

Large deviation bounds for functionals of Viterbi paths

Arka P. Ghosh* Elizabeth Kleiman† Alexander Roitershtein‡

February 16, 2009; Revised December 18, 2009

Abstract

In a number of applications, the underlying stochastic process is modeled as a finite-state discrete-time Markov chain that cannot be observed directly and is represented by an auxiliary process. The maximum *a posteriori* (MAP) estimator is widely used to estimate states of this hidden Markov model through available observations. The MAP path estimator based on a finite number of observations is calculated by the Viterbi algorithm, and is often referred to as the Viterbi path. It was recently shown in [2, 3] and [16, 17] (see also [12, 15]) that under mild conditions, the sequence of estimators of a given state converges almost surely to a limiting regenerative process as the number of observations approaches infinity. This in particular implies a law of large numbers for some functionals of hidden states and finite Viterbi paths. The aim of this paper is to provide the corresponding large deviation estimates.

MSC2000: primary 60J20, 60F10; secondary 62B15, 94A15.

Keywords: hidden Markov models, maximum *a posteriori* path estimator, Viterbi algorithm, large deviations, regenerative processes.

1 Introduction and statement of results

Let $(X_n)_{n \geq 0}$ be a discrete-time irreducible and aperiodic Markov chain in a finite state space $\mathcal{D} = \{1, \dots, d\}$, $d \in \mathbb{N}$. We interpret $X = (X_0, X_1, \dots)$ as an underlying state process that cannot be observed directly, and consider a sequence of accessible observations $Y = (Y_0, Y_1, \dots)$ that serves to produce an estimator for X . For instance, Y_n can represent the measurement of a signal X_n disturbed by noise. Given a realization of X , the sequence Y is formed by independent random variables Y_n valued in a measurable space $(\mathcal{Y}, \mathcal{S})$, with Y_n dependent only on the single element X_n of the sequence X .

*Department of Statistics and Department of Mathematics, Iowa State University, Ames, IA 50011, USA; e-mail: apghosh@iastate.edu

†Department of Computer Science and Department of Mathematics, Iowa State University, Ames, IA 50011, USA; e-mail: ellerne@iastate.edu

‡Department of Mathematics, Iowa State University, Ames, IA 50011, USA; e-mail: roiterst@iastate.edu

Formally, let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and assume the existence of a probability space (Ω, \mathcal{F}, P) that carries both the Markov chain X as well as an i.i.d. sequence of (Ω, \mathcal{F}) -valued random variables ω_n , $n \in \mathbb{N}_0$ independent of X , such that

$$Y_n = h(X_n, \omega_n) \quad (1)$$

for some measurable function $h : \mathcal{D} \times \Omega \rightarrow \mathcal{Y}$. We assume that X is stationary under P . Let p_{ij} , $i, j \in \mathcal{D}$, be the transition matrix of the Markov chain X . Then the sequence of pairs $(X_n, Y_n)_{n \in \mathbb{N}_0}$ forms, under the law P , a stationary Markov chain with transition kernel defined for $n \in \mathbb{N}_0$, $i, j \in \mathcal{D}$, $y \in \mathcal{Y}$, $A \in \mathcal{S}$ by

$$K(i, y; j, A) := P(X_{n+1} = j, Y_{n+1} \in A | X_n = i, Y_n = y) = p_{ij} P((j, \omega_0) \in h^{-1}(A)) \quad (2)$$

Notice that $K(i, y; j, A)$ is in fact independent of the value of y .

The sequence $(X_n, Y_n)_{n \geq 0}$ constructed above is called a *Hidden Markov Model* (HMM). We refer to monographs [4, 6, 7, 9, 10, 13, 19] for a general account on HMM. The Hidden Markov Models have numerous applications in various areas such as communications engineering, bio-informatics, finance, applied statistics, and more. An extensive list of applications, for instance to coding theory, speech recognition, pattern recognition, satellite communications, and bio-informatics, can be also found in [2, 3, 16, 17].

One of the basic problems associated with HMM is to determine the most likely sequence of hidden states $(X_n)_{n=0}^M$ that could have generated a given output $(Y_n)_{n=0}^M$. In other words, given a vector $(y_n)_{n=0}^M \in \mathcal{Y}^M$, $M \in \mathbb{N}_0$, the problem is to find a feasible sequence of states $U_0^{(M)}, U_1^{(M)}, \dots, U_M^{(M)}$ such that

$$\begin{aligned} & P(X_0 = U_0^{(M)}, X_1 = U_1^{(M)}, \dots, X_M = U_M^{(M)} | Y_0 = y_0, Y_1 = y_1, \dots, Y_M = y_M) \\ &= \max_{(x_0, \dots, x_M) \in \mathcal{D}^M} P(X_0 = x_0, X_1 = x_1, \dots, X_M = x_M | Y_0 = y_0, Y_1 = y_1, \dots, Y_M = y_M), \end{aligned}$$

which is equivalent to

$$P(X_n = U_n^{(M)}, Y_n = y_n : 0 \leq n \leq M) = \max_{(x_0, \dots, x_M) \in \mathcal{D}^M} P(X_n = x_n, Y_n = y_n : 0 \leq n \leq M) \quad (3)$$

A vector $(U_n^{(M)})_{n=0}^M$ that satisfies (3) is called the *maximum a-posteriori* (MAP) path estimator. It can be efficiently calculated by the Viterbi algorithm [8, 19]. In general, the estimator is not uniquely defined even for a fixed outcome of observations $(y_n)_{n=0}^M$. Therefore, we will assume throughout that an additional selection rule (for example, according to the lexicographic order) is applied to produce $(U_n^{(M)})_{n=0}^M$. The MAP path estimator is used for instance in cryptanalysis, speech recognition, machine translation, and statistics, see [7, 8, 9, 19] and also references in [2, 3, 16, 17].

In principle, for a fixed $k \in \mathbb{N}$, the first k entries of $(U_i^{(M)})_{i=0}^M$ might vary significantly when M increases and approaches infinity. Unfortunately, it seems that very little is known about the asymptotic properties of the MAP path estimator for general HMM. However, recently it was shown in [2, 3] that under certain mild conditions on the transition kernel of an HMM, which were further amended in [16, 17], there exists a strictly increasing sequence of integer-valued non-negative random variables $(T_n)_{n \geq 1}$ such that the following properties

hold (In fact, we use here a slightly strengthened version of the corresponding results in [16, 17]. The proof that the statement holds in this form under the assumptions introduced in [16, 17] is given in Lemma 2.2 below).

Condition RT.

- (i) $(T_n)_{n \geq 1}$ is a (delayed) sequence of positive integers *renewal times*, that is the increments $\tau_n := T_{n+1} - T_n$, $n \geq 1$, form an i.i.d. sequence which is furthermore independent of T_1 .
- (ii) Both T_1 and τ_1 have finite moments of every order. In fact, there exist positive constants $a > 0$ and $b > 0$ such that

$$P(T_1 > t) \leq ae^{-bt} \quad \text{and} \quad P(\tau_1 > t) \leq ae^{-bt}, \quad \text{for all } t > 0. \quad (4)$$

- (iii) There exist a positive integer $r \geq 0$ such that for any fixed $i \geq 1$,

$$U_m^{(k)} = U_m^{(T_i+r)} \quad \text{for all } k \geq T_i + r \text{ and } m \leq T_i. \quad (5)$$

In particular, for any $n \geq 0$, the following limit exists and the equality holds with probability one:

$$U_n := \lim_{k \rightarrow \infty} U_n^{(k)} = U_n^{(T_{k_n}+r)}, \quad (6)$$

where

$$k_n = \min\{i \in \mathbb{N} : T_i + r \geq n\}, \quad (7)$$

that is k_n is the unique sequence of integers such that $T_{k_n-1} < n - r \leq T_{k_n}$ for all $n \geq 0$.

The interpretation of the above property is that at a random time $T_i + r$, blocks of estimators $(U_{T_{i-1}+1}^{(k)}, \dots, U_{T_i}^{(k)})$ (here and henceforth we make the convention that $T_0 = 0$) are fixed for all $k \geq T_i + r$ regardless of the future observations $(Y_j)_{j \geq T_i+r}$.

- (iv) For $n \geq 0$, let

$$Z_n := (X_n, U_n). \quad (8)$$

Then the sequence $Z := (Z_n)_{n \geq 0}$ forms a *regenerative process* with respect to the embedded renewal structure $(T_n)_{n \geq 0}$. That is, the random blocks

$$W_n = (Z_{T_n}, Z_{T_n+1}, \dots, Z_{T_{n+1}-1}), \quad n \geq 0, \quad (9)$$

are independent and, moreover, W_1, W_2, \dots (but possibly not W_0) are identically distributed.

Following [2, 3, 16, 17] we refer to the sequence $U = (U_n)_{n \geq 0}$ as the *infinite Viterbi path*. Condition RT yields a nice asymptotic behavior of the sequence U_n , see Proposition 1 in [2] for a summary and for instance [11] for a general account of regenerative processes. In particular, it implies the law of large numbers that we describe below.

Let $\xi : \mathcal{D}^2 \rightarrow \mathbb{R}$ be a real-valued function, and for all $n \geq 0$, $i = 0, 1, \dots, n$, define

$$\xi_i^{(n)} = \xi(X_i, U_i^{(n)}) \quad \text{and} \quad \xi_n = \xi(X_n, U_n), \quad (10)$$

and

$$\widehat{S}_n = \sum_{i=0}^n \xi_i^{(n)} \quad \text{and} \quad S_n = \sum_{i=0}^n \xi_i. \quad (11)$$

An important practical example, which we borrow from [2], is

$$\xi(x, y) = \mathbf{1}_{\{x \neq y\}} = \begin{cases} 1 & \text{if } x \neq y, \\ 0 & \text{if } x = y \end{cases}$$

In this case S_n and \widehat{S}_n count the number of places where the realization of the state X_i differs from the estimators U_i and $U_i^{(n)}$ respectively.

If Condition RT holds, then (see [2, 16]) there is a unique probability measure Q on the infinite product space $(\mathcal{D}^2)^{\mathbb{N}_0}$ where the sequence $(Z_n)_{n \geq 0}$ is defined, which makes $(W_n)_{n \geq 0}$ introduced in (9) into an i.i.d. sequence. Furthermore, under Condition RT, the following limits exist and the equalities hold:

$$\mu := \lim_{n \rightarrow \infty} \frac{S_n}{n} = \lim_{n \rightarrow \infty} \frac{\widehat{S}_n}{n} = \frac{E_Q(\widehat{S}_{T_2} - \widehat{S}_{T_1})}{E_Q(T_2 - T_1)}, \quad P - \text{a.s. and } Q - \text{a.s.}, \quad (12)$$

where E_Q denotes the expectation with respect to measure Q . In fact, Q is the conditional measure given by

$$Q(\cdot) = P(Z \in \cdot | T_1 = 0), \quad (13)$$

where it is assumed that $(Z_n)_{n \geq 0}$ is extended into a stationary sequence $(Z_n)_{n \in \mathbb{Z}}$.

The goal of this paper is to provide the following complementary large deviation estimates to the above law of large numbers.

Theorem 1.1. *Let Assumption 1.3 hold. Then either $Q(\widehat{S}_{T_2-1} = \mu T_2) = 1$ or there exist a constant $\gamma \in (0, \mu)$ and a function $I(x) : (\mu - \gamma, \mu + \gamma) \rightarrow [0, +\infty)$ such that*

- (i) $I(x)$ is lower semi-continuous and convex, $I(\mu) = 0$, and $I(x) > 0$ for $x \neq \mu$.
- (ii) $\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) = -I(x)$ for all $x \in (\mu, \mu + \gamma)$.
- (iii) $\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \leq nx) = -I(x)$ for all $x \in (\mu - \gamma, \mu)$.

Assumption 1.3 stated below is a set of specific conditions on HMM that ensures the existence of a sequence $(T_n)_{n \geq 1}$ satisfying Condition RT. The rate function $I(x)$ is specified in (18) below.

The proof of Theorem 1.1 is included in Section 2. First we show that Assumption 1.3 ensures that Condition RT is satisfied for an appropriate sequence of random times $(T_n)_{n \geq 1}$. This is done in Lemma 2.2 below. Then we show that as long as the non-degeneracy condition $Q(\widehat{S}_{T_2-1} = \mu T_2) \neq 1$ is satisfied, the existence of the regeneration structure described by Condition RT implies the large deviation result stated in Theorem 1.1. Related results for non-delayed regenerative processes (in which case W_0 defined in (9) would be distributed the same as the rest of the random blocks $W_n, n \geq 1$) can be found for instance in [1, 14, 18]. We cannot apply general large deviation results for regenerative processes directly to our setting

for the following two reasons. First, in our case the first block W_0 is in general distributed differently from the rest of the blocks W_n , $n \geq 1$, defined in (9). Secondly, the estimators $U_i^{(n)}$ for $i > T_{k_{n-1}} + r$ differ in general from the limiting random variables U_i . In principle, the contribution of these two factors to the tails asymptotic on the large deviation scale could be significant because both W_0 and $\sum_{i=T_{k_{n-1}}+r+1}^n \xi_i^{(n)}$ (where the last sum is assumed to be empty if $T_{k_{n-1}} + r + 1 > n$) have exponential tails. However, it turns out that this is not the case, and the large deviation result for our model holds with the same “classical” rate function $I(x)$ as it does for the non-delayed purely regenerative processes in [1, 14, 18]. In fact, only a small modification is required in order to adapt the proof of a large deviation result for regenerative processes in [1] to our model.

We next state our assumptions on the underlying HMM. The set of assumptions that we use is taken from [16, 17]. We assume throughout that the distribution of Y_n , conditioned on $X_n = i$, has a density with respect to a reference measure ν for all $i \in \mathcal{D}$. That is, there exist measurable functions $f_i : \mathcal{Y} \rightarrow \mathbb{R}_+$, $i \in \mathcal{D}$, such that

$$P(Y_0 \in A | X_0 = i) = \int_{\mathcal{Y}} f_i(y) \nu(dy), \quad A \in \mathcal{S}, i \in \mathcal{D}. \quad (14)$$

For each $i \in \mathcal{D}$, let $G_i = \{y \in \mathcal{Y} : f_i(y) > 0\}$. That is, the closure of G_i is the support of the conditional distribution $P(Y_0 \in \cdot | X_0 = i)$.

Definition 1.2. *We call a subset $C \subset \mathcal{D}$ a cluster if*

$$\min_{j \in C} P\left(Y_0 \in \bigcap_{i \in C} G_i | X_0 = j\right) > 0 \quad \text{and} \quad \max_{j \notin C} P_j\left(Y_0 \in \bigcap_{i \in C} G_i | X_0 = j\right) = 0.$$

In other words, a cluster is a maximal subset of states C such $\nu(G_C) > 0$, where $G_C := \bigcap_{i \in C} G_i$.

Assumption 1.3. [16, 17] *For $k \in \mathcal{D}$, let $H_k^* = \max_{j \in \mathcal{D}} p_{jk}$. Assume that*

$$P(f_k(Y_0)H_k^* > \max_{i \neq k} f_i(Y_0)H_i^* | X_0 = k) > 0, \quad \forall k \in \mathcal{D}. \quad (15)$$

Moreover, there exists a cluster $C \subset \mathcal{D}$ and $m \in \mathbb{N}$ such that the m -th power of the sub-stochastic matrix $H_C = (p_{ij})_{i,j \in C}$ is strictly positive.

Assumption 1.3 is taken from the conditions of Lemma 3.1 in [16] and [17], and it is slightly weaker than the one used in [2, 3]. This assumption is satisfied if $C = \mathcal{D}$ and

$$P(f_k(Y_0) > a \max_{i \neq k} f_i(Y_0) | X_0 = k) > 0, \quad \forall k \in \mathcal{D}, \forall a > 0. \quad (16)$$

The latter condition holds for instance if f_i is the density of a Gaussian random variable centered at i with variance $\sigma^2 > 0$.

For a further discussion of Assumption 1.3 and examples of HMM that satisfy this assumption see Section III in [17] and also the beginning of Section 2 below, where some results of [17] are summarized and their relevance to Condition RT is explained.

2 Proof of Theorem 1.1

Recall the random variables $\xi_i^{(n)}$ defined in (10). For the rest of the paper we will assume, actually without loss of generality, that $\xi_i^{(n)} > 0$ for all $n \geq 0$ and integers $i \in [0, n]$. Indeed, since \mathcal{D} is a finite set, the range of the function ξ is bounded, and hence if needed we can replace ξ by $\bar{\xi} = M + \xi$ with some $M > 0$ large enough such that $\xi_i^{(n)} + M > 0$ for all $n \geq 0$, and $i \in [0, n]$.

We start with the definition of the sequence of random times $(T_n)_{n \geq 1}$ that satisfies Condition RT. A similar sequence was first introduced in [2] in a slightly less general setting. In this paper, we use a version of the sequence which is defined in Section 4 of [16]. The authors of [16] do not explicitly show that Condition RT is fulfilled for this sequence. However, only a small modification of technical lemmas from [16] or [17] is required to demonstrate this result.

Roughly speaking, for some integers $M \in \mathbb{N}$ and $r \in [0, M - 1]$, the random times $\theta_n := T_n + r - M + 1$ are defined, with the additional requirement $\theta_{n+1} - \theta_n > M$, as successive hitting time of a set in the form $\mathbf{q} \times \mathcal{A}$, $\mathbf{q} \in \mathcal{D}^M$, $\mathcal{A} \subset \mathcal{Y}^M$, for the Markov chain formed by the vectors of pairs $R_n = (X_i, Y_i)_{i=n}^{n+M-1}$.

More precisely, the following is shown in [16, 17]. See Lemmas 3.1 and 3.2 in either [16] or [17] for (i) and (ii), and Section 4 [p. 202] of [16] for (iii). We notice that a similar regeneration structure was first introduced in [2] under a more stringent condition than Assumption 1.3.

Lemma 2.1. [16, 17] *Let Assumption 1.3 hold. Then there exist integer constants $M \in \mathbb{N}$, $r \in [0, M - 1]$, a state $l \in \mathcal{D}$, a vector of states $\mathbf{q} \in \mathcal{D}^M$, and a measurable set $\mathcal{A} \subset \mathcal{Y}^M$ such that the following hold:*

(i) $P((X_n)_{n=0}^{M-1} = \mathbf{q}, (Y_n)_{n=0}^{M-1} \in \mathcal{A}) > 0,$

(ii) *If for some $k \geq 0$, $(Y_i)_{i=k}^{k+M-1} \in \mathcal{A}$, then $U_j^{(m)} = U_j^{(k+M-1)} = l$ for any $j \leq k + M - 1 - r$ and $m \geq k + M - 1$.*

Furthermore,

(iii) *Let $\theta_1 = \inf\{k \geq 0 : (X_i)_{i=k}^{k+M-1} = \mathbf{q} \text{ and } (Y_i)_{i=k}^{k+M-1} \in \mathcal{A}\}$, and for $n \geq 1$, $\theta_{n+1} = \inf\{k \geq \theta_n + M : (X_i)_{i=k}^{k+M-1} = \mathbf{q} \text{ and } (Y_i)_{i=k}^{k+M-1} \in \mathcal{A}\}$. Define*

$$T_n = \theta_n + M - 1 - r. \tag{17}$$

For $n \geq 0$, let $L_n := (Y_n, U_n)$. Then the sequence $(L_n)_{n \geq 0}$ forms a regenerative process with respect to the embedded renewal structure $(T_n)_{n \geq 0}$. That is, the random blocks

$$J_n = (L_{T_n}, L_{T_n+1}, \dots, L_{T_{n+1}-1}), \quad n \geq 0,$$

are independent and, moreover, J_1, J_2, \dots (but possibly not J_0) are identically distributed. Furthermore, there is a deterministic function $g : \bigcup_{n \in \mathbb{N}} \mathcal{Y}^n \rightarrow \bigcup_{n \in \mathbb{N}} \mathcal{D}^n$ such that $(U_{T_n}, U_{T_n+1}, \dots, U_{T_{n+1}-1}) = g(Y_{T_n}, Y_{T_n+1}, \dots, Y_{T_{n+1}-1})$.

In fact, using Assumption 1.3, the set \mathcal{A} is designed in [16, 17] in such a way that once a M -tuple of observable variables $(Y_i)_{i=n}^{n+M-1}$ that belongs to \mathcal{A} occurs, the value of the estimator $U_{n+M-1-r}^{(m)}$ is set to l when $m = n + M - 1$, and, moreover, it remains unchanged for all $m \geq n + M - 1$ regardless of the values of the future observations $(Y_i)_{i \geq n+M}$. The elements of the set \mathcal{A} are called barriers in [16, 17]. Using basic properties of the Viterbi algorithm, it is not hard to check that the existence of the barriers implies claims (ii) and (iii) of the above lemma. In fact, the Viterbi algorithm is a dynamic programming procedure which is based on the fact that for any integer $k \geq 2$ and states $i, j \in \mathcal{D}$ there exists a deterministic function $g_{k,i,j} : \mathcal{Y}^k \rightarrow \mathcal{D}^k$ such that $(U_n^{(m)}, U_{n+1}^{(m)}, \dots, U_{n+k-1}^{(m)}) = g_{k,i,j}(Y_n, Y_{n+1}, \dots, Y_{n+k-1})$ for all $n \geq 0$ and $m \geq n + k - 1$ once it is decided that $U_n^{(m)} = i$ and $U_{n+k}^{(m)} = j$. See [16, 17] for details.

We have:

Lemma 2.2. *Let Assumption 1.3 hold. Then Condition RT is satisfied for the random times T_n defined in (17).*

Proof. We have to show that properties (i)–(iv) listed in the statement of Condition RT hold for $(T_n)_{n \in \mathbb{N}}$.

(i)-(ii) Notice that $(\theta_n)_{n \geq 1}$ defined in (iii) of Lemma 2.1 is (besides the additional requirement that $\theta_{n+1} - \theta_n > M$) essentially the sequence of successive hitting times of the set $\mathbf{q} \times \mathcal{A}$ for the Markov chain formed by the pairs of M -vectors $(X_i, Y_i)_{i=n}^{n+M-1}$, $n \geq 0$. Therefore, (i) and (ii) follow from Lemma 2.1-(i) and the fact that the M -vectors $(X_i)_{i=n}^{n+M-1}$, $n \geq 0$, form an irreducible Markov chain on the finite space $\mathcal{D}_M = \{\mathbf{x} \in \mathcal{D}^M : P((X_i)_{i=0}^{M-1} = \mathbf{x}) > 0\}$ while the distribution of the observation Y_i depends on the value of X_i only.

(iii) Follows from (ii) of Lemma 2.1 directly.

(iv) The definition of the HMM together with (17) imply that $(X_n, Y_n)_{n \geq 0}$ is a regenerative process with respect to the renewal sequence $(T_n)_{n \geq 1}$. The desired result follows from this property combined with the fact that $(U_{T_n}, U_{T_n+1}, \dots, U_{T_{n+1}-1})$ is a deterministic function of $(Y_{T_n}, Y_{T_n+1}, \dots, Y_{T_{n+1}-1})$ according to part (iii) of Lemma 2.1.

The proof of the lemma is completed. \square

We next define the rate function $I(x)$ that appears in the statement of Theorem 1.1. The rate function is standard in the large deviation theory of regenerative processes (see for instance [1, 14, 18]). Recall from (13) the conditional measure Q which makes $W = (W_n)_{n \geq 0}$ defined in (9) into a stationary sequence of random blocks. For $n \geq 1$ let $\tau_n = T_n - T_{n-1}$ and let (\tilde{S}, τ) be a random pair distributed under the measure Q identically to any of $(\sum_{k=T_{n-1}}^{T_n-1} \xi_k, \tau_n)$, $n \geq 1$. For any constants $\alpha \in \mathbb{R}$, $\Lambda \in \mathbb{R}$, and $x \geq 0$, set

$$\Gamma(\alpha, \Lambda) = \log E_Q(\exp(\alpha \tilde{S} - \Lambda \tau)),$$

and define

$$\Lambda(\alpha) = \inf\{\Lambda \in \mathbb{R} : \Gamma(\alpha, \Lambda) \leq 0\} \quad \text{and} \quad I(x) = \sup_{\alpha \in \mathbb{R}} \{\alpha x - \Lambda(\alpha)\}, \quad (18)$$

using the usual convention that the infimum over an empty set is $+\infty$.

We summarize some properties of the above defined quantities in the following lemma. Recall the definition of μ from (12).

Lemma 2.3. *The following hold:*

(i) $\Lambda(\alpha)$ and $I(x)$ are both lower semi-continuous convex functions on \mathbb{R} . Moreover, $I(\mu) = 0$ and $I(x) > 0$ for $x \neq \mu$.

(ii) If there is no $c > 0$ such that $Q(\widehat{S}_{T_2-1} = cT_2) = 1$, then

(a) $\Lambda(\alpha)$ is strictly convex in an neighborhood of 0.

(b) $I(x)$ is finite in some neighborhood of μ , and for each x in that neighborhood we have $I(x) = (\alpha_x x - \Lambda(\alpha_x))$ for some α_x such that $(x - \mu)\alpha_x > 0$ for $x \neq \mu$ and $\lim_{x \rightarrow \mu} \alpha_x = 0$.

Proof.

(i) The fact that $\Lambda(\alpha)$ and $I(x)$ are both lower semi-continuous convex functions is well-known, see for instance p. 2884 in [1]. Furthermore, (4) implies that $\Gamma(\alpha, \Lambda(\alpha)) = 0$ for α in a neighborhood of 0. Moreover, the dominated convergence theorem implies that Γ has continuous partial derivatives in a neighborhood of $(0, 0)$. Therefore, by the implicit function theorem, $\Lambda(\alpha)$ is differentiable in this neighborhood. In particular, $\Lambda(0) = 0$, and

$$\Lambda'(0) = -\frac{\frac{\partial \Gamma}{\partial \alpha}(0, 0)}{\frac{\partial \Gamma}{\partial \Lambda}(0, 0)} = \mu.$$

Since Λ is a convex function, this yields the second part of (i), namely establishes that $I(\mu) = 0$ and $I(x) > 0$ for $x \neq \mu$.

(ii) If there is no constants $c, b \in \mathbb{R}$ such that $Q(\widehat{S}_{T_2-1} - cT_2 = b) = 1$, we can use the results of [1] for both (ii-a) and (ii-b), see Lemma 3.1 and the preceding paragraph on p. 2884 of [1].

It remains to consider the case when $Q(\widehat{S}_{T_2-1} - cT_2 = b) = 1$ with $b \neq 0$ but there are no $\tilde{c} > 0$ such that $Q(\widehat{S}_{T_2-1} - \tilde{c}T_2 = 0) = 1$. In this case, using the definition (18) and the estimate (4), we have for α in a neighborhood of 0,

$$\begin{aligned} 1 &= E_Q[\exp(\alpha S_{T_2-1} - \Lambda(\alpha)T_2)] = E_Q[\exp(\alpha(b + cT_2) - \Lambda(\alpha)T_2)] \\ &= e^{\alpha b} E_Q[\exp(T_2(\alpha c - \Lambda(\alpha)))]. \end{aligned}$$

That is, for α in a neighborhood of 0,

$$E_Q[\exp(T_2(\alpha c - \Lambda(\alpha)))] = e^{-\alpha b}. \quad (19)$$

Suppose $\Lambda(\alpha)$ is not strictly convex in any neighborhood of 0, that is $\Lambda(\alpha) = k\alpha$ for some $k \in \mathbb{R}$ and all α in a one-sided (say positive) neighborhood of zero. Let $\alpha_1 > 0$ and $\alpha_2 > \alpha_1$

be two values of α in the aforementioned neighborhood of zero for which (19) is true. Using Jensen's inequality $E_Q(X^{\frac{\alpha_2}{\alpha_1}}) \geq (E_Q(X))^{\frac{\alpha_2}{\alpha_1}}$, with

$$X = \exp(T_2(\alpha_1 c - \Lambda(\alpha_1))) = \exp(T_2 \alpha_1 (c - k)),$$

one gets that the identity (19) can hold only if $Q(T_2 = c_0) = 1$ for some $c_0 > 0$. But under this condition and the assumption $Q(\widehat{S}_{T_2-1} - cT_2 = b) = 1$, we would have

$$Q(\widehat{S}_{T_2-1} - T_2(c c_0 + b)/c_0 = 0) = 1$$

in contradiction to what we have assumed in the beginning of the paragraph. This completes the proof of part (a), namely shows that $\Lambda(\alpha)$ is strictly convex in a neighborhood of 0 provided that there is no $c > 0$ such that $Q(\widehat{S}_{T_2-1} - cT_2 = 0) = 1$.

To complete the proof of the lemma, it remains to show that part (b) of (ii) holds. Although the argument is standard, we give it here for the sake of completeness. We remark that in contrast to the "usual assumptions", we consider properties of $\Lambda(\alpha)$ and $I(x)$ only in some neighborhoods of 0 and μ respectively, and not in the whole domains where these functions are finite. Recall that $\Lambda(\alpha)$ is an analytic and strictly convex function of α in a neighborhood of zero (see for instance [1] or [18]). In particular, $\Lambda'(\alpha)$ exists and is strictly increasing in this neighborhood. Since $\Lambda'(0) = \mu$, it follows that for all x in a neighborhood of μ , there is a unique α_x such that $x = \Lambda'(\alpha_x)$. Since $\Lambda'(\alpha)$ is a continuous increasing function, we have $\lim_{x \rightarrow \mu} \alpha_x = 0$ and $\alpha_x(x - \mu) > 0$. Furthermore, since $\Lambda(\alpha)$ is convex, we obtain $\Lambda(\alpha) - \Lambda(\alpha_x) \geq x(\alpha - \alpha_x)$, and hence $x\alpha_x - \Lambda(\alpha_x) \geq x\alpha - \Lambda(\alpha)$ for all $\alpha \in \mathbb{R}$. This yields $I(x) = x\alpha_x - \Lambda(\alpha_x)$, completing the proof of the lemma. \square

Remark 2.4. Part (ii)-(b) of the above lemma shows that the rate function $I(x)$ is finite in a neighborhood of μ as long as there is no $c > 0$ such that $Q(\widehat{S}_{T_2-1} - cT_2 = 0) = 1$. On the other hand, if $Q(\widehat{S}_{T_2-1} = \mu T_2) = 1$, then the definition of $\Lambda(\alpha)$ implies $\Lambda(\alpha) = \mu\alpha$ for all $\alpha \in \mathbb{R}$, and hence $I(x) = +\infty$ for $x \neq \mu$.

Part of the claim (i) of Theorem 1.1 is in the part (i) of the above lemma. Therefore we now turn to the proof of parts (ii) and (iii) of Theorem 1.1. We start from the observation that the large deviation asymptotic for the lower tail $P(\widehat{S}_n \leq nx)$ with $x < \mu$, can be deduced from the corresponding results for the upper tail $P(\widehat{S}_n \geq nx)$, $x > \mu$. Indeed, let $\bar{\xi}(i, j) = \mu - \xi(i, j)$ and $\bar{S}_n = \sum_i^n \bar{\xi}(X_i, U_i^{(n)})$. Then

$$P(\widehat{S}_n \leq nx) = P(\bar{S}_n \geq n(\mu - x)). \quad (20)$$

Furthermore, for the function $\bar{\xi}$ we have

$$\bar{\Gamma}(\alpha, \Lambda) = \log E(\exp(\alpha(\mu\tau - W) - \Lambda\tau)) = \log E(\exp(-\alpha W - (\Lambda - \alpha\mu)\tau)),$$

and hence $\bar{\Lambda}(\alpha) := \inf\{\Lambda \in \mathbb{R} : \bar{\Gamma}(\alpha, \Lambda) \leq 0\} = \Lambda(-\alpha) + \mu\alpha$, which in turn implies

$$\bar{I}(x) := \sup\{\alpha \in \mathbb{R} : \alpha x - \bar{\Lambda}(\alpha)\} = \sup\{\alpha \in \mathbb{R} : -\alpha(\mu - x) - \Lambda(-\alpha)\} = I(\mu - x).$$

Therefore, part (iii) of Theorem 1.1 can be derived from part (ii) applied to the auxiliary function $\bar{\xi}$.

It remains to prove part (ii) of the theorem. For the upper bound we adapt the proof of Lemma 3.2 in [1].

Proposition 2.5. *Let Assumption 1.3 hold and suppose in addition that $\xi(i, j) > 0$ for all $i, j \in \mathcal{D}$ and $Q(\widehat{S}_{T_2-1} = \mu T_2) \neq 1$. Let $I(x)$ be as defined in (18). Then there exists a constant $\gamma > 0$ such that $\limsup_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \leq -I(x)$ for all $x \in (\mu, \mu + \gamma)$.*

Proof. Recall k_n from (7). For $n \in \mathbb{N}$ and integers $i \in [0, k_n)$, let

$$\widetilde{S}_i = \sum_{j=0}^{T_i-1} \xi(X_j, U_j^{(n)}),$$

with the convention that $\widetilde{S}_1 = 0$ if $T_1 = 0$. Further, recall the random sequence k_n from (7), and set $R_n = \widehat{S}_n - \widetilde{S}_{k_n-1} = \sum_{i=T_{k_n-1}}^n \xi(X_i, U_i^{(n)})$.

We use the following series of estimates, where $\alpha > 0$ and $\Lambda > 0$ are arbitrary positive parameters small enough to ensure that all the expectations below exist (due to the tail estimates assumed in (4)):

$$\begin{aligned} P(\widehat{S}_n \geq nx) &= P(\widetilde{S}_{k_n-1} + R_n \geq nx) = \sum_{j=0}^{n-1} P(\widetilde{S}_j + R_n \geq nx; T_j < n - r \leq T_{j+1}) \\ &\leq \sum_{j=0}^{n-1} P(\widetilde{S}_j + R_n \geq nx; T_j < n) \leq \sum_{j=0}^{n-1} P(\alpha(\widetilde{S}_j + R_n - nx) + \Lambda(n - T_j) \geq 0) \\ &\leq \sum_{j=0}^{n-1} E[\exp(\alpha(\widetilde{S}_j - xn) - \Lambda(T_j - n) + \alpha R_n)] \\ &= E[\exp(\alpha(\widetilde{S}_1 + R_n))] \cdot \exp(-(\alpha x - \Lambda)n) \cdot \sum_{j=0}^{n-1} \exp((j-1)\Gamma(\alpha, \Lambda)) \\ &\leq E[\exp(\alpha(\widetilde{S}_1 + R_n))] \cdot \exp(-(\alpha x - \Lambda)n) \cdot \frac{1 - e^{\Gamma(\alpha, \Lambda)n}}{1 - e^{\Gamma(\alpha, \Lambda)}}. \end{aligned}$$

By the Cauchy-Schwartz inequality,

$$E[\exp(\alpha(\widetilde{S}_1 + R_n))] \leq \sqrt{E[\exp(2\alpha\widetilde{S}_1)] \cdot E[\exp(2\alpha R_n)]}.$$

Therefore, using any $M > 0$ such that $\xi(i, j) < M$ for all $i, j \in \mathcal{D}$,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{1}{n} \log E[\exp(\alpha(\widetilde{S}_1 + R_n))] &= \limsup_{n \rightarrow \infty} \frac{1}{2n} \log E[\exp(2\alpha R_n)] \\ &\leq \lim_{n \rightarrow \infty} \frac{1}{n} \log E[\exp(2\alpha M(n - T_{k_n-1}))] = 0, \end{aligned}$$

where the first equality is due to (4), while in the last step we used the renewal theorem which shows that the law of $n - T_{k_n-1}$ converges, as n goes to infinity, to a (proper) limiting distribution. Therefore, if $\alpha > 0$ and $\Lambda > 0$ are small enough and $\Gamma(\alpha, \Lambda) \leq 0$, then

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \leq -(\alpha x - \Lambda).$$

It follows then from the definition of $\Lambda(\alpha)$ given in (18) that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \leq -(\alpha x - \Lambda(\alpha)). \quad (21)$$

The rest of the proof is standard. Let

$$A := \sup\{\alpha > 0 : E(e^{\alpha \widetilde{S}_1}) < \infty, E(e^{\alpha S_2}) < \infty\}.$$

It follows from (21) that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \leq - \sup_{0 < \alpha < A} (\alpha x - \Lambda(\alpha)),$$

and therefore $\limsup_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \leq -I(x)$ provided that $\alpha_x \in (0, A)$, where α_x is defined in the statement of Lemma 2.3. Since by part (ii-b) of Lemma 2.3, we have $\lim_{x \downarrow \mu} \alpha_x = 0$, this completes the proof of Proposition 2.5. \square

For the lower bound in part (ii) of Theorem 1.1 we have

Proposition 2.6. *Let Assumption 1.3 hold and suppose in addition that $\xi(i, j) > 0$ for all $i, j \in \mathcal{D}$. Let $I(x)$ be as defined in (18). Then $\liminf_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq nx) \geq -I(x)$ for all $x > \mu$.*

Using the assumption that $\xi(i, j) > 0$ the proof of this proposition can be done by relying on the arguments used in the proof of Lemma 3.4 in [1] nearly verbatim. For the sake of completeness we sketch the proof here. First, notice that the proposition is trivially true in the degenerate case $Q(\widehat{S}_{T_2-1} = \mu T_2) = 1$ where $I(x) = +\infty$ (see Remark 2.4). Assume now that $Q(\widehat{S}_{T_2-1} = \mu T_2) \neq 1$. Then, as in Lemma 3.1 of [1], we have:

$$\inf_{0 < \gamma < 1} \gamma \Gamma^*(x/\gamma, 1/\gamma) = I(x), \quad (22)$$

where $\Gamma^*(u, v) := \sup_{\alpha, \Lambda \in \mathbb{R}} \{\alpha u - \Lambda v - \Gamma(\alpha, \Lambda)\}$. Notice that since the infimum in (22) is taken over strictly positive values of γ , the proof of this identity given in [1] works verbatim in our case. This is because considering only strictly positive values of γ makes the first renewal block, which is special and difficult to control in our case, irrelevant to the proof of the identity (22).

Since $\xi_i^{(n)}$ are assumed to be positive numbers, we have for any $\varepsilon > 0$, $\gamma \in (0, 1)$, and $x > \mu$:

$$\begin{aligned} P(\widehat{S}_n \geq xn) &\geq P(\widehat{S}_{T_{[\gamma n]}} - \widehat{S}_1 \geq xn; T_{[\gamma n]} < n) \\ &\geq P(\widehat{S}_{T_{[\gamma n]}} - \widehat{S}_1 \geq \frac{x + \varepsilon}{\gamma} [\gamma n]; T_{[\gamma n]} < \frac{1}{\gamma} [\gamma n]), \end{aligned} \quad (23)$$

where, as usual, $[x]$ denotes the integer part of $x \in \mathbb{R}$, that is $[x] = \max\{z \in \mathbb{Z} : z \leq x\}$.

Applying Cramer's large deviation theorem [5] to a sequence of i.i.d 2-dimensional vectors $(\widehat{S}_{T_{[\gamma n]}} - \widehat{S}_1 \geq \frac{x + \varepsilon}{\gamma} [\gamma n]; T_{[\gamma n]})_{n \in \mathbb{N}}$ we obtain

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_{T_{[\gamma n]}} - \widehat{S}_1 \geq \frac{x + \varepsilon}{\gamma} [\gamma n]; T_{[\gamma n]} < \frac{1}{\gamma} [\gamma n]) = \Gamma^*((x + \varepsilon)/\gamma, 1/\gamma)$$

Letting ε go to zero and using (23), we obtain

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P(\widehat{S}_n \geq xn) \geq - \inf_{0 < \gamma < 1} \gamma \Gamma^*(x/\gamma, 1/\gamma),$$

which completes the proof of Proposition 2.6 in virtue of (22).

Propositions 2.5 and 2.6 combined together yield the claim of part (ii) of Theorem 1.1, completing the proof of the theorem.

References

- [1] J. Bucklew, *The blind simulation problem and regenerative processes*, IEEE Trans. Inform. Theory **44** (1998), 2877–2891.
- [2] A. Caliebe, *Properties of the maximum a posteriori path estimator in hidden Markov models*, IEEE Trans. Inform. Theory **52** (2006), 41–51.
- [3] A. Caliebe and U. Rösler, *Convergence of the maximum a posteriori path estimator in hidden Markov models*, IEEE Trans. Inform. Theory **48** (2002), 1750–1758.
- [4] O. Cappé, E. Moulines, T. Rydén, *Inference in Hidden Markov Models*, Springer, 2005.
- [5] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications, 2nd edition*, Applications of Mathematics **38**, Springer-Verlag, New York, 1998.
- [6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge Univ. Press, 1998.
- [7] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Trans. Inform. Theory **48** (2002), 1518–1569.
- [8] G. D. Forney, *The Viterbi algorithm*, Proceedings of the IEEE **61** (1973), 268–278.
- [9] X. Huang, Y. Ariki, and M. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh Univ. Press, 1990.
- [10] F. Jelinek, *Statistical Methods for Speech Recognition*, MIT Press, 2001.
- [11] V. Kalashnikov, *Topics on Regenerative Processes*, CRC Press, Boca Raton, FL, 1994.
- [12] A. Koloydenko, M. Käärik, and J. Lember, *On adjusted Viterbi training*, Acta Appl. Math. **96** (2007), 309–326.
- [13] A. Krogh, *Computational Methods in Molecular Biology*, Amsterdam: North-Holland, 1998.
- [14] T. Kuczek and K. N. Crank, *A large-deviation result for regenerative processes*, J. Theor. Probab. **4** (1991), 551–561.

- [15] J. Lember and A. Koloydenko, *Adjusted Viterbi training*, Probab. Engrg. Inform. Sci **21** (2007), 451–475.
- [16] J. Lember and A. Koloydenko, *The adjusted Viterbi training for hidden Markov models*, Bernoulli **14** (2008), 180–206.
- [17] J. Lember and A. Koloydenko, *A constructive proof of the existence of Viterbi processes*, preprint, <http://www.arxiv.org/pdf/0804.2138>.
- [18] P. Ney and E. Nummelin, *Markov additive processes II. Large deviations*, Ann. Probab. **15** (1987), 593–609.
- [19] L. R. Rabiner, *A tutorial on Hidden Markov Models and selected applications in speech recognition*, Proceedings of the IEEE **77** (1989), 257-286.