (2.1) yields

$$\frac{1}{n}\sum_{t=1}^{n} f(Z_t) \to Ef(Z^*_1) = \frac{1}{ET_1} E\big[\sum_{t=1}^{T_1} f(Z^o_t)\big] = \frac{1}{ET_1} E\big[\sum_{t=S_0+1}^{S_1} f(Z_t)\big] \tag{2.5}$$

a.s. and in L_1.

2.2 Infinite Viterbi alignment

2.2.1 One-sided process

Definition 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*

$$g : \mathcal{X}^\infty \to S^\infty \tag{2.6}$$

such that for almost every realization $x^\infty \in \mathcal{X}^\infty$ the following statement holds: for every $k \in \mathbb{N}$ there exists $m \geq k$ (depending on x^∞) 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.

The existence of an infinite alignment is in general 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 is trivial if every observation is classified independently. The existence of an infinite alignment allows to study asymptotic properties of the alignment. Usually it is done via the corresponding *alignment process* $\{G_t\}_{t=1}^\infty := g(X)$.

In the following, we consider the existence of infinite Viterbi alignments. Under rather restrictive assumptions on HMMs the existence of an infinite Viterbi alignment

was first proven in [3]. In [19], the existence of an infinite Viterbi alignment was proven 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 $G_s := \{x \in \mathcal{X} : f_s(x) > 0\}$.

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

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

Hence, a cluster is a maximal subset of states such that $G_C = \cap_{s \in C} G_s$, 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.

Let C be a cluster. The existence of C implies the existence of a set $\mathcal{X}_o \subset \cap_{s \in C} G_s$ and $\epsilon > 0$, $M < \infty$ such that $\mu(\mathcal{X}_o) > 0$, and $\forall x \in \mathcal{X}_o$ the following statements hold: (i) $\epsilon < \min_{s \in C} f_s(x)$; (ii) $\max_{s \in C} f_s(x) < M$; (iii) $\max_{s \notin C} f_s(x) = 0$. For proof, see [19].

In the following, we introduce two assumptions on HMMs that are needed for the existence of an infinite Viterbi alignment.

A1 (cluster-assumption): There exists a cluster $C \subset S$ such that the substochastic 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 $l \in S$,

$$P_l\left(\left\{x \in \mathcal{X} : \; f_l(x) p_l^* > \max_{s, s \neq l} f_s(x) p_s^*\right\}\right) > 0, \quad p_l^* = \max_j p_{j,l}, \, \forall l \in S. \quad (2.7)$$

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 $s \in S$, $G_s = \mathcal{X}$. Thus, **A1** is more general than the *strong mixing condition* (Assumption 4.2.21 in [4]) and also weaker than Assumption 4.3.29 in [4]. Note that **A1** implies the aperiodicity of Y, but not vice versa.

The assumption **A2** is more technical in nature. In [14] it was shown that for a two-state HMM, (2.7) 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 [18, 19]. 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 [18, 19, 16, 14].

In the following, let $\tilde{V}^n = v^n(X^n)$, where v^n is a finite Viterbi alignment. The results of the present paper are largely based on the following theorem, which has been proved in [19, 18]. See also Lemma 2.1 in [9].

Theorem 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 $\{S_t\}_{t=0}^{\infty}$;*

R2 *there exists a nonnegative integer $m < \infty$ such that for every $j \geq 0$, $\tilde{V}_t^n = V_t$ for all $n \geq S_j + m$ and $t \leq S_j$.*

Proof. The required infinite alignment is constructed piecewisely. The construction and main proof are given in [19]. The piecewise construction guarantees **R2**. The regenerativity and positive recurrence is shown in Section 4 of [18]. The aperiodicity follows from the aperiodicity of Y that follows from **A1**. □

In what follows, we always assume that the finite Viterbi alignments $v^n : \mathcal{X}^n \to S^n$ are chosen according to Theorem 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 $\tilde{Z}^n := \{(\tilde{V}_t^n, X_t, Y_t)\}_{t=1}^n$ satisfies by **R2** the following property: $\tilde{Z}_t^n = Z_t$ for every $t = 1, \ldots, S_{k(n)}$, where $k(n) = \max\{k \geq 0 : S_k + m \leq n\}$.

Let $p \in \mathbb{N}$ and $g_p : \mathcal{Z}^p \to \mathbb{R}$ be measurable. We 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

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

The random variables $M_p, M_{p+1}, \ldots$ are identically distributed, but for $p > 1$ not necessarily independent.

The following theorem generalizes Theorem 3.1 of Chapter VI in [1]. The proof is based on the same argument. Recall that Z^* is a stationary version of Z.

Theorem 3. *Let g_p be such that $EM_p < \infty$ and $E|g_p(Z_1^*, \ldots, Z_p^*)| < \infty$. Then*

$$\frac{1}{n-p+1} \sum_{i=p}^{n} \tilde{U}_i^n \to EU_p = Eg_p(Z_1^*, \ldots, Z_p^*) \quad \textit{a.s. and in } L_1. \tag{2.8}$$

Proof.

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

Since $S_{k(n)} \nearrow \infty$ a.s., from (2.3) we know that

$$\frac{1}{S_{k(n)}}\sum_{i=p}^{S_{k(n)}} U_i \to Eg_p(Z_1^*,\dots,Z_p^*) \quad \text{a.s. and in } L_1. \tag{2.9}$$

Note that

$$\frac{S_{k(n)}}{n-p+1} = \frac{S_{k(n)}}{k(n)}\frac{k(n)}{n-p+1}.$$

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

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

Combining this with (2.9) 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

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

The theorem is proven if we can show that, as $k \to \infty$,

$$\frac{M_k}{k} \to 0 \quad \text{a.s. and in } L_1.$$

By the Borel-Cantelli lemma this holds if for every $\epsilon > 0$,

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

because the random variables M_k, $k \geq p$, are indentically distributed. Clearly, $E[\frac{M_k}{k}] \to 0$, so by Scheffe's theorem $\frac{M_k}{k} \to 0$ in L_1 as well. □

2.2.2 Double-sided infinite Viterbi alignment

Definition 2. *Let for every $z_1, z_2 \in \mathbb{Z}$, $g_{z_1}^{z_2} : \mathcal{X}^{[z_1,z_2]} \to S^{[z_1,z_2]}$ 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*

$$g : \mathcal{X}^{\mathbb{Z}} \to S^{\mathbb{Z}} \tag{2.10}$$

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 $m \geq k$ (depending on $x_{-\infty}^{\infty}$) 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.

The piecewise construction of the infinite Viterbi alignment allows the double-sided extension as well.

Theorem 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 $\{S_t\}_{t=-\infty}^{\infty}$;*

RD2 *there exists a nonnegative integer $m < \infty$ such that for every $j \geq 0$, $\tilde{V}_t^n = V_t$ for all $n \geq S_j + m$ and $S_0 \leq i \leq S_j$;*

RD3 *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** and **RD2** is the same as in Theorem 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 [19]). □

In the following, the finite Viterbi alignments $v_{z_1}^{z_2}$ are chosen to be consistent. The property **RD3** is important. Since X is an ergodic process, from **RD3** 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. Let Z^* denote the restriction of $\{(X_t, Y_t, V_t)\}_{t=-\infty}^{\infty}$ to the nonnegative integers, i.e. $Z^* = \{(X_t, Y_t, V_t)\}_{t=1}^{\infty}$. By **RD2**, the restriction of Z^* is a stationary version of Z as in **R1**. Thus $(X_0, Y_0, V_0) \stackrel{D}{=} (X_1^*, Y_1^*, V_1^*) = Z_1^*$ and in the following, we shall often use this. Note that the one-sided Viterbi process V in **R1** is not defined at time zero so that the random variable V_0 always implies the double-sided, hence stationary, case.

2.3 Smoothing probabilities

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 [16] it is shown that under **A1** these probabilities satisfy the following exponential forgetting inequalities:

$$\|\mathbf{P}(Y_t\in\cdot|X_1^{\infty})-\mathbf{P}(Y_t\in\cdot|X_{-\infty}^{\infty})\|\le C\rho^t \quad \text{a.s.}, \tag{2.11}$$

$$\|\mathbf{P}(Y_t\in\cdot|X_1^{\infty})-\mathbf{P}(Y_t\in\cdot|X_1^{n})\|\le C\rho^{n-t} \quad \text{a.s.}, \tag{2.12}$$

where $t\ge 1$, C is a finite positive random variable and $\rho\in(0,1)$. Here $\|\cdot\|$ stands for the total variation distance.

In what follows, we shall use the notation $p_t(j|x_{-\infty}^{\infty}):=\mathbf{P}(Y_t=j|X_{-\infty}^{\infty}=x_{-\infty}^{\infty})$.

3 Convergence of R_1-risk

Let the loss function be defined as in (1.4) and let v^n be a consistently chosen Viterbi alignment. If the underlying Markov chain would not be hidden, the *empirical risk of the Viterbi alignment* could be directly calculated as follows:

$$R_1(Y^n,X^n)=\frac{1}{n}\sum_{t=1}^{n} l(Y_t,v_t^n(X^n))=\frac{1}{n}\sum_{t=1}^{n} l(Y_t,\tilde V_t^n). \tag{3.1}$$

The conditional expectation of $R_1(Y^n,X^n)$ given X^n is the random variable $R_1(v,X^n)=E[R_1(Y^n,\tilde V^n)|X^n]$. Since S is finite and $l:S\times S\to\mathbb{R}$ is bounded, from Theorem 3 and (2.5) it follows that

$$R_1(Y^n,\tilde V^n)\to El(Y_0,V_0)=\frac{1}{ET_1}E\Big(\sum_{t=S_0+1}^{S_1} l(Y_t,V_t)\Big)=:R_1 \quad \text{a.s. and in } L_1. \tag{3.2}$$

We shall call the constant R_1 *asymptotic Viterbi risk*. It depends on the model (Y,X) and on the loss function l, only. For $l(s,s')=I_{\{s'\neq s\}}$, the actual risk is the average number of mistakes made by the Viterbi alignment:

$$R_1(Y^n,\tilde V^n)=\frac{1}{n}\sum_{t=1}^{n} I_{\{Y_t\neq\tilde V_t^n\}}, \tag{3.3}$$

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

To show the convergence of $R_1(v,X_n)$ we use the following lemma (see Theorem 9.4.8 in [5]). To our knowledge, the idea of considering the R_1-type limits for the Viterbi alignment has been first mentioned in [2], the convergence of the empirical risk is also stated in [9].

Lemma 1. *Let X_n be bounded random variables such that $X_n \to 0$ a.s. Let $\{\mathcal{F}_n\}_{n=1}^{\infty}$ be a filtration. Then*

$$E[X_n|\mathcal{F}_n] \to 0 \quad a.s. \tag{3.4}$$

The following theorem is the first main result of this paper. A similar result for the PMAP-alignment, namely the convergence of $R_1(X^n)$ to a constant, is proven in [16].

Theorem 5. *Let $\{(Y_t, X_t)\}_{t=1}^{\infty}$ be an ergodic HMM satisfying* **A1** *and* **A2**. *Then there exists a constant $R_1 \geq 0$ such that the empirical risk and the risk of the Viterbi alignment both converge to R_1 almost surely and in L_1:*

$$\lim_n R_1(Y^n, X^n) = \lim_n R_1(v, X^n) = R_1 \quad a.s. \text{ and in } L_1. \tag{3.5}$$

Moreover, the expected risk of Viterbi alignments converges to R_1 as well: $ER_1(v, X^n) \to R_1$.

Proof. The convergence of the empirical risk is (3.2). To show that $R_1(v, X^n) \to R_1$ a.s., apply Lemma 1 with $X_n := R_1(Y^n, X^n) - R_1$. Clearly, $R_1(Y^n, X^n) - R_1$ is bounded and by (3.2) it goes to 0 a.s. Thus, by (3.4),

$$|E[R_1(Y^n, X^n) - R_1|X^n]| = |E[R_1(Y^n, X^n)|X^n] - R_1| = |R_1(v, X^n) - R_1| \to 0 \quad \text{a.s.}$$

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

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

For the convergence of $\bar{R}_1$ we use Theorem 4. Recall that the double-sided infinite alignment v is a stationary coding. Consider the function $f : \mathcal{X}^{\mathbb{Z}} \to S^{\mathbb{Z}}$, 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 $t = 1, 2, \ldots$,

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

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

$$-\frac{1}{n}\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty}) \to -E\big(\ln \mathbf{P}(Y_0 = V_0 | X_{-\infty}^{\infty})\big) =: \bar{R}_1 \quad \text{a.s. and in } L_1, \tag{4.1}$$

provided the expectation is finite. Recall the inequalities (2.11) and (2.12). Unfortunately these bounds do not immediately hold for the logarithms. The following lemma uses the inequality $|\ln a - \ln b| \leq \frac{1}{\min\{a,b\}}|a - b|$, provided that $a, b > 0$.

Lemma 2. *Suppose that for an $\alpha > 0$*

$$E\Big(\frac{1}{\mathbf{P}(Y_0 = V_0|X_{-\infty}^{\infty})}\Big)^{\alpha} < \infty\,. \tag{4.2}$$

Then

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = V_t|X^n) \to \bar{R}_1 \quad \text{a.s.} \tag{4.3}$$

Proof. Let $\xi_t := \mathbf{P}(Y_t = V_t|X_{-\infty}^{\infty})$, $\eta_t^n := \mathbf{P}(Y_t = V_t|X_1^n)$, $\eta_t := \mathbf{P}(Y_t = V_t|X_1^{\infty})$ and let $\beta = \frac{1}{\alpha}$. Recall that $\{\xi_t\}$ is a stationary ergodic process. The assumption (4.2) ensures that $E|\ln \xi_0| < \infty$. Hence, by assumption,

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

Thus, the sequence ξ_t, $t = 1, 2, \ldots$, satisfies $\mathbf{P}(\xi_t > \frac{1}{t^{\beta}} \quad \text{ev}) = 1$. From (2.11) it follows that $\mathbf{P}(\eta_t > \frac{1}{2t^{\beta}} \quad \text{ev}) = 1$. Thus, almost surely $|\ln \eta_t - \ln \xi_t| \le C2t^{\beta}\rho^t$ eventually. Since $-\frac{1}{n}\sum_{t=1}^{n} \ln \xi_t \to \bar{R}_1$ a.s., we now have

$$-\frac{1}{n}\sum_{t=1}^{n} \ln \eta_t \to \bar{R}_1 \quad \text{a.s.} \tag{4.4}$$

Let $m = \lfloor \frac{n}{2} \rfloor$ and note that by (2.12) it holds that $|\eta_t^n - \eta_t| \le C\rho^{n-t} \le C\rho^m \le C\rho^t$ a.s., provided $t = 1, \ldots, m$. Let (random) T be so big that $\eta_t > \frac{1}{2t^{\beta}}$ when $t > T$. Let (random) M be so big that $C\rho^m < \frac{1}{4m^{\beta}}$ when $m > M$. Thus, for $T < t \le m$ and $m > M$ it holds that $C\rho^m < \frac{1}{4m^{\beta}} \le \frac{1}{4t^{\beta}}$. The inequality (2.12) then implies that $\min\{\eta_t, \eta_t^n\} \ge \frac{1}{4t^{\beta}}$ and $|\ln \eta_t^n - \ln \eta_t| \le (4t^{\beta}C)\rho^m$. Hence, as m goes to infinity,

$$\begin{aligned}\left|\frac{1}{m}\sum_{t=1}^{m} \ln \eta_t^n - \frac{1}{m}\sum_{t=1}^{m} \ln \eta_t\right| &\le \frac{1}{m}\sum_{t=1}^{T} |\ln \eta_t^n - \ln \eta_t| + \frac{1}{m}\sum_{t=T+1}^{m} |\ln \eta_t^n - \ln \eta_t| \\ &\le \frac{1}{m}\sum_{t=1}^{T} |\ln \eta_t^n - \ln \eta_t| + \frac{1}{m}\sum_{t=T+1}^{m} (4t^{\beta}C)\rho^m \to 0\,.\end{aligned}$$

From (4.4) it now follows that $-\frac{1}{m}\sum_{t=1}^{m} \ln \eta_t^n \to \bar{R}_1$ a.s. In other words, we have proven that

$$-\frac{2}{n}\sum_{t=1}^{\lfloor \frac{n}{2} \rfloor} \ln \mathbf{P}(V_t = Y_t|X^n) \to \bar{R}_1 \quad \text{a.s.}$$

By **RD3**, the process (X, Y, V) is stationary. Hence, for every n,

$$-\frac{2}{n}\sum_{t=\lfloor \frac{n}{2} \rfloor + 1}^{n} \ln \mathbf{P}(V_t = Y_t|X^n) \stackrel{D}{=} -\frac{2}{n}\sum_{t=1}^{\lfloor \frac{n}{2} \rfloor} \ln \mathbf{P}(V_t = Y_t|X^n). \tag{4.5}$$

Thus the left hand side of (4.5) tends to $\bar{R}_1$ a.s. as well, so that

$$\lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{n} \ln \mathbf{P}(V_t = Y_t|X^n) \to \bar{R}_1 \quad \text{a.s.} \tag{4.6}$$

□

Let C be the cluster that satisfies the assumptions of **A1** and let $\mathcal{X}_o$ be the corresponding set. The following proposition is proven in Appendix.

Proposition 1. *Let $x_{-\infty}^{\infty} \in \mathcal{X}^{\mathbb{Z}}$ be such that for some $u, v \in \mathbb{N}$, $x_{-u}^{-u+r} \in \mathcal{X}_o^{r+1}$, $x_{v-r}^{v} \in \mathcal{X}_o^{r+1}$ 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 $0 < B < \infty$ such that*

$$p_0\big(v_0|x_{-\infty}^{\infty}\big) \geq c\exp[-B(u+v)]. \tag{4.7}$$

Lemma 3. *There exists $\alpha > 0$ such that for every $t \in \mathbb{Z}$,*

$$E\Big(\frac{1}{\mathbf{P}(Y_t = V_t|X_{-\infty}^{\infty})}\Big)^{\alpha} < \infty\,. \tag{4.8}$$

Proof. Let U and V be the following stopping times:

$$U = \min\{t \geq r+1 : X_{-t}^{-t+r} \in \mathcal{X}_o^{r+1}\}, \quad V = \min\{t \geq r+1 : X_{t-r}^{t} \in \mathcal{X}_o^{r+1}\}. \tag{4.9}$$

Since X is stationary, the stopping times U and V are identically distributed. Because for every $s \in S$, $\lim_n \mathbf{P}(Y_0 = s|X_{-n}^n) = \mathbf{P}(Y_0 = s|X_{-\infty}^{\infty})$ a.s., from the inequality (4.7) it follows that

$$\mathbf{P}(Y_0 = V_0|X_{-\infty}^{\infty}) \geq c\exp[-B(U+V)] \quad \text{a.s.} \tag{4.10}$$

It is not hard to see that for some positive constants a and b and for every $k = 1, 2, \ldots$,

$$\mathbf{P}(U > k) \leq a\exp(-bk),$$

see, e.g. [9]. This inequality implies that for $\alpha > 0$ small enough, $E(e^{\alpha U}) < \infty$. By the Cauchy-Schwartz inequality, for sufficiently small α,

$$E\big(e^{\alpha(U+V)}\big) = E\big(e^{\alpha U}e^{\alpha V}\big) \leq \Big(E\big(e^{2\alpha U}\big)E\big(e^{2\alpha V}\big)\Big)^{\frac{1}{2}} < \infty. \tag{4.11}$$

The inequalities (4.10) and (4.11) imply (4.8) for $t = 0$. By the stationarity of (X, Y), (4.8) holds for arbitrary t. □

The proof of Proposition 1 reveals that it holds also for a finite sequence of observations x^n. Moreover, the following corollary holds.

Corollary 2. *Let $x^n \in \mathcal{X}^n$ be such that for some $w < n - r$, $x_w^{w+r} \in \mathcal{X}_o^{r+1}$. Let $\tilde{v}_t = v_t(x^n)$. Then there exist $c > 0$ and $0 < D < \infty$ such that for every t, $w < t \leq n$,*

$$p_t(\tilde{v}_t|x^n) \geq c\exp[-D(n-w)]. \tag{4.12}$$

The proof of Corollary 2 follows the one of Proposition 1 and is sketched in Appendix.

Theorem 6. *Let $\{(Y_t, X_t)\}_{t=1}^{\infty}$ be an ergodic HMM satisfying* **A1** *and* **A2**. *Then there exists a constant $\bar{R}_1$ such that*

$$\lim_{n\to\infty} \bar{R}_1(v, X^n) = \lim_{n\to\infty} -\frac{1}{n}\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) \to \bar{R}_1 \quad \text{a.s. } \textit{and in } L_1. \tag{4.13}$$

Proof. Without loss of generality, we can consider a double-sided HMM $\{(Y_t, X_t)\}_{t=-\infty}^{\infty}$ 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 $\{S_t\}_{t\ge 0}$ is the renewal process as in Theorem 4. Thus,

$$\sum_{t=1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) =$$
$$\sum_{t=1}^{S_0-1} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) + \sum_{t=S_0}^{S_{k(n)}} \ln \mathbf{P}(Y_t = V_t | X^n) + \sum_{t=S_{k(n)}+1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n).$$

By (4.3), it suffices to prove that

$$\frac{1}{n}\sum_{t=S_{k(n)}+1}^{n} \ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n) \to 0 \quad \text{a.s.} \tag{4.14}$$

For every $k \ge 0$, let

$$M_k = \max_{S_k < i \le S_{k+1}} |\ln \mathbf{P}(Y_{S_k+1} = \tilde{V}_{S_k+1}^i | X^n) + \cdots + \ln \mathbf{P}(Y_i = \tilde{V}_i^i | X^n)|.$$

The random variables M_k are iid. As in the proof of Theorem 3, for (4.14) it now suffices to show that $EM_k < \infty$ for every $k \ge 0$.

We shall consider S_1. The construction of S_k implies that there exists an integer m such that $m > r+1$ and for every k, the observations $X_{S_k-m}, \ldots, X_{S_k-m+r}$ belong to $\mathcal{X}_o$ (see [19]). Recall that we are considering the case $n \le S_2$. Hence, for every t such that $S_1 < t \le n$, by (4.12),

$$|\ln \mathbf{P}(Y_t = \tilde{V}_t^n | X^n)| \le D(n - S_1 + m) \le D(S_2 - S_1 + m),$$

implying that

$$|M_1| \le D(S_2 - S_1 + m)^2. \tag{4.15}$$

The renewal times $S_2 - S_1$ have all moments (see [9, 18]), hence $EM_1 < \infty$. □

Remark. Note that the approach of the present section can be easily applied to prove the convergence of the R_1-risk: $R_1(v, X^n) \to R_1$ a.s. Indeed, the counterpart of (4.1) is

$$\frac{1}{n}\sum_{t=1}^{n} \mathbf{P}(Y_t = V_t | X_{-\infty}^{\infty}) \to E\big(\mathbf{P}(Y_0 = V_0 | X_{-\infty}^{\infty})\big) =: 1 - R_1 \quad \text{a.s.and in } L_1. \tag{4.16}$$

The inequalities (2.11) and (2.12) immediately imply

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

and 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. Then $\max_s\mathbf{P}(Y_t=s|X^n)\geq|S|^{-1}$. By Birkhoff's theorem,

$$-\frac{1}{n}\sum_{t=1}^{n}\max_{s\in S}\ln\mathbf{P}(Y_t=s|X_{-\infty}^{\infty})\to\bar{R}_1^*\quad\text{a.s. and in } L_1\,,\tag{4.17}$$

where $\bar{R}_1^*$ is a constant. The inequalities (2.11) and (2.12) imply that

$$|\max_s\ln\mathbf{P}(Y_t=s|X^n)-\max_s\ln\mathbf{P}(Y_t=s|X_{-\infty}^{\infty})|\leq C|S|(\rho^n+\rho^{n-t})\quad\text{a.s.}$$

Thus, the convergence (4.17) implies the convergence

$$\bar{R}_1(X^n)=-\frac{1}{n}\sum_{t=1}^{n}\max_{s\in S}\ln\mathbf{P}(Y_t=s|X^n)\to\bar{R}_1^*\quad\text{a.s. and in } L_1.\tag{4.18}$$

Hence, the following corollary holds.

Corollary 3. *There exists a constant $\bar{R}_1^*$ such that (4.18) holds.*

5 Convergence of log-likelihood

Let Q_s be the conditional measure $\mathbf{P}(X_0\in\cdot|V_0=s)$, $s\in S$. As it follows from Theorem 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,

$$\frac{\sum_{t=1}^{n}I_{A\times s}(X_t,\tilde{V}_t^n)}{\sum_{t=1}^{n}I_s(\tilde{V}_t^n)}\to Q_s(A)\quad\text{a.s.}\tag{5.1}$$

This convergence is the basis of the adjusted Viterbi training introduced in [17, 18]. For every Q_s-integrable g,

$$E\big(g(X_0)I_s(V_0)\big)=E\big(g(X_0)|V_0=s\big)\mathbf{P}(V_0=s)=m_s\int g(x)Q_s(dx),\tag{5.2}$$

where $m_s := \mathbf{P}(V_0 = s)$.

Suppose now that the logarithms of the conditional densities f_s are Q_s-integrable for every s. As shown in [12], this holds if $\ln f_s$ is P_s-integrable. Then, by Theorem 3 and (5.2), for every state $s \in S$

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n} \ln f_s(X_t) I_s(\tilde{V}_t^n) &\to E\big(\ln f_s(X_0) I_s(V_0)\big) \\ &= m_s \int \ln f_s(x) Q_s(dx) \quad \text{a.s. and in } \mathrm{L}_1. \end{aligned} \tag{5.3}$$

Let $p(x^n|s^n)$ be the conditional likelihood of observing x^n given that $\{Y^n = s^n\}$. Thus,

$$\ln p(x^n|s^n) = \sum_{t=1}^{n} \ln f_{s_t}(x_t) = \sum_{t=1}^{n} \ln f_1(x_t) I_1(s_t) + \cdots + \sum_{t=1}^{n} \ln f_{|S|}(x_t) I_{|S|}(s_t). \tag{5.4}$$

Applying (5.4) to the Viterbi alignment, from the convergence (5.3) follows that under the assumption that $\ln f_s$ is Q_s-integrable,

$$\frac{1}{n} \ln p(X^n | Y^n = v(X^n)) \to \sum_{s \in S} m_s \int \ln f_s(x) Q_s(dx) \quad \text{a.s. and in } L_1. \tag{5.5}$$

Recall that $\bar{R}_\infty(X^n) = -\frac{1}{n} \ln \mathbf{P}(Y^n = \tilde{V}^n | X^n)$.

Theorem 7. *Let for every $s \in S$ the function $\ln f_s$ be P_s-integrable. Then*

$$-\bar{R}_\infty(X^n) \to \sum_{s \in S} m_s \int \ln f_s(x) Q_s(dx) + E[\ln p_{V_1^* V_2^*}] + H_X =: -\bar{R}_\infty \quad \text{a.s } \textit{and in } \mathrm{L}_1, \tag{5.6}$$

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

Proof. Let $p(x^n)$ be the likelihood of x^n. Then

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

Thus,

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

The first term of the RHS converges by (5.5). For the second term use the Markov property

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

where $\pi_s = \mathbf{P}(Y_1 = s)$. Since $\tilde{V}^n$ is a path with positive likelihood, $p_{\tilde{V}_t^n, \tilde{V}_{t+1}^n} > 0$ a.s. for every t. Because the number of states is finite, there exists a constant $M > 0$ such that for every i,

$$-\ln p_{\tilde{V}_i^n \tilde{V}_{i+1}^n} < M \quad \text{a.s.}$$

Hence, the assumptions of Theorem 3 hold and, with $p_{\tilde{V}_0^n \tilde{V}_1^n} = \pi_{\tilde{V}_1^n}$, we get

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

Finally, by the Shannon-McMillan-Breiman theorem,

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

□

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

$$\frac{1}{n} \ln \mathbf{P}(Y^n | X^n) \to \sum_{s \in S} \pi_s \int \ln f_s(x) P_s(dx) - H_Y + H_X =: -\bar{R}_\infty^Y \quad \text{a.s. and in } L_1, \tag{5.7}$$

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,

$$\bar{R}_\infty \leq \bar{R}_\infty^Y, \tag{5.8}$$

and the difference of those two numbers shows how much the Viterbi alignment "over-estimates" the likelihood.

Acknowledgements

J. Lember is supported by Estonian science foundation grant no 7553.

References

[1] S. Asmussen. *Applied Probability and Queues.* Springer, 2003.

[2] A. Caliebe. Properties of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory*, 52(1):41–51, 2006.

[3] A. Caliebe and U. Rösler. Convergence of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory*, 48(7):1750–1758, 2002.

[4] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer, 2005.

[5] K.-L. Chung. *A Course in Probability Theory*. Academic Press, 1974.

[6] P. Doukhan. *Mixing: Properties and Examples*. Springer, 1994.

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

[8] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Trans. Inform. Theory*, 48(6):1518–1569, 2002.

[9] A. Ghosh, E. Kleiman, and A. Roitershtein. Large deviation bounds for functionals of Viterbi paths. http://www.public.iastate.edu/∼roiterst/papers.html, 2009.

[10] F. Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, 2001.

[11] V. Kalashnikov. *Topics on Regenerative Processes*. CRC, 1994.

[12] A. Koloydenko, M. Käärik, and J. Lember. On adjusted Viterbi training. *Acta Appl. Math.*, 96(1-3):309–326, 2007.

[13] A. Koloydenko and J. Lember. Segmentation with hidden Markov model. In preparation.

[14] A. Koloydenko and J. Lember. Infinite Viterbi alignments in the two-state hidden Markov models. *Acta Comment. Univ. Tartu. Math.*, 12:109–124, 2008.

[15] T. Koski. *Hidden Markov Models for Bioinformatics*. Kluwer Academic Publishers, Dordrecht, 2001.

[16] J. Lember. On approximation of smoothing probabilities for hidden Markov models. http://arxiv.org/abs/0910.4636, 2009. Submitted.

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

[18] J. Lember and A. Koloydenko. The Adjusted Viterbi training for hidden Markov models. *Bernoulli*, 14(1):180–206, 2008.

[19] J. Lember and A. Koloydenko. A constructive proof of the existence of Viterbi processes. *IEEE Trans. Inform. Theory*, 2010. To appear.

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

[21] F. Och and H. Ney. Improved statistical alignment models. In *Proc. 38th Ann. Meet. Assoc. Comput. Linguist.*, pages 440 – 447. Assoc. Comput. Linguist., 2000.

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

[23] H. Thorisson. *Coupling, Stationarity, and Regeneration*. Springer, 2000.

A Appendix

A.1 Preliminaries

Let us start with some notation. 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,

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

We also define

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

The last two notations are standard in the HMM literature, see e.g. [8, 4]. Let

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

Finally, let

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

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

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

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

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

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, the inequalities (A.1) imply that for every state j

$$\frac{\sigma(x^m, j|i_1)}{\sigma(x^m, j|i_2)} \leq \frac{\max_{y^r \in C^r : y_r = s_1} p(x^r, y^r|i_1)}{\max_{y^r \in C^r : y_r = s_1} p(x^r, y^r|i_2)} \leq \frac{A}{a}. \tag{A.2}$$

Similarly, if x^m is such that the last r elements belong to $\mathcal{X}_o$, i.e. $x_{m-r+1}^m \in \mathcal{X}^r$, 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.1) it follows that

$$\frac{\sigma(x^m, j_1|i)}{\sigma(x^m, j_2|i)} \leq \frac{\sigma(x_{m-r+2}^m, j_1|s_1)}{\sigma(x_{m-r+2}^m, j_2|s_1)} \leq \frac{A}{a}. \tag{A.3}$$

A.2 Proof of Proposition 1

Proof. 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 $u, v \in \{1, \dots, n-1\}$ and for an arbitrary state, let it be 1,

$$\begin{aligned}\gamma_0(x_{-n}^n, i) &= \sum_{s_1, s_2, s_3, s_4 \in S} \alpha(x_{-n}^{-u}, s_1)\beta(x_{-u+1}^{-1}, s_2|s_1)\times\\ &\qquad \times P(s_2, 1)f_1(x_0)\beta(x_1^{v-1}, s_3|1)P(s_3, s_4)\alpha(s_4, x_v^n)\\ &\geq \sum_{s_1, s_4 \in S} \alpha(x_{-n}^{-u}, s_1)\sigma(x_{-u+1}^{-1}, 1|s_1)f_1(x_0)\sigma(x_1^{v-1}, s_4|1)\alpha(s_4, x_v^n)\\ &\geq p(x_{-n}^{-u})\big(\min_s \sigma(x_{-u+1}^{-1}, 1|s)\big)f_1(x_0)\big(\min_s \sigma(x_1^{v-1}, s|1)\big)p(x_v^n).\end{aligned}$$

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

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

implying that for every state i_o,

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

Thus,

$$\begin{aligned}\gamma_0(x_{-n}^n, 1) \geq p(x_{-n}^{-u}) \frac{\big(\min_s \sigma(x_{-u+1}^{-1}, 1|s)\big)}{\sigma(x_{-u+1}^{-1}, 1|a)} \sigma(x_{-u+1}^{-1}, i_o|a) f_{i_o}(x_0) \times \\ \times \sigma(x_1^{v-1}, b|i_o) \frac{\big(\min_s \sigma(x_1^{v-1}, s|1)\big)}{\sigma(x_1^{v-1}, b|1)} p(x_v^n). \qquad \text{(A.4)}\end{aligned}$$

Note that for every x_k^m,

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

Therefore, for every $i_o \in S$

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

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

$$\begin{aligned}\gamma_0(x_{-n}^n, i_o) \quad \leq \quad & p(x_n^{-u}) |S|^{u-1} \big(\max_{s \in C} \sigma(x_{-u+1}^{-1}, i_o|s)\big) \times \\ & \times f_{i_o}(x_0) |S|^{v-1} \big(\max_{s \in C} \sigma(x_1^{v-1}, s|i_o)\big) p(x_v^n). \qquad \text{(A.5)}\end{aligned}$$

The same holds for (A.4), implying that

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

The inequalities (A.2) and (A.3) imply that the ratios above are bounded below by $\frac{a}{A}$ that does not depend on the observations. Thus, there exist constants $c_1 := \big(\frac{a}{A}\big)^4$ and $0 < B < \infty$ (not depending on the data) such that for every state i_o,

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

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

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

Because $p_0(1|x_{-n}^n) \to p_0(1|x_{-\infty}^{\infty})$, the inequality (4.7) follows by taking $c = \frac{c_1}{|S|}$. □

A.3 Proof of Corollary 2

Proof. The proof is analogous to the proof of Proposition 1. Using the same notations we obtain that for every t, $w < t < n$,

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

For every $i_o \in S$,

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

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

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

Thus,

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

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