Bernoulli **17**(2), 2011, 609–627
DOI: 10.3150/10-BEJ294

On the Viterbi process with continuous state space

PAVEL CHIGANSKY and YAACOV RITOV

Department of Statistics, The Hebrew University, Mount Scopus, Jerusalem 91905, Israel.
E-mail: pchiga@mscc.huji.ac.il; yaacov.ritov@gmail.com

This paper deals with convergence of the maximum a posterior probability path estimator in hidden Markov models. We show that when the state space of the hidden process is continuous, the optimal path may stabilize in a way which is essentially different from the previously considered finite-state setting.

Keywords: hidden Markov models; MAP path estimator; Viterbi algorithm

1. Introduction

Consider a standard hidden Markov model (X, Y), where $X = (X_n)_{n\in\mathbb{Z}_+}$ and $Y = (Y_n)_{n\in\mathbb{Z}_+}$ are the *hidden* state and the *observation* processes, respectively. The state process X is Markov with values in a subset $\mathcal{S} \subseteq \mathbb{R}$, transition probability Q and initial distribution $\mathcal{M}$: for all measurable subsets $A \subseteq \mathcal{S}$,

$$\begin{aligned} \mathbb{P}(X_1 \in A) &= \mathcal{M}(A), \\ \mathbb{P}(X_n \in A|X_{n-1}) &= Q(X_{n-1}, A), \qquad \mathbb{P}\text{-a.s.}, n > 1. \end{aligned}$$

We shall consider either countable $\mathcal{S}$, in which case $q(u, v) := Q(u, \{v\})$ and $\mu(u) := \mathcal{M}(\{u\})$, or $\mathcal{S} = \mathbb{R}$, assuming that $Q(u, \mathrm{d}v)$ and $\mathcal{M}(\mathrm{d}u)$ have densities $q(u, v)$ and $\mu(u)$ with respect to the Lebesgue measure. The precise meaning of $q(u, v)$ and $\mu(u)$ should be obvious from the context.

The observed process Y forms a sequence of conditionally independent random variables, given $X_{1:\infty} = (X_1, X_2, \ldots)$, with the *observation* density p:

$$\mathbb{P}(Y_n \in B|X_{1:\infty}) = \int_B p(X_n, y)\,\mathrm{d}y, \qquad \mathbb{P}\text{-a.s.},$$

for any Borel $B \subseteq \mathbb{R}$.

The path estimation problem is to reconstruct the trajectory of the hidden process[1] $X_{1:n} = (X_1, \ldots, X_n)$, given the realization of $Y_{1:n} = (Y_1, \ldots, Y_n)$ for a fixed horizon $n \ge 1$. If $\mathcal{S}$ is a

[1]Hereafter, for $x \in \mathbb{R}^n$, x_m stands for the mth entry of x and $x_{k:m}$, $k \le m$, denotes the vector $x = (x_k, \ldots, x_m)$; $|x_{1:n}| = \max_i |x_i|$ and $\|x_{1:n}\| = \sqrt{\sum_{i=1}^n x_i^2}$.

1350-7265

discrete set, a natural estimator is the maximizer of the a posterior probability (MAP estimator):

$$\hat{X}^n_{1:n} := \operatorname*{argmax}_{x_{1:n}\in\mathcal{S}^n} \mathbb{P}(X_{1:n} = x_{1:n}|Y_{1:n}),$$

where the optimal path is chosen according to the lexicographical order on $\mathcal{S}^n$, induced by an order on $\mathcal{S}$, whenever the maximum is not unique. The obtained path minimizes the probability of error among all estimators depending on $Y_{1:n}$, that is,

$$\mathbb{P}(\hat{X}^n_{1:n} \neq X_{1:n}) \le \mathbb{P}(\xi_{1:n} \neq X_{1:n}) \qquad \text{for all } \sigma\{Y_1,\dots,Y_n\}\text{-measurable } \xi_{1:n}.$$

By Bayes' formula,

$$\mathbb{P}(X_{1:n} = x_{1:n}|Y_{1:n}) = \frac{L_n(x_{1:n}, Y_{1:n})}{\sum_{u_{1:n}\in\mathcal{S}^n} L_n(u_{1:n}, Y_{1:n})},$$

where L_n is the "posterior" likelihood:

$$L_n(x_{1:n}; y_{1:n}) = \mu(x_1)p(x_1, y_1)\prod_{m=2}^{n} q(x_{m-1}, x_m)p(x_m, y_m), \qquad x_{1:n}\in\mathcal{S}^n, \tag{1.1}$$

and hence

$$\hat{X}^n_{1:n} = \operatorname*{argmax}_{x_{1:n}\in\mathcal{S}^n} L_n(x_{1:n}, Y_{1:n}).$$

Due to the product structure of L_n, the search for the maximizing path can be carried out efficiently by a dynamic programming procedure, called the *Viterbi algorithm*, after A. Viterbi, who introduced it in the context of error correction codes.

When the next observation, Y_{n+1}, is added, the optimal path may change entirely, that is, for any $m = 1,\dots,n$, $\hat{X}^{n+1}_{1:m}$ is, in general, different from $\hat{X}^n_{1:m}$. In practical terms, the latter means that[2] $\#\mathcal{S}$ optimal path candidates of length n are to be kept in memory at each time n. This motivates the question of whether the optimal path stabilizes as the number of observations grows to infinity or, more precisely, whether the limit

$$\hat{X}_{1:m} = \lim_{n\to\infty} \hat{X}^n_{1:m} \tag{1.2}$$

exists $\mathbb{P}$-a.s. for each fixed $m \ge 1$. If such a limit exists, it defines a random process with paths in $\mathcal{S}^\infty$, named (in [13]) *the Viterbi process*.

An affirmative answer to this question was given in [5] (see also [10]) under a sufficient condition (see (2.1) below) which also ensures that the limit sequence $\hat{X} = (\hat{X}_m)_{m\ge1}$ is a regenerative process. More precisely, a sequence of stopping times can be constructed (see [4]), splitting the process $\hat{X}$ into cycles that are i.i.d. and independent of the initial delay. In particular, by the regenerative property, $\hat{X}$ satisfies the classical limit laws, such as the law of large numbers (LLN) and the central limit theorem (CLT).

[2] $\#A$ stands for the cardinality of a set A.

In fact, the existence of such renewal times under the condition (2.1) can be deduced by a simple argument (reproduced, for completeness, in Section 2). A more delicate construction in [12,13] verifies (1.2) under conditions weaker than (2.1).

In this paper, we revisit the question of the existence of the limit (1.2) for hidden Markov models (HMMs) with continuous state spaces, that is, when $\mathcal{S} = \mathbb{R}$ and for each $u \in \mathbb{R}$, the transition kernel $Q(u, \mathrm{d}v)$ and the initial distribution $\mathcal{M}(\mathrm{d}v)$ have densities $q(u, v)$ and $\mu(v)$, respectively, with respect to the Lebesgue measure. By Bayes' formula, the conditional law of the vector $X_{1:n}$ given $Y_{1:n}$ has the density ψ_n with respect to the Lebesgue measure on $\mathbb{R}^n$:

$$\psi_n(x_{1:n}) := \frac{L_n(x_{1:n}; Y_{1:n})}{\int_{\mathbb{R}^n} L_n(u_{1:n}; Y_{1:n})\,\mathrm{d}u_1 \cdots \mathrm{d}u_n},$$

with L_n defined as in (1.1). The MAP path estimator is

$$\hat{X}^n_{1:n} := \operatorname*{argmax}_{x_{1:n}\in\mathbb{R}^n} \psi_n(x_{1:n}) = \operatorname*{argmax}_{x_{1:n}\in\mathbb{R}^n} L_n(x_{1:n}; Y_{1:n}),$$

where, as in (1.2), the maximum is chosen according to the lexicographical order on $\mathbb{R}^n$ (induced, e.g., by $<$ on $\mathbb{R}$) in case of ambiguity.

Note that for any $\sigma\{Y_1, \dots, Y_n\}$-measurable random vector $\xi_{1:n}$ and $\varepsilon > 0$,

$$\mathbb{P}(|X_{1:n} - \xi_{1:n}| \le \varepsilon) = \mathbb{E}\mathbb{P}(|X_{1:n} - \xi_{1:n}| \le \varepsilon | Y_{1:n}) = \mathbb{E} \int_{[-\varepsilon,\varepsilon]^n} \psi_n(x_{1:n} + \xi_{1:n})\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

and hence the estimator $\hat{X}^n_{1:n}$ is optimal in the sense that

$$\lim_{\varepsilon\to 0} \varepsilon^{-n}\mathbb{P}(|X_{1:n} - \xi_{1:n}| \le \varepsilon) = \mathbb{E}\psi_n(\xi_{1:n}) \le \mathbb{E} \max_{x_{1:n}\in\mathbb{R}^n} \psi_n(x_{1:n}) = \lim_{\varepsilon\to 0} \varepsilon^{-n}\mathbb{P}(|X_{1:n} - \hat{X}^n_{1:n}| \le \varepsilon)$$

whenever interchanging the expectation and the limit is possible. Roughly, this means that $\hat{X}^n_{1:n}$ yields the best "small" credible intervals among all other path estimates.[3]

As in state estimation problems such as filtering, the exact calculation of $\hat{X}^n_{1:n}$ is impossible beyond a number of models with a special structure, most notably Kalman's linear Gaussian setting. A number of efficient numerical techniques, such as particle filters, have been developed (see, e.g., [6]) to approximate the conditional law of the hidden state process. In this paper, we are concerned with the convergence properties of the MAP paths, leaving the computational issues for further investigation.

In Section 2, we explore, through a number of examples, various patterns of convergence encountered in (1.2), when the hidden state space is continuous. We also give an example of HMM, for which the MAP path does not converge as the estimation time horizon increases. In Section 3, we prove a more general result, deducing the existence of the limit (1.2) from certain strong log-concavity of the transition and observation densities. The Appendix contains a lemma which is used in the proof of the main result and may be of independent interest. Finally, a short discussion of the results appears in Section 4.

[3] In fact, this optimality interpretation turns out to be meaningful even in the infinite-dimensional function space; see [16,17].

2. Examples

Let us briefly recall the essential elements of the proof in the finite setting $\mathcal{S} = \{1, \dots, d\}$. For simplicity, consider an irreducible finite (and thus recurrent) chain X and define

$$D_i = \{y \in \mathbb{R} : q(x_1, i)p(i, y)q(i, x_3) > q(x_1, x_2)p(x_2, y)q(x_2, x_3), \forall x_2 \neq i, x_1, x_3 \in \mathcal{S}\}.$$

Suppose that, for a pair of states j_0 and i_0,

$$\int_{D_{i_0}} p(j_0, y)\,\mathrm{d}y > 0. \tag{2.1}$$

Recall the definition of L_n in (1.1) and note that on the event $A_m = \{X_m = j_0, Y_m \in D_{i_0}\}$, with a fixed $m > 1$ and all $n > m$,

$$\begin{aligned} L_n(x_{1:n}, Y_{1:n}) &= L_{m-1}(x_{1:m-1}, Y_{1:m-1}) \\ &\quad \times q(x_{m-1}, x_m)p(x_m, Y_m)q(x_m, x_{m+1})L_{m+1,n}\bigl(x_{(m+1):n}, Y_{(m+1):n}\bigr) \\ &\le L_{m-1}(x_{1:m-1}, Y_{1:m-1}) \\ &\quad \times q(x_{m-1}, i_0)p(i_0, Y_m)q(i_0, x_{m+1})L_{m+1,n}\bigl(x_{(m+1):n}, Y_{(m+1):n}\bigr) \end{aligned}$$

for an appropriate function $L_{m+1:n}$ and where equality is attained only at a path $x_{1:m}$ with $x_m = i_0$. Hence, the mth entry of the optimal path must equal i_0 for any $n \ge m$, that is, $\hat{X}^n_m = i_0$. But, then, given $\hat{X}^n_m$, the first m entries of the optimal path depend only on the values of $Y_1, \dots, Y_m$ and are not affected by Y_k, $k > m$. Hence, the limit (1.2) exists on the event A_m. Since the chain (X, Y) is recurrent, for any fixed m, one of the events $A_{m+1}, A_{m+2}, \dots$ occurs $\mathbb{P}$-a.s. and thus (1.2) holds $\mathbb{P}$-a.s.

Using the same basic idea, let $\tau(k)$, $k \ge 0$, be the times at which the chain (X, Y) revisits the set $\{j_0\} \times D_{i_0}$:

$$\begin{aligned} &\tau(0) = 1, \\ &\tau(k) = \inf\{n > \tau(k-1) : X_n = j_0, Y_n \in D_{i_0}\}, \qquad k \ge 1. \end{aligned}$$

By construction, for any k, on the event $\{\tau(k) \le n\}$,

$$L(x_{1:n}; Y_{1:n}) \le L(x^\tau_{1:n}; Y_{1:n}) \qquad \forall x_{1:n} \in \mathcal{S}^n,$$

where $x^\tau_{1:n}$ is the vector which coincides with $x_{1:n}$ at all but the indices $\tau(1), \dots, \tau(k)$, where its entries equal i_0.

The upper bound is attained if $L(x_{1:n}; Y_{1:n})$ is maximized over $x_{1:n}$, constrained to $x_{\tau(1)} = \cdots = x_{\tau(k)} = i_0$. Since each $x_{\tau(\ell)}$, $\ell = 1, \dots, k$, appears in the product $L(x_{1:n}; Y_{1:n})$ in three adjacent terms, the optimal choice for each segment $x_{\tau(\ell-1)+1:\tau(\ell)-1}$, $\ell = 1, \dots, k$, is determined only by the values of $Y_{\tau(\ell-1)+1}, \dots, Y_{\tau(\ell)-1}$. Hence, in particular, the limit $\lim_{n\to\infty} \hat{X}^n_{1:m}$ exists on any of the events $\{\tau(k-1) < m \le \tau(k) < \infty\}$, $k \ge 1$. By recurrence of j_0 and the condition (2.1), $\mathbb{P}(\tau(k) < \infty) = 1$ and $\lim_{k\to\infty} \tau(k) = \infty$, $\mathbb{P}$-a.s., which verifies the existence of the limit (1.2).

The stopping times $\tau(k)$, $k \ge 1$, form a renewal process, with respect to which both (X, Y) and $\hat{X} = (\hat{X}_m)_{m\ge 1}$ are regenerative (see [4] for more details). As pointed out in [12], the condition (2.1) can be quite restrictive, especially when the transition matrix is sparse. The convergence in (1.2) and the regenerative property are verified in [12] under less conservative conditions, using a more sophisticated construction of the renewal times.

In summary, both [5] and [12] deduce the existence of the limit in (1.2) from the explicit construction of stopping times, based on the discreteness of the hidden process state space. The following example shows that this still may be possible in HMMs with continuous state spaces.

Example 2.1. Consider a linear HMM with Laplacian state and Gaussian observation noises:

$$\mu(u) = \frac{1}{4}\mathrm{e}^{-|u|/2}, \qquad q(u, v) = \frac{1}{4}\mathrm{e}^{-|u-v|/2}, \qquad p(x, y) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-y)^2/2}.$$

In this case, the MAP path is given by

$$\hat{X}^n_{1:n} = \operatorname*{argmin}_{x_{1:n}\in\mathbb{R}^n}\left(|x_1| + (x_1 - Y_1)^2 + \sum_{m=2}^{n} |x_{m-1} - x_m| + (x_m - Y_m)^2\right).$$

Consider the function $x \mapsto f(x) := |a - x| + (x - y)^2 + |x - b|$ for fixed $a, b, y \in \mathbb{R}$. Suppose, without loss of generality, that $a \le b$ and note that f, being strictly convex, is minimized at a unique point $x^* = \operatorname{argmin}_{x\in\mathbb{R}} f(x)$. If $y \in [a, b]$, then, clearly, $x^* \in [a, b]$ and since $f(x) = -a + (y - x)^2 + b$ on this interval, we have $x^* = y$. Consider the case $y \le a$ and suppose $x^* < a$. For $x < a$, $f(x) = a - x + (y - x)^2 + b - x$ and hence $x^* = y + 1$. By strict convexity, this implies that $x^* = y + 1$ if $y < a - 1$ and that $x^* \ge a$ otherwise. Clearly, $x^* \le b$, that is, $x^* \in [a, b]$, which, in turn, implies that $x^* = a$ for $y \in [a - 1, a)$. Similar calculations reveal that $x^* = y - 1$ if $y > b + 1$ and $x^* = b$ if $y \in (b, b + 1]$.

To summarize, $x^* \in [y - 1, y + 1]$ for any $a, b, y \in \mathbb{R}$ and $x^* = y$, whenever $a \le y \le b$. In particular, $\hat{X}^n_{m-1} \in [Y_{m-1} - 1, Y_{m-1} + 1]$ and $\hat{X}^n_{m+1} \in [Y_{m+1} - 1, Y_{m+1} + 1]$ for any $n \ge m + 1$. Hence, on the event

$$A_m := \{Y_{m-1} + 1 \le Y_m \le Y_{m+1} - 1\},$$

$Y_m \in [\hat{X}^n_{m-1}, \hat{X}^n_{m+1}]$ and, consequently, $\hat{X}^n_m = Y_m$. This, in turn, implies that $\hat{X}^n_{1:m} = \hat{X}^{m+1}_{1:m}$ for all $n \ge m + 1$ and the existence of the limit (1.2) on any of A_k, $k \ge m + 1$. Clearly, the A_k's occur infinitely often and hence, as in the discrete case, $\hat{X}^n_{1:m}$ ceases to change, starting from some random, but $\mathbb{P}$-a.s. finite, time n. In particular, (1.2) holds $\mathbb{P}$-a.s.

However, splitting the optimal trajectory into unrelated segments is not the only way to get the convergence in (1.2): the following example shows that the limit may exist without ever being actually attained.

Example 2.2. Consider the linear Gaussian HMM with

$$\mu(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \qquad q(u,v) = \frac{1}{\sqrt{2\pi}} e^{-(u-v)^2/2}, \qquad p(x,y) = \frac{1}{\sqrt{2\pi}} e^{-(x-y)^2/2}.$$

In this case, the conditional law of $X_{1:n}$, given $Y_{1:n}$, is Gaussian and hence

$$\hat{X}^n_{1:n} = \mathbb{E}(X_{1:n}|Y_{1:n}).$$

For any fixed $m \ge 1$, the process $\hat{X}^n_{1:m} = \mathbb{E}(X_{1:m}|Y_{1:n})$, $n \ge m$, is a uniformly integrable vector-valued martingale and hence the limit (1.2) exists by the martingale convergence. In fact, Kalman linear filtering theory (see, e.g., [11]) guarantees that in this case (of controllable and observable dynamics) the stronger $\mathbb{P}$-a.s. exponential convergence holds (see also Remark 3.2 below).

Moreover, $\mathbb{E}(X_{1:m}|Y_{1:n})$ is a deterministic linear map of $Y_{1:n}$ and a calculation reveals that it actually depends on each one of the components in $Y_{1:n}$. Since $Y_{1:n}$ is a non-degenerate Gaussian vector,

$$\mathbb{P}(\hat{X}^n_j = \hat{X}^{n'}_j, \text{for some } j \le m) = 0$$

for any $n' > n \ge m$.

Finally, the next example demonstrates that a finite limit in (1.2) may not exist, even when the hidden state chain is positive recurrent and has countably many states. In fact, it also shows that the optimal MAP path may not be an adequate estimate: in this case, a trajectory of a positive recurrent chain V is estimated as a constant trajectory, diverging to infinity, as $n \to \infty$.

Example 2.3. Consider the HMM with the hidden state process $X_n = (U_n, V_n)$, consisting of independent components U and V. The process $U = (U_n)_{n\ge1}$ is a sequence of i.i.d. random variables uniformly distributed over [0, 1].

$V = (V_n)_{n\ge1}$ is a random walk on positive integers with reflecting boundary at $\{1\}$ and the transition probabilities $P(1,1) = 1-\varepsilon$, $P(1,2) = \varepsilon$ and, for $i \ge 2$,

$$P(i,j) = \begin{cases} \varepsilon \dfrac{(i/(i+1))^2}{1+(i/(i+1))^2}, & j = i+1, \\ \varepsilon \dfrac{1}{1+(i/(i+1))^2}, & j = i-1, \\ 1-\varepsilon, & j = i, \end{cases} \tag{2.2}$$

where $\varepsilon > 0$ is a small fixed constant (in fact, we shall later choose $\varepsilon < e^{-2}/(1+e^{-2}) = 0.119\ldots$). V is a positive recurrent Markov chain with the unique invariant distribution

$$\pi(j) = \begin{cases} \frac{1}{5} C\big(1 + \big(\frac{1}{2}\big)^2\big), & j = 1, \\ C \dfrac{1}{j^2}\left(1 + \left(\dfrac{j}{j+1}\right)^2\right), & j > 1, \end{cases} \tag{2.3}$$

where C is the normalization constant, independent of ε. We shall assume that V is stationary, that is, it is started from $V_1 \sim \pi$. Stationarity is not really required in what follows and is solely a matter of aesthetics (e.g., $\mathbb{P}(V_1 = j) = C/j^2$ will work as well).

Let $a_0 = 0$, $a_i = 8\sum_{j=1}^{i}(1/9)^j$, $i = 1, 2, \dots,$ and set $A_i = [a_{i-1}, a_i)$, $i \geq 1$. Denote by $\ell_i = 8(1/9)^i$ the length of the interval A_i and note that $[0, 1) = \bigcup_{i=1}^{\infty} A_i$.

Now, consider the observation density

$$p((u, v), y) = \mathbf{1}_{\{y\in[0,1]\}}\mathbf{1}_{\{u\notin\bigcup_{i=1}^{v} A_i\}} + \sum_{i=1}^{v} \ell_i^{-1}\mathbf{1}_{\{(u,y)\in A_i\times A_i\}}.$$

As we show below, the MAP estimates of $U_{1:n}$ and $V_{1:n}$ are given by[4]:

$$\begin{aligned} \hat{U}_m^n &= \sum_{j=1}^{\infty} a_{j-1}\mathbf{1}_{\{Y_m\in A_j\}}, \qquad m = 1, \dots, n, \\ \hat{V}_m^n &= \begin{cases} 2, & j^*(n) = 1, \\ j^*(n), & j^*(n) > 1, \end{cases} \end{aligned} \tag{2.4}$$

where $j^*(n) := \max\{j : \sum_{k=1}^{n} \mathbf{1}_{\{Y_k\in A_j\}} > 0\}$. Since all A_j's have positive Lebesgue measure, $j^*(n) \nearrow \infty$ as $n \to \infty$ and, consequently, for any fixed $m \geq 1$,

$$\lim_{n\to\infty} \hat{V}_m^n = \lim_{n\to\infty} j^*(n) = \infty, \qquad \mathbb{P}\text{-a.s.}$$

Before proving (2.4), we shall briefly explain why the optimal path of such a form should be anticipated. Note that since U_i's are uniformly distributed in $[0, 1]$, the choice of $\hat{U}_i^n$'s influences the likelihood (1.1) only through the observation densities. More precisely, whenever $\{Y_m \in A_i\}$ is observed, the maximal gain of ℓ_i^{-1} is obtained if $\hat{U}_m^n \in A_i$ and $\hat{V}_m^n \geq i$ are chosen. On the other hand, the transition probabilities of (2.2) favor paths $\hat{V}_{1:n}^n$ without jumps. Hence, the optimal path $\hat{V}_{1:n}^n$ should be constant and large enough to allow access to the narrowest A_i visited by Y_m's so far, that is, greater or equal to $j^*(n)$. However, if constant $\hat{V}_{1:n}^n$ is chosen, it cannot be too large, as this would decrease the likelihood through the term $\pi(\hat{V}_1^n)$, due to the fast tail decay of the initial distribution π. This heuristics is implemented by an appropriate balancing between all the ingredients of the model.

We shall first check (2.4) in the case $j^*(n) > 1$. To this end, consider the ratio

$$\frac{L_n((u_{1:n}, v_{1:n}), Y_{1:n})}{L_n((\hat{U}_{1:n}^n, \hat{V}_{1:n}^n), Y_{1:n})} = \frac{\pi(v_1)}{\pi(j^*(n))}\prod_{m=2}^{n}\frac{P(v_{m-1}, v_m)}{P(j^*(n), j^*(n))}\prod_{m=1}^{n}\frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)} \tag{2.5}$$

for an arbitrary $u_{1:n}$ and $v_{1:n}$. Let N be the number of jumps in $v_{1:n}$ and $v^*(n) = \max_{k=1,\dots,n} v_k$. Note that $P(v_{m-1}, v_m) = 1 - \varepsilon$ when $v_{m-1} = v_m$ and $P(v_{m-1}, v_m) \leq \varepsilon$ otherwise. Hence, as

[4]The choice of $\hat{U}_m^n$ is not unique, unless the lexicographic order is imposed: for example, $\hat{U}_m^n := Y_m$ yields the same value of the likelihood.

$P(j^*(n), j^*(n)) = 1 - \varepsilon$,

$$\prod_{m=2}^{n} \frac{P(v_{m-1}, v_m)}{P(j^*(n), j^*(n))} \le \left(\frac{\varepsilon}{1-\varepsilon}\right)^N.$$

Further, note that on the event $\{Y_m \in A_i\}$, $p((u_m, v_m), Y_m) \le 1 \vee \ell_i^{-1} = \ell_i^{-1}$ and $p((\hat{U}_m^n, j^*(n)), Y_m) = \ell_i^{-1}$, thus

$$\frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)} \le 1.$$

Moreover, on $\{Y_m \in A_{j^*(n)}\}$,

$$\frac{p((u_m, v_m), Y_m)}{p((\hat{U}_m^n, j^*(n)), Y_m)} \le \frac{\mathbf{1}_{\{v^*(n)<j^*(n)\}} + \ell_{j^*(n)}^{-1}\mathbf{1}_{\{v^*(n)\ge j^*(n)\}}}{\ell_{j^*(n)}^{-1}} \le \frac{\ell_{v^*(n)\wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}}.$$

Plugging these inequalities into (2.5), we get

$$\begin{aligned}\frac{L_n((u_{1:n}, v_{1:n}), Y_{1:n})}{L_n((\hat{U}_{1:n}^n, \hat{V}_{1:n}^n), Y_{1:n})} &\le \frac{\pi(v_1)}{\pi(j^*(n))}\left(\frac{\varepsilon}{1-\varepsilon}\right)^N \frac{\ell_{v^*(n)\wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}} \\ &= \frac{\pi(v_1)}{\pi(v^*(n))}\tilde{\varepsilon}^N \frac{\pi(v^*(n))}{\pi(j^*(n))}\frac{\ell_{v^*(n)\wedge j^*(n)}^{-1}}{\ell_{j^*(n)}^{-1}},\end{aligned} \tag{2.6}$$

where we define $\tilde{\varepsilon} := \varepsilon/(1-\varepsilon)$ for the purposes of brevity. Since $N \ge v^*(n) - v_1$,

$$\begin{aligned}\frac{\pi(v_1)}{\pi(v^*(n))}\tilde{\varepsilon}^N &\le \frac{\pi(v_1)}{\pi(v^*(n))}\tilde{\varepsilon}^{v^*(n)-v_1} \le \left(\frac{v^*(n)}{v_1}\right)^2 \frac{1+(v_1/(v_1+1))^2}{1+(v^*(n)/(v^*(n)+1))^2}\tilde{\varepsilon}^{v^*(n)-v_1} \\ &\le \left(\frac{v^*(n)}{v_1}\right)^2 \tilde{\varepsilon}^{v^*(n)-v_1},\end{aligned}$$

where, in the second inequality, we have used the expression for $\pi(j)$, $j > 1$, from (2.3). In fact, the inequality is also true for $v^*(n) = v_1 = 1$, as both the right- and left-hand sides become 1, and for $v^*(n) > v_1 = 1$, as $\pi(1)$ is less than $C\frac{1}{j^2}(1 + (\frac{j}{j+1})^2)$ evaluated at $j := 1$.

The function $x \mapsto x^2\tilde{\varepsilon}^x$ attains its maximum at $x^* = 2/\log\tilde{\varepsilon}^{-1}$ and is strictly decreasing on (x^*, ∞). Hence, with $\tilde{\varepsilon} < \mathrm{e}^{-2}$, that is, with $\varepsilon < \mathrm{e}^{-2}/(1+\mathrm{e}^{-2})$, for any $y > x \ge 1$, $(y/x)^2\tilde{\varepsilon}^{y-x} < 1$ and hence

$$\frac{\pi(v_1)}{\pi(v^*(n))}\tilde{\varepsilon}^N \le 1. \tag{2.7}$$

The equality holds if and only if $v_{1:n}$ is a constant path, that is, $v_m = v^*(n)$ for all $m = 1, \dots, n$.

Further, if $v^*(n) \le j^*(n)$, then

$$\frac{\pi(v^*(n))}{\pi(j^*(n))} \frac{\ell^{-1}_{v^*(n)\wedge j^*(n)}}{\ell^{-1}_{j^*(n)}} \le \left(\frac{j^*(n)}{v^*(n)}\right)^2 \frac{1+(v^*(n)/(v^*(n)+1))^2}{1+(j^*(n)/(j^*(n)+1))^2}(1/9)^{j^*(n)-v^*(n)} \tag{2.8}$$

$$\le \left(\frac{j^*(n)}{v^*(n)}\right)^2 (1/9)^{j^*(n)-v^*(n)} \le 1, \tag{2.9}$$

where the latter inequality holds since $1/9 < \mathrm{e}^{-2}/(1+\mathrm{e}^{-2})$.

The sequence $\pi(j)$ attains its unique maximum at $j := 2$ and is strictly decreasing for $j \ge 2$. Hence, if $v^*(n) > j^*(n) \ge 2$, then

$$\frac{\pi(v^*(n))}{\pi(j^*(n))} \frac{\ell^{-1}_{v^*(n)\wedge j^*(n)}}{\ell^{-1}_{j^*(n)}} < 1.$$

Plugging (2.7) and (2.8) into (2.6) yields the following inequality for any $u_{1:n}$ and $v_{1:n}$:

$$L_n((u_{1:n}, v_{1:n}), Y_{1:n}) \le L_n((\hat{U}^n_{1:n}, \hat{V}^n_{1:n}), Y_{1:n}),$$

which saturates if and only if $v_m = j^*(n)$, $m = 1, \ldots, n$, thus verifying the optimality of (2.4) on the event $\{j^*(n) > 1\}$.

We shall omit the details in the case $\{j^*(n) = 1\}$, which is treated similarly: the optimal value $\hat{V}^n_m = 2$ is obtained since $\pi(j)$ is maximal at $j = 2$. Of course, as $j^*(n)$ eventually leaves the state 1, the exact value is irrelevant for the main point of the present example, that is, the divergence $\lim_{n\to\infty} \hat{V}^n_m = \infty$.

3. Convergence in the case of log-concave densities

In this section, we establish the existence of the limit (1.2), deducing it from certain strong log-concavity properties of the densities q and p. Hereafter, the following assumptions are in force:

(a1) the initial state density μ is a $C^2(\mathbb{R})$ log-concave function on $\mathbb{R}$ and $-\log \mu(u) \ge 0$;
(a2) the hidden state transition density q is a $C^2(\mathbb{R}^2)$ log-concave function, namely[5] $q(u, v) \propto \mathrm{e}^{-\alpha(u,v)}$, where $\alpha(u, v)$ is a non-negative twice continuously differentiable convex function on $\mathbb{R}^2$;
(a3) the observation density p is a $C^2(\mathbb{R})$ log-concave function in the first argument: $p(x, y) \propto \mathrm{e}^{-\gamma(x,y)}$, where, for each $y \in \mathbb{R}$, the function $x \mapsto \gamma(x, y)$ is non-negative, twice continuously differentiable and strongly convex on $\mathbb{R}$ with $x_*(y) := \operatorname{argmin}_{x\in\mathbb{R}} \gamma(x, y) \in (-\infty, \infty)$ and

$$\frac{\partial^2}{\partial x^2}\gamma(x, y) \ge \kappa > 0 \qquad \forall x, y \in \mathbb{R},$$

with a constant κ;

[5] $f \propto g$ means that f/g is constant.

(a4) for some constant C,

$$-\overline{\lim_{n\to\infty}}\frac{1}{n}\log L_n(X_{1:n}, Y_{1:n}) \le C, \qquad \mathbb{P}\text{-a.s.};$$

(a4) there is a non-decreasing function $g:\mathbb{R}_+ \mapsto \mathbb{R}_+$, growing to $+\infty$ not faster than a polynomial, such that for all $M > 0$,

$$\alpha(x, y) \le M \quad \Longrightarrow \quad \left|\frac{\partial^2}{\partial x \partial y}\alpha(x, y)\right| \le g(M) \qquad \forall x, y \in \mathbb{R}.$$

Remark 3.1. The log-concavity assumptions (a1)–(a3) are quite restrictive. For example, if $Y_n = h(X_n) + w_n$ with $w_n \sim N(0, 1)$, then

$$\frac{\partial^2}{\partial x^2}\gamma(x, y) = \frac{1}{2}\frac{\partial^2}{\partial x^2}\big(y - h(x)\big)^2 = (h'(x))^2 - \big(y - h(x)\big)h''(x),$$

which typically will not admit the uniform lower bound of (a3), unless h is linear, that is, $h''(x) \equiv 0$.

If the assumption (a3) is satisfied, then it implies that $\gamma_*(y) := \gamma(x_*(y), y) \in (-\infty, \infty)$ for all $y \in \mathbb{R}$ and, moreover,

$$\gamma(x, y) - \gamma_*(y) \ge \tfrac{1}{2}\kappa(x - x_*)^2 \qquad \forall x, y \in \mathbb{R}, \tag{3.1}$$

which is essential to our approach.

Assuming that $-\log \mu(u)$, $\alpha(u, v)$ and $\gamma(x, y)$ are non-negative is equivalent to assuming that they are lower-bounded by a constant, that is, that the corresponding densities are bounded.

The assumption (a4) is typically satisfied if the state process X is positively recurrent (explicit recurrence tests can be found in [14]; see also [9]). Finally, (a5) is a technical assumption which is satisfied in most models of practical interest.

Example 3.1. All of the above assumptions are satisfied for the linear HMM

$$\begin{aligned} X_n &= aX_{n-1} + v_n, \qquad n \ge 1, \\ Y_n &= bX_n + w_n, \end{aligned}$$

where $|a| < 1$ and $b \ne 0$ are constants and $v = (v_n)_{n\ge 1}$ and $w = (w_n)_{n\ge 1}$ are independent sequences of i.i.d. random variables with

$$X_0, v_n \sim f_v(x) \propto \mathrm{e}^{-|x|^{2+\delta}} \quad \text{and} \quad w_n \sim f_w(x) \propto \mathrm{e}^{-x^2(1+c|x|^{\delta'})}$$

for some $\delta \ge 0$ and $\delta' \ge 0$, $c \ge 0$.

Theorem 3.1. *The limit in* (1.2) *exists* $\mathbb{P}$*-a.s.*

Proof. To keep the notation simple, we shall prove the convergence in (1.2) for $m = 1$, that is, the limit $\lim_{n\to\infty} \hat{X}_1^n$ exists $\mathbb{P}$-a.s. As will be clear from the proof below, the same arguments imply convergence of $\lim_{n\to\infty} \hat{X}_i^n$ for any $i \le m$ and hence of (1.2) for any fixed $m \ge 1$.

To check $\lim_{n\to\infty} \hat{X}_1^n$, $\mathbb{P}$-a.s., we shall show that on a set of probability one, the series

$$\hat{X}_1^n = \hat{X}_1^1 + \sum_{i=2}^{n} (\hat{X}_1^k - \hat{X}_1^{k-1})$$

is convergent. The proof hinges on the system of inequalities (3.6) and (3.7), which stem from the log-concavity properties assumed in (a1)–(a3). A pigeonhole principle type of argument (Lemma A.1) shows that a sequence satisfying such inequalities must decay at least polynomially backward in time, which, in turn, yields the desired conclusion.

To this end, introduce[6]

$$\begin{aligned} h_n(x_{1:n}) &:= -\log L_n(x_{1:n}, Y_{1:n}) \\ &= -\log \mu(x_1) + \gamma(x_1, Y_1) + \sum_{m=2}^{n} \bigl(\alpha(x_{m-1}, x_m) + \gamma(x_m, Y_m)\bigr). \end{aligned} \tag{3.2}$$

By assumptions (a1)–(a3), $\lim_{R\to\infty} \inf_{\|x_{1:n}\|=R} h_n(x_{1:n}) \to \infty$ and, for any $n \ge 1$, the function

$$x_{1:n} \mapsto h_n(x_{1:n}) + \alpha(x_n, u) \tag{3.3}$$

attains its global minimum at

$$\tilde{X}_{1:n}^n(u) := \operatorname*{argmin}_{x_{1:n}} \bigl(h_n(x_{1:n}) + \alpha(x_n, u)\bigr), \qquad u \in \mathbb{R}.$$

The Hessian matrix of the function defined in (3.3) is positive definite uniformly over $x_{1:n} \in \mathbb{R}^n$ and hence the minimum is unique and $\tilde{X}_{1:n}^n(u)$ is the solution of

$$\operatorname{grad}\bigl(h_n(x_{1:n}) + \alpha(x_n, u)\bigr) = 0.$$

The Jacobian matrix of the function on the left-hand side of this equation with respect to the vector $x_{1:n}$ coincides with the aforementioned Hessian matrix and hence is invertible at any $u \in \mathbb{R}$. Thus, by the implicit function theorem, $u \mapsto \tilde{X}_{1:n}^n(u)$ is continuously differentiable on $\mathbb{R}$.

The usual dynamical programming argument yields the following chain rules:

$$\begin{aligned} \tilde{X}_j^n(x) &= \tilde{X}_j^m(\tilde{X}_{m+1}^n(x)), \qquad x \in \mathbb{R}, j < n, m = j, \dots, n, \\ \hat{X}_j^n &= \tilde{X}_j^m(\hat{X}_{m+1}^n). \end{aligned} \tag{3.4}$$

[6]For $k > \ell$, $\sum_{i=k}^{\ell} \cdots = 0$ is understood.

Hence, for $j < n$, and $j \le m < n$,

$$\begin{aligned}\hat{X}_j^{n+1} - \hat{X}_j^n &= \tilde{X}_j^m(\hat{X}_{m+1}^{n+1}) - \tilde{X}_j^m(\hat{X}_{m+1}^n) \\ &= (\hat{X}_{m+1}^{n+1} - \hat{X}_{m+1}^n)\int_0^1 \frac{\partial}{\partial s}\tilde{X}_j^m\big(s\hat{X}_{m+1}^{n+1} + (1-s)\hat{X}_{m+1}^n\big)\,\mathrm{d}s.\end{aligned} \tag{3.5}$$

The following lemma is the key to a bound on the integrand in (3.5).

Lemma 3.1. *Assume* (a1)–(a3). *Then, for* $j = 1, \dots, n-1$,

$$\left\| \frac{\partial}{\partial x}\tilde{X}_{1:j}^n(x)\right\|^2 \le \frac{2}{\kappa}\left|\mathcal{D}_{12}\alpha(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x))\frac{\partial}{\partial x}\tilde{X}_{j+1}^n(x)\frac{\partial}{\partial x}\tilde{X}_j^n(x)\right| \tag{3.6}$$

and

$$\left\| \frac{\partial}{\partial x}\tilde{X}_{1:n}^n(x)\right\|^2 \le \frac{2}{\kappa}\left|\mathcal{D}_{12}\alpha(\tilde{X}_n^n(x), x)\frac{\partial}{\partial x}\tilde{X}_n^n(x)\right|, \tag{3.7}$$

where $\mathcal{D}_{12}\alpha(x,y) := \frac{\partial^2}{\partial x\,\partial y}\alpha(x,y)$ *and* κ *is as in assumption* $(a3)$.

Proof. Recall that the function (3.3) is strongly convex and the spectral norm of its Hessian is lower bounded by κ. Hence, for any $1 \le j < n$ and $u, v \in \mathbb{R}$, by (3.1),

$$\frac{\kappa}{2}\|\tilde{X}_{1:j}^j(v) - \tilde{X}_{1:j}^j(u)\|^2 \le h_j(\tilde{X}_{1:j}^j(v)) + \alpha(\tilde{X}_j^j(v), u) - h_j(\tilde{X}_{1:j}^j(u)) - \alpha(\tilde{X}_j^j(u), u)$$

since, by definition, the minimum of $h_j(x_{1:j}) + \alpha(x_j, u)$ over $x_{1:j}$ is attained at $\tilde{X}_{1:j}^j(u)$. Further, by the definition of $\tilde{X}_{1:j}^j(v)$,

$$h_j(\tilde{X}_{1:j}^j(v)) + \alpha(\tilde{X}_j^j(v), v) \le h_j(\tilde{X}_{1:j}^j(u)) + \alpha(\tilde{X}_j^j(u), v),$$

which gives

$$\frac{\kappa}{2}\|\tilde{X}_{1:j}^j(v) - \tilde{X}_{1:j}^j(u)\|^2 \le -\alpha(\tilde{X}_j^j(v), v) + \alpha(\tilde{X}_j^j(u), v) + \alpha(\tilde{X}_j^j(v), u) - \alpha(\tilde{X}_j^j(u), u). \tag{3.8}$$

Plugging $v := \tilde{X}_{j+1}^n(x+h)$ and $u := \tilde{X}_{j+1}^n(x)$ into this with $x \in \mathbb{R}$ and using the chain rule (3.4), we get

$$\begin{aligned}\frac{\kappa}{2}\|\tilde{X}_{1:j}^n(x+h) - \tilde{X}_{1:j}^n(x)\|^2 \le{}& -\alpha\big(\tilde{X}_j^n(x+h), \tilde{X}_{j+1}^n(x+h)\big) + \alpha\big(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x+h)\big) \\ &+ \alpha\big(\tilde{X}_j^n(x+h), \tilde{X}_{j+1}^n(x)\big) - \alpha(\tilde{X}_j^n(x), \tilde{X}_{j+1}^n(x)).\end{aligned}$$

Since all of the functions appearing in the latter inequality are twice continuously differentiable, dividing by h^2 and taking $h \to 0$ gives the bound (3.6). Similarly, with $j := n$, $v := x + h$ and $u := x$, (3.8) yields (3.7). □

By assumption (a4),

$$\Omega' := \left\{ \varlimsup_{n\to\infty} \frac{1}{n} \sum_{j=2}^{n} \bigl(\alpha(X_{j-1}, X_j) + \gamma(X_j, Y_j)\bigr) \le C \right\}$$

is an event of full probability and hence it is enough to verify the claimed convergence for all $\omega \in \Omega'$. Clearly, for an $\omega \in \Omega'$,

$$-\log \mu(X_1) + \gamma(X_1, Y_1) + \sum_{j=2}^{n} \bigl(\alpha(X_{j-1}, X_j) + \gamma(X_j, Y_j)\bigr) \le 2Cn \qquad \forall n \ge N(\omega)$$

for an integer $N(\omega) < \infty$. Then, $\hat{X}^n_{1:n}$, being a minimizer, a fortiori satisfies

$$-\log \mu(\hat{X}^n_1) + \gamma(\hat{X}^n_1, Y_1) + \sum_{j=2}^{n} \bigl(\alpha(\hat{X}^n_{j-1}, \hat{X}^n_j) + \gamma(\hat{X}^n_j, Y_j)\bigr) \le 2Cn \qquad \forall n \ge N. \tag{3.9}$$

Hence, for a large fixed constant $M > 4C$ and any $n \ge N$,

$$\#\{j : \alpha(\hat{X}^n_{j-1}, \hat{X}^n_j) + \gamma(\hat{X}^n_j, Y_j) > M\} \le \frac{2Cn}{M} =: \rho n.$$

Similarly,

$$\#\{j : \alpha(\hat{X}^{n+1}_{j-1}, \hat{X}^{n+1}_j) + \gamma(\hat{X}^{n+1}_j, Y_j) > M\} \le \frac{2C(n+1)}{M} = \rho(n+1).$$

There is then an index $m \in [n - 2\rho n, n]$ such that

$$\alpha(\hat{X}^n_{m-1}, \hat{X}^n_m) + \gamma(\hat{X}^n_m, Y_m) \le M \quad \text{and} \quad \alpha(\hat{X}^{n+1}_{m-1}, \hat{X}^{n+1}_m) + \gamma(\hat{X}^{n+1}_m, Y_m) \le M,$$

and, by the assumption (a3),

$$\begin{aligned}
|\hat{X}^{n+1}_m - \hat{X}^n_m| &\le \Bigl|\hat{X}^{n+1}_m - \operatorname*{argmin}_{x\in\mathbb{R}} \gamma(x, Y_m)\Bigr| + \Bigl|\hat{X}^n_m - \operatorname*{argmin}_{x\in\mathbb{R}} \gamma(x, Y_m)\Bigr| \\
&\le \sqrt{\frac{2}{\kappa}} \bigl(\gamma(\hat{X}^{n+1}_m, Y_m) - \gamma_*(Y_m)\bigr)^{1/2} + \sqrt{\frac{2}{\kappa}} \bigl(\gamma(\hat{X}^n_m, Y_m) - \gamma_*(Y_m)\bigr)^{1/2} \\
&\le \sqrt{\frac{2}{\kappa}\gamma(\hat{X}^n_m, Y_m)} + \sqrt{\frac{2}{\kappa}\gamma(\hat{X}^{n+1}_m, Y_m)} \le \sqrt{\frac{8M}{\kappa}}.
\end{aligned}$$

Plugging this estimate into (3.5), we get (for $j := 1$)

$$|\hat{X}^{n+1}_1 - \hat{X}^n_1| \le \sqrt{\frac{8M}{\kappa}} \int_0^1 \left| \frac{\partial}{\partial s} \tilde{X}^{m-1}_1\bigl(s\hat{X}^{n+1}_m + (1-s)\hat{X}^n_m\bigr) \right| \mathrm{d}s. \tag{3.10}$$

Introduce

$$\check{X}_m^m(s) := s\hat{X}_m^{n+1} + (1-s)\hat{X}_m^n,$$
$$\check{X}_j^m(s) := \tilde{X}_j^{m-1}(\check{X}_m^m(s)), \qquad j = 1, \dots, m-1,$$

and define

$$c_j(s) := \frac{2}{\kappa}|\mathcal{D}_{12}\alpha(\check{X}_j^m(s), \check{X}_{j+1}^m(s))|, \qquad j < m,$$
$$b_j(s) := \left|\frac{\partial}{\partial x}\tilde{X}_j^{m-1}(\check{X}_m^m(x))\Big|_{x:=s}\right|.$$

Then, from (3.6) and (3.7) (the dependence on s is now omitted for brevity),

$$\sum_{i=1}^{j} b_i^2 \le c_j b_j b_{j+1}, \qquad j < m-1, \tag{3.11}$$

$$\sum_{i=1}^{m-1} b_i^2 \le c_{m-1} b_{m-1} \tag{3.12}$$

and (3.10) reads

$$|\hat{X}_1^{n+1} - \hat{X}_1^n| \le \sqrt{\frac{8M}{\kappa}} \int_0^1 b_1(s)\,\mathrm{d}s. \tag{3.13}$$

Lemma 3.2. *For any* $s \in [0,1]$, $x > 0$ *and* $g(\cdot)$ *as in* (a5),

$$\#\left\{j < m : c_j(s) > \frac{2}{\kappa}g(x)\right\} \le \frac{4C}{x(1-2\rho)}m. \tag{3.14}$$

Proof. The function $u \mapsto \min_{x_{1:n}}(h_n(x_{1:n}) + \alpha(x_n, u))$ is convex and hence

$$\begin{aligned}
\sum_{j=2}^{m}\alpha(\check{X}_{j-1}^m, \check{X}_j^m) &\le h_{m-1}(\check{X}_{1:m-1}^m) + \alpha(\check{X}_{m-1}^m, \check{X}_m^m) \\
&= \min_{x_{1:m-1}}\bigl(h_{m-1}(x_{1:m-1}) + \alpha\bigl(x_{m-1}, s\hat{X}_m^{n+1} + (1-s)\hat{X}_m^n\bigr)\bigr) \\
&\le s\min_{x_{1:m-1}}\bigl(h_{m-1}(x_{1:m-1}) + \alpha(x_{m-1}, \hat{X}_m^{n+1})\bigr) \\
&\quad + (1-s)\min_{x_{1:m-1}}\bigl(h_{m-1}(x_{1:m-1}) + \alpha(x_{m-1}, \hat{X}_m^n)\bigr) \\
&= s\bigl(h_{m-1}(\hat{X}_{1:m-1}^{n+1}) + \alpha(\hat{X}_{m-1}^{n+1}, \hat{X}_m^{n+1})\bigr) \\
&\quad + (1-s)\bigl(h_{m-1}(\hat{X}_{1:m-1}^n) + \alpha(\hat{X}_{m-1}^n, \hat{X}_m^n)\bigr) \\
&\le 2C(n+1),
\end{aligned}$$

where the latter inequality follows from (3.9). Hence,

$$\#\{j \le m : \alpha(\check{X}_{j-1}^m, \check{X}_j^m) > x\} \le \frac{2C(n+1)}{x}$$

and, since $m \ge (1-2\rho)n$, (3.14) follows from the assumption (a5). □

Now, by Corollary A.1 in the Appendix, applied to (3.11)–(3.12) and (3.14), for any $\beta > 1$, there is a constant C_β such that

$$b_1 \le C_\beta m^{-\beta} \le C_\beta (1-2\rho)^{-\beta} n^{-\beta} \tag{3.15}$$

for all sufficiently large n and, thus, by (3.13), the sequence $|\hat{X}_1^{n+1} - \hat{X}_1^n|$, $n \ge 1$, is summable, which verifies the existence of the limit (1.2). □

Remark 3.2. When the hidden state process is a Gaussian autoregression, that is, when $\alpha(x, y) = \frac{1}{2}(y - bx)^2$ with a constant $b \ne 0$, $|\mathcal{D}_{12}\alpha(x, y)| \equiv b$ and Lemma A.1(1) implies the exponential bound in (3.15), confirming the results deducible from Kalman linear filtering theory.

4. Concluding remarks

As indicated by the examples of Section 2 and the partial results of Theorem 3.1, the convergence in (1.2) appears to be a non-trivial issue. Analogous problems have been discussed in the engineering literature. In fact, the MAP path estimation can be viewed as an optimal control problem, in which one is required to minimize the cost functional $h_n(x_{1:n})$ defined in (3.2), where the term $\alpha(x_{m-1}, x_m)$ is interpreted as the cost incurred by the control effort (needed to move from x_{m-1} to x_m) and $\gamma(x_m, Y_m)$ is the cost paid for the deviation of the state from Y_m. This setting appears in [1], Chapter I, Section 1.7, as the "smoothing" problem and, in the control literature, is often referred to as the *tracking* problem. From the control theory perspective, the existence of the limit in (1.2) means that the optimal control and the corresponding optimal trajectory cease to depend on the future values of the exogenous signal Y.

Among other related questions, the convergence (1.2) of the optimal trajectory is part of the "asymptotic control theory" program initiated by R. Kalman, R. Bellman and R. Bucy, at the dawn of modern control theory. In the linear state/quadratic cost (LQ) setting of R. Kalman, the control problem admits an elegant closed-form solution for each fixed horizon n and the study of the limit (1.2) reduces to the stability analysis of the associated Riccati equation (a comprehensive treatment of the LQ problem can be found in, e.g., [11]).

To the best of our knowledge, asymptotic analysis beyond the LQ case has been carried out only for a limited number of nonlinear models. Bellman and Bucy [2] found a remarkable explicit solution to a quite general scalar continuous-time control problem, amenable to asymptotic analysis. A vector control problem with linear state dynamics and convex costs was studied in [3].

While much progress has been made in the optimal control theory on the *infinite horizon* (see, e.g., [7,15]), we were not able to track any results directly applicable to the question under consideration.

Another possible connection, remaining elusive at the moment, is to the stability theory of nonlinear filtering equations, developed during the last decade (see, e.g., the survey [8]).

Appendix: A supporting lemma

Lemma A.1. *Consider the system of inequalities*

$$\begin{aligned} &\sum_{i=1}^{j} b_i^2 \le b_j b_{j+1} c_j, \qquad j = 1, \dots, n-1, \\ &\sum_{i=1}^{n} b_i^2 \le b_n c_n, \end{aligned} \tag{A.1}$$

where b_i and c_i, $i = 1, \dots, n$, are non-negative real numbers, and let θ and θ' be arbitrary positive constants:

(1) *If $c_i \le \theta$, $i = 1, \dots, n$, then*

$$b_1 \le \sqrt{\theta \mathrm{e}} \exp\left(-\frac{n}{2\mathrm{e}(\theta^2 \vee \theta)}\right) \qquad \text{for } n \ge \theta^2 \mathrm{e}. \tag{A.2}$$

(2) *If, for a non-decreasing non-negative function $g : \mathbb{R}_+ \mapsto \mathbb{R}_+$,*

$$\#\{i \le n : c_i \ge g(x)\} \le \frac{\theta n}{x} \qquad \forall x > 0, \tag{A.3}$$

and $c_n \le \theta'$, then, for any $p \in (0, 1)$ and $\ell > \theta$,

$$b_1 \le \sqrt{g(\ell)} n^{-p\ell/(4\theta)} \qquad \text{for } n > \left(\frac{\ell(\theta'^2 \vee g(\ell))}{\theta}\right)^{1/(1-p)}. \tag{A.4}$$

(3) *If only (A.3) holds, then, for any $p \in (0, 1)$ and $\ell > \theta$,*

$$b_1 \le g(2\theta n)\sqrt{g(\ell)} n^{-p\ell/(4\theta)} \qquad \text{for } n > \left(\frac{\ell(1 \vee g(\ell))}{\theta}\right)^{1/(1-p)}. \tag{A.5}$$

Proof. (1) The second inequality in (A.1) and $c_n \le \theta$ together imply that $b_n^2 \le b_n \theta$ and, in turn, that $b_1^2 + \cdots + b_n^2 \le \theta^2$. Fix a constant $\eta \in (0, 1)$ and let $m_1 := \lfloor \theta^2/\eta \rfloor$. Then, at most half of the b_i's with $i \in [n - 2m_1, n]$ are greater than $\sqrt{\eta}$ and hence there is an index $k_1 \in [n - 2m_1, n]$ such that $b_{k_1} \le \sqrt{\eta}$ and $b_{k_1+1} \le \sqrt{\eta}$. The inequality corresponding to $j := k_1$ in (A.1) then gives the bound $b_1^2 + \cdots + b_{k_1}^2 \le b_{k_1} b_{k_1+1} c_{k_1} \le \eta\theta$.

Similarly, let $m_2 := \lfloor \theta/\eta \rfloor$. There is then an index $k_2 \in [k_1 - 2m_2 : k_1]$ such that $b_{k_2} \le \eta$ and $b_{k_2+1} \le \eta$ and, again applying (A.1), $b_1^2 + \cdots + b_{k_2}^2 \le b_{k_2} b_{k_2+1} c_{k_2} \le \eta^2\theta$. This argument can be iterated at least

$$\left\lfloor \frac{n}{2(m_1 \vee m_2)} \right\rfloor = \left\lfloor \frac{\eta n}{2(\theta^2 \vee \theta)} \right\rfloor$$

times and thus

$$b_1^2 \le \theta \eta^{\lfloor 1/2\eta n/(\theta^2 \vee \theta) \rfloor} \le \frac{\theta}{\eta}\big((\eta^\eta)^{1/2/(\theta^2 \vee \theta)}\big)^n.$$

The best rate is obtained at $\eta := \mathrm{e}^{-1}$, which yields the bound (A.2).

(2) For a fixed $\ell \ge \theta$, by (A.3),

$$\#\{i \le n : c_i \ge g(\ell)\} \le \frac{\theta n}{\ell} := rn \tag{A.6}$$

and thus at least half of the c_i's with $i \in [n - 2rn, n]$ do not exceed $g(\ell)$. Fix a constant $p \in (0, 1)$ and let $\eta := n^{-p/2}$. Suppose that for all $i \in [n - 2rn, n]$ such that $c_i \le g(\ell)$, either $b_i \ge \eta$ or $b_{i+1} \ge \eta$, or both. Then,

$$\#\{i \in [n - 2rn : n] : b_i \ge \eta\} \ge rn.$$

But, on the other hand, by the second inequality in (A.1) and as $c_n \le \theta'$, $b_n^2 \le b_n\theta'$ and $b_1^2 + \cdots + b_n^2 \le \theta'^2$, we have

$$\#\{i \in [n - 2rn : n] : b_i \ge \eta\} \le \frac{\theta'^2}{\eta^2} = \theta'^2 n^p.$$

This contradicts the previous estimate if n is large enough, namely, if $n > (\ell\theta'^2/\theta)^{1/(1-p)}$. Thus, for such n, there is an index $m_1 \in [n - 2rn : n]$ such that $b_{m_1} \le \eta$, $b_{m_1+1} \le \eta$ and $c_{m_1} \le g(\ell)$.

Now, by the inequality in (A.1) corresponding to $j := m_1$,

$$b_1^2 + \cdots + b_{m_1}^2 \le b_{m_1} b_{m_1+1} c_{m_1} \le \eta^2 g(\ell) \tag{A.7}$$

for which the above consideration can be repeated. Namely, by (A.6), there are at least rn indices $i \in [m_1 - 2rn, m_1]$ for which $c_i \le g(\ell)$. Suppose that, for all of them, either $b_i \ge \eta^2$ or $b_{i+1} \ge \eta^2$, or both. Then

$$\#\{i \in [m_1 - 2rn : m_1] : b_i \ge \eta^2\} \ge rn,$$

while (A.7) implies that

$$\#\{i \in [m_1 - 2rn : m_1] : b_i \ge \eta^2\} \le \frac{\eta^2 g(\ell)}{\eta^4} = n^p g(\ell),$$

which is a contradiction for n large enough, that is, for $n > (\ell g(\ell)/\theta)^{1/(1-p)}$. Hence, there is an $m_2 \in [m_1 - 2rn : m_1]$ such that $b_{m_2} \le \eta^2$, $b_{m_2+1} \le \eta^2$ and $c_{m_2} \le g(\ell)$, and thus, by (A.1), we have

$$b_1^2 + \cdots + b_{m_2}^2 \le \eta^4 g(\ell).$$

This argument can be iterated at least $\lfloor 1/(2r) \rfloor$ times, which yields the bound

$$b_1^2 \le g(\ell)(\eta^{1/2r})^2 = g(\ell) n^{-p\ell/2\theta}.$$

(3) Note that $b_i' := b_i / g(2\theta n)$, $i = 1, \dots, n$, satisfy the inequalities (A.1) with the c_i's replaced by $c_i' := c_i$, $i = 1, \dots, n-1$, and $c_n' := c_n / g(2\theta n)$. By (A.3),

$$\#\{i \le n : c_i \ge g(2\theta n)\} \le \frac{\theta n}{2\theta n} = 1/2,$$

that is, all c_i's are less than $g(2\theta n)$ and, in particular, $c_n \le g(2\theta n)$, that is, $c_n' \le 1$. Moreover, assuming that $g(2\theta n) \ge 1$, we have

$$\#\{i \le n : c_i' \ge g(x)\} \le \#\{i \le n : c_i \ge g(x)\} \le \frac{\theta n}{x} \qquad \forall x > 0.$$

Hence, by (A.4), we have

$$b_1' \le \sqrt{g(\ell)} n^{-p\ell/(2\theta)} \qquad \text{for } n > \left(\frac{\ell(1 \vee g(\ell))}{\theta} \right)^{1/(1-p)},$$

which, in turn, gives (A.5). □

Corollary A.1. *Under the assumption* (A.3) *with* $g(\cdot)$ *growing to* $+\infty$ *not faster than a polynomial, for any* $\beta > 1$, *there is a constant* C_β, *such that*

$$b_1 \le C_\beta n^{-\beta}$$

for all sufficiently large n.

Proof. This follows from (3) of Lemma A.1. □

Acknowledgements

The authors are grateful to M. Margaliot and R. Van Handel for the enlightening discussion on the issues raised in this paper and appreciate the thorough proofreading by the referees. Research of P. Chigansky was supported by an ISF Grant 314/09. Research of Yaacov Ritov was supported by an ISF grant.

References

[1] Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton Univ. Press. MR0090477
[2] Bellman, R. and Bucy, R. (1964). Asymptotic control theory. *J. Soc. Indust. Appl. Math. Ser. A Control* **2** 11–18. MR0168401
[3] Bucy, R.S. (1966). New results in asymptotic control theory. *SIAM J. Control* **4** 397–402. MR0213931

[4] Caliebe, A. (2006). Properties of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory* **52** 41–51. MR2237334

[5] Caliebe, A. and Rösler, U. (2002). Convergence of the maximum a posteriori path estimator in hidden Markov models. *IEEE Trans. Inform. Theory* **48** 1750–1758. MR1929991

[6] Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models*. New York: Springer. MR2159833

[7] Carlson, D., Haurie, A. and Leizarowitz, A. (1991). *Infinite Horizon Optimal Control: Deterministic and Stochastic Systems*, 2nd ed. Berlin: Springer.

[8] Chigansky, P., Liptser, R. and Van Handel, R. (2010). Intrinsic methods in filter stability. In *Handbook of Nonlinear Filtering* (D. Crisan and B. Rozovskii, eds.). Oxford: Oxford Univ. Press. To appear.

[9] Genon-Catalot, V., Jeantheau, T. and Larédo, C. (2000). Stochastic volatility models as hidden Markov models and statistical applications. *Bernoulli* **6** 1051–1079. MR1809735

[10] Kogan, J.A. (1996). Hidden Markov models estimation via the most informative stopping times for the Viterbi algorithm. In *Image Models (and Their Speech Model Cousins) (Minneapolis, MN, 1993/1994)*. *IMA Vol. Math. Appl.* **80** 115–130. New York: Springer. MR1435746

[11] Kwakernaak, H. and Sivan, R. (1972). *Linear Optimal Control Systems*. New York: Wiley. MR0406607

[12] Lember, J. and Koloydenko, A. (2008). The adjusted Viterbi training for hidden Markov models. *Bernoulli* **14** 180–206. MR2401659

[13] Lember, J. and Koloydenko, A. (2010). A constructive proof of the existence of Viterbi processes. *IEEE Trans. Inform. Theory* **56** 2017–2033.

[14] Meyn, S. and Tweedie, R.L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge: Cambridge Univ. Press. MR2509253

[15] Zaslavski, A.J. (2006). *Turnpike Properties in the Calculus of Variations and Optimal Control*. *Nonconvex Optimization and Its Applications* **80**. New York: Springer. MR2164615

[16] Zeitouni, O. and Dembo, A. (1987). A maximum a posteriori estimator for trajectories of diffusion processes. *Stochastics* **20** 221–246. MR0878313

[17] Zeitouni, O. and Dembo, A. (1988). An existence theorem and some properties of maximum a posteriori estimators of trajectories of diffusions. *Stochastics* **23** 197–218. MR0928355

Received September 2009 and revised March 2010