![HAL open science]

Online EM Algorithm for Hidden Markov Models

Olivier Cappé

Online EM Algorithm for Hidden Markov Models

Olivier Cappé

LTCI, Telecom ParisTech & CNRS

Abstract

Online (also called "recursive" or "adaptive") estimation of fixed model parameters in hidden Markov models is a topic of much interest in times series modelling. In this work, we propose an online parameter estimation algorithm that combines two key ideas. The first one, which is deeply rooted in the Expectation-Maximization (EM) methodology consists in reparameterizing the problem using complete-data sufficient statistics. The second ingredient consists in exploiting a purely recursive form of smoothing in HMMs based on an auxiliary recursion. Although the proposed online EM algorithm resembles a classical stochastic approximation (or Robbins-Monro) algorithm, it is sufficiently different to resist conventional analysis of convergence. We thus provide limited results which identify the potential limiting points of the recursion as well as the large-sample behavior of the quantities involved in the algorithm. The performance of the proposed algorithm is numerically evaluated through simulations in the case of a noisily observed Markov chain. In this case, the algorithm reaches estimation results that are comparable to that of the maximum likelihood estimator for large sample sizes.

Keywords Hidden Markov Models, Expectation-Maximization Algorithm, Online Estimation, Recursive Estimation, Stochastic Approximation, Smoothing

1   Introduction

Hidden Markov modelling is a key concept of statistical time series analysis, which has had a wide-ranging practical impact over the latest forty years. Hidden Markov models (HMMs) in their classical form (i.e., when the state variable is finite-valued) are sufficiently simple to give rise to efficient inference procedures while allowing for useful modelling of various practical situations. Ever since the pioneering contributions of Baum and Eagon (1967), Baum et al. (1970), the EM (Expectation-Maximization) algorithm has been the method of choice for parameter inference in HMMs. The EM algorithm is a dedicated numerical optimization routine which aims at maximizing the (log) likelihood of a batch of observations. It tends to be preferred to its alternatives due to its robustness and ease of implementation.

This contribution is devoted to online parameter estimation for HMMs, in which the available observations are only scanned once and never stored, allowing for a continuous adaptation of the parameters along a potentially infinite data stream. In the case of HMMs, online parameter estimation is a challenging task due to the non-trivial dependence between the observations. The EM-inspired methods proposed so far have been either based on finite-memory approximations of the required smoothing computations (Krishnamurthy and Moore, 1993) or on finite-memory approximations of the data log-likelihood itself (Rydén, 1997). An alternative consists in using gradient-based methods (Le Gland and Mevel, 1997) which do not directly follow the principles of

1

the EM algorithm. Kantas et al. (2009) provide a comprehensive recent review of these methods, including more advanced aspects for models that require the use of simulation-based methods. Recently, Mongillo and Denève (2008) proposed an online version of the EM algorithm for HMMs in the case where both the states and observations take a finite number of values. The key ingredient of this algorithm is a recursion which allows for recursive computation of smoothing functionals required by the EM algorithm. However, this recursion appears to be very specific and its potential application to more general types of HMMs is not addressed by Mongillo and Denève (2008).

The purpose of this paper is to build on the idea of Mongillo and Denève (2008) in light of the framework introduced by Cappé and Moulines (2009) for online EM estimation in the case of independent observations. The first contribution of the paper is an algorithm that extends the proposal of Mongillo and Denève (2008) to general HMMs, with possibly continuous observations. This algorithm is based on the key observation that the recursion used in Mongillo and Denève (2008) is an instance of the recursive smoothing scheme for sum functionals introduced by Zeitouni and Dembo (1988), Elliott et al. (1995). Although a complete analysis of the proposed algorithm is currently lacking, we provide a first result that identifies the possible limiting points of the algorithm. These coincide with the stationary points of a limiting EM mapping that may be interpreted as the limit of the EM recursion under an infinite number of observations. This interpretation which generalizes the argument of Cappé and Moulines (2009) for the case of independent observations also provides some interesting insight regarding the behavior of the batch EM algorithm when used with a large number of observations.

The remaining of the paper is organized as follows. Section 2 opens with a brief review of our modelling assumptions and of smoothing computations in HMMs. The proposed online EM algorithm is then introduced in Section 2.3. Section 3 is devoted to a discussion of the online EM algorithm and, in particular, of its connections with previous works and of its numerical complexity. Section 4 contains preliminary results pertaining to the convergence of the method, with corresponding proofs to be found in the appendix. Finally, in Section 5 we apply the online EM algorithm to the estimation of the parameters of a Markov chain observed in Gaussian noise and illustrate its performance through numerical simulations.

2 Online EM Algorithm for HMMs

2.1 Model and Notations

It is assumed that the state and observation sequences, $(X_t, Y_t)_{t \in \mathbb{Z}}$ are generated under a *stationary* Hidden Markov model with unknown parameter $\theta_\star$, where (X_t) takes its values in some finite set $\mathcal{X}$. The notations $\ell_{\theta_\star}$, $p_{\theta_\star}$, $\mathrm{P}_{\theta_\star}$ and $\mathrm{E}_{\theta_\star}$ refer to, respectively, the likelihood, the joint density of the states and observations, the probability, and, the expectation under the model parameterized by $\theta_\star$. In practice $\theta_\star$ is unknown and one only has access to the observation sub-sequence $(Y_t)_{t \geq 0}$, where, by convention, the initial observation time is taken to be 0. In this context, the initial pdf ν of X_0 is arbitrary and $\ell_{\nu,\theta}$, $p_{\nu,\theta}$, $\mathrm{P}_{\nu,\theta}$ and $\mathrm{E}_{\nu,\theta}$ are used to denote, respectively, the likelihood, the joint density of the states and observations, the probability, and, the expectation under the non-stationary model that has initial pdf ν and θ as parameter. Hence, P_θ refers to the probability under θ of the stationary HMM process $(X_t, Y_t)_{t \in \mathbb{Z}}$ while $\mathrm{P}_{\nu,\theta}$ denotes the probability of the non-stationary process $(X_t, Y_t)_{t \in \mathbb{N}}$ which is started with initial distribution ν for X_0. Note that ν itself is not considered as a model parameter as it cannot be estimated consistently from a single trajectory (see Chapters 10 and 12 of Cappé et al., 2005 for further discussion of this issue). The

state transition matrix and state conditional pdf (probability density function of Y_t given X_t) that characterize the HMM are denoted, respectively, by $q_\theta(x, x')$ and $g_\theta(x, y)$.

To make the EM recursion explicit, a key requirement is that the model belongs to an exponential family. In the following, we will thus make the following assumptions.

Assumption 1.

(i) Exponential Family

$$p_\theta(x_t, y_t | x_{t-1}) = h(x_t, y_t) \exp \left(\langle \psi(\theta), s(x_{t-1}, x_t, y_t) \rangle - A(\theta) \right) , \tag{1}$$

where $\langle \cdot \rangle$ denotes the scalar product, $s : (x', x, y) \in \mathcal{X}^2 \times \mathcal{Y} \mapsto s(x', x, y) \in \mathcal{S}$ is the vector of complete-data sufficient statistics, $\psi(\cdot)$ is the (non-necessarily invertible) function that maps θ to the natural parameterization and $A(\cdot)$ is the log-partition function.

(ii) Explicit M-Step *For all $S \in \mathcal{S}$, the complete-data maximum likelihood equation*

$$\nabla_\theta \psi(\theta) S - \nabla_\theta A(\theta) = 0 ,$$

where ∇_θ denotes the gradient, has a unique solution denoted by $\bar{\theta}(S)$.

Assumption 1–(ii) states that the function $\bar{\theta} : S \in \mathcal{S} \mapsto \bar{\theta}(S) \in \Theta$ that returns the complete-data maximum likelihood estimator corresponding to any feasible value of the sufficient statistics is available in closed-form. Note that the form used in (1) is in fact slightly more general than the HMM case. Indeed, if both q_θ and g_θ belong to exponential families:

$$q_\theta(x', x) = h^q(x', x) \exp \left(\langle \psi^q(\theta), s^q(x', x) \rangle - A^q(\theta) \right) ,$$
$$g_\theta(x, y) = h^g(x, y) \exp \left(\langle \psi^g(\theta), s^g(x, y) \rangle - A^g(\theta) \right) , \tag{2}$$

we then have $A(\theta) = A^q(\theta) + A^g(\theta)$,

$$\psi(\theta) = \begin{pmatrix} \psi^q(\theta) \\ \psi^g(\theta) \end{pmatrix} \text{ and } s(x', x, y) = \begin{pmatrix} s^q(x', x) \\ s^g(x, y) \end{pmatrix} .$$

We will not make use of this specific structure when describing the algorithm and we thus stick to the concise representation of (1). Sections and 3.2 and 5 provide more details on the nature of the function s in specific examples of HMMs.

Under Assumption 1, the k-th iteration of the usual EM algorithm applied to observations $Y_{0:n} = (Y_0, \dots, Y_n)$ takes the following familiar form:

E-Step Compute

$$S_{k+1} = \frac{1}{n} \mathrm{E}_{\nu, \theta_k} \left[\left. \sum_{t=1}^n s(X_{t-1}, X_t, Y_t) \right| Y_{0:n} \right] . \tag{3}$$

M-Step Update the parameter estimate to $\theta_{k+1} = \bar{\theta}(S_{k+1})$.

To avoid unnecessary notational complexity, we have omitted in (3) the initial term $\frac{1}{n} \log p_{\nu, \theta}(x_0, y_0)$ from the normalized complete-data log-likelihood. The influence of this term is vanishing with n and it is not necessary to take it into account for online estimation.

2.2 Recursive Form of Smoothing

We briefly recall here a key ingredient of the proposed algorithm which makes it possible to compute recursively the normalized sum $\frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t)\mid Y_{0:n}\right]$ by use of an auxiliary recursion. Curiously, this idea which dates back to, at least, Zeitouni and Dembo (1988) and has been extensively studied by Elliott et al. (1995) remains largely under-exploited (see discussion in Chapter 4 of Cappé et al., 2005).

In addition to the usual filter,

$$\phi_{n,\nu,\theta}(x) = \mathrm{P}_{\nu,\theta}\left(X_n = x\mid Y_{0:n}\right), \tag{4}$$

define the following intermediate quantity

$$\rho_{n,\nu,\theta}(x) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\left.\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t)\right| Y_{0:n}, X_n = x\right]. \tag{5}$$

Obviously, these two quantities allow the computation of the sum of interest as

$$\sum_{x\in\mathcal{X}} \phi_{n,\nu,\theta}(x)\rho_{n,\nu,\theta}(x) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\left.\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t)\right| Y_{0:n}\right].$$

The appeal of this decomposition is that $\phi_{n,\nu,\theta}$ and $\rho_{n,\nu,\theta}$ can be updated recursively according to the following proposition.

Proposition 1. Initialization *For $x \in \mathcal{X}$, set*

$$\phi_{0,\nu,\theta}(x) = \frac{\nu(x)g_\theta(x, Y_0)}{\sum_{x'\in\mathcal{X}}\nu(x')g_\theta(x', Y_0)},$$

$$\rho_{0,\nu,\theta}(x) = 0.$$

Recursion *For $n \geq 0$ and $x \in \mathcal{X}$, it holds that*

$$\phi_{n+1,\nu,\theta}(x) = \frac{\sum_{x'\in\mathcal{X}}\phi_{n,\nu,\theta}(x')q_\theta(x', x)g_\theta(x, Y_{n+1})}{\sum_{x',x''\in\mathcal{X}^2}\phi_{n,\nu,\theta}(x')q_\theta(x', x'')g_\theta(x'', Y_{n+1})} \tag{6}$$

$$\rho_{n+1,\nu,\theta}(x) = \sum_{x'\in\mathcal{X}}\left\{\frac{1}{n+1}s(x', x, Y_{n+1})\right.$$

$$\left. + \left(1 - \frac{1}{n+1}\right)\rho_{n,\nu,\theta}(x')\right\}\frac{\phi_{n,\nu,\theta}(x')q_\theta(x', x)}{\sum_{x''\in\mathcal{X}}\phi_{n,\nu,\theta}(x'')q_\theta(x'', x)}. \tag{7}$$

In Proposition 1 above, the rightmost term in (7),

$$r_{n+1,\nu,\theta}(x'|x) = \frac{\phi_{n,\nu,\theta}(x')q_\theta(x', x)}{\sum_{x''\in\mathcal{X}}\phi_{n,\nu,\theta}(x'')q_\theta(x'', x)},$$

corresponds to the *backward retrospective probability* $\mathrm{P}_{\nu,\theta}(X_n = x'|X_{n+1} = x, Y_{0:n})$, which does not depend on the newly available observation Y_{n+1}. The main argument in proving Proposition 1 is to check that

$$\mathrm{P}_{\nu,\theta}(X_t = x_t, X_{t+1} = x_{t+1}|X_{n+1} = x_{n+1}, Y_{0:n+1}) =$$

$$\sum_{x_n\in\mathcal{X}} \mathrm{P}_{\nu,\theta}(X_t = x_t, X_{t+1} = x_{t+1}|X_n = x_n, Y_{0:n})\mathrm{P}_{\nu,\theta}(X_n = x_n|X_{n+1} = x_{n+1}, Y_{0:n}),$$

for all indices $0 \leq t \leq n - 1$ which implies the claimed result by summation on t.

2.3 Online EM Algorithm

Proposition 1 constitutes a recursive rewriting of the computation required to carry out the E-step in the batch EM algorithm. By analogy with the case of independent observations studied in Cappé and Moulines (2009), the proposed online EM algorithm for HMMs takes the following form.

Algorithm 1. *Chose a decreasing sequence $(\gamma_n)_{n \geq 1}$ of step-sizes, which satisfy the usual stochastic approximation requirement that $\sum_{n \geq 1} \gamma_n = \infty$ and $\sum_{n \geq 1} \gamma_n^2 < \infty$. Also select a parameter initialization $\hat{\theta}_0$ and a minimal number of observations $n_{\min}$ required before performing the first parameter update.*

Initialization *Compute, for $x \in \mathcal{X}$,*

$$\hat{\phi}_0(x) = \frac{\nu(x) g_{\hat{\theta}_0}(x, Y_0)}{\sum_{x' \in \mathcal{X}} \nu(x') g_{\hat{\theta}_0}(x', Y_0)} \,,$$

$$\hat{\rho}_0(x) = 0 \,.$$

Recursion *For $n \geq 0$,*

Compute, for $x \in \mathcal{X}$,

$$\hat{\phi}_{n+1}(x) = \frac{\sum_{x' \in \mathcal{X}} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x, Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} \,, \tag{8}$$

$$\hat{\rho}_{n+1}(x) = \sum_{x' \in \mathcal{X}} \left\{ \gamma_{n+1} s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \hat{\rho}_n(x') \right\} \frac{\hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x)}{\sum_{x'' \in \mathcal{X}} \hat{\phi}_n(x'') q_{\hat{\theta}_n}(x'', x)} \,. \tag{9}$$

If $n \geq n_{\min}$, update the parameter according to

$$\hat{\theta}_{n+1} = \bar{\theta} \left(\sum_{x \in \mathcal{X}} \hat{\rho}_{n+1}(x) \hat{\phi}_{n+1}(x) \right) \,, \tag{10}$$

otherwise, set $\hat{\theta}_{n+1} = \hat{\theta}_n$.

In Algorithm 1, the role of $n_{\min}$ is only to guarantee that the M-step update is numerically well-behaved as it is well-known that in most models the maximum likelihood estimation equation is degenerate for very small numbers of observations and, hence, the function $\bar{\theta}$ may not be properly defined. For this purpose, a very small value of $n_{\min}$ is usually sufficient (for instance, $n_{\min} = 20$ is used in the simulations of Section 5.3).

3 Discussion

We first discuss connections of the proposed algorithm with earlier works before discussing its numerical complexity in more details.

Quite obviously, Algorithm 1 is intended to generalize the online EM algorithm of Cappé and Moulines (2009) to the case of dependent observations. In the case of independent observations, (8) and (9) reduces to a simpler recursion of the form:

$$\hat{S}_{n+1} = \gamma_{n+1} \mathrm{E}_{\bar{\theta}(\hat{S}_n)} \left[s(X_{n+1}, Y_{n+1}) | Y_{n+1} \right] + (1 - \gamma_{n+1}) \hat{S}_n \,, \tag{11}$$

which can be analyzed using the arguments developed for the analysis of stochastic approximation (or Robbins-Monro) schemes under Markovian random perturbations. The HMM framework however implies several key differences. First, in HMMs it is necessary to maintain an approximate filter $\hat{\phi}_n$ through (8) which, hopefully, becomes an acceptable proxy for $\phi_{n,\nu,\hat{\theta}_n}$ when approaching convergence. More importantly, as it is no more possible to compute the conditional expectation of the complete-data sufficient statistics online, we require the auxiliary recursion of (9). Although directly inspired by the exact recursive smoothing formula of (7), this auxiliary update cannot be put in the usual stochastic approximation form. And thus, in contrast to the online EM algorithm of Cappé and Moulines (2009), Algorithm 1 cannot be analyzed using off-the-shelf mathematical arguments. Both the need to maintain an estimate of the filter and the presence of the backward retrospective probability in (9) constitute significant departures from the usual stochastic approximation framework. Thus a complete analysis of the convergence of Algorithm 1 is a challenging mathematical problem and we provide in Section 4 below important —though limited— arguments in that direction.

Algorithm 1 is related to the EM-based approach suggested in Eq. (51) of Kantas et al. (2009) which however lacks the idea of a recursive implementation of smoothing (in this work, the authors propose to use a particle filtering method to approximate the smoothing functional). Chapters 4 and 5 of Ford (1998) and Elliott et al. (2002) consider the use of Proposition 1 for online estimation of the parameter of a Markov chain observed in additive noise —an example that we will discuss in more detail in Section 5— and for a linear state-space model with known noise characteristics. Ford (1998) however does not acknowledge the generality of the approach and the role of complete-data sufficient statistics and is constrained to the choice of $\gamma_n = n^{-1}$, which in simulations gives poor performances compared to alternative choices of step-sizes (see Section 5 below). To the best of our knowledge, the first instance of a particular case of Algorithm 1 appears in the work of Mongillo and Denève (2008) who considered the specific case of finite-valued observations. Mongillo and Denève (2008) however entirely rediscovered Proposition 1 under an equivalent form discussed below and thus failed to identify some of the general principles underpinning Algorithm 1.

3.1 Comparison with the Algorithm of Mongillo and Denève (2008)

This section is devoted to a more detailed analysis of the difference between Algorithm 1 applied to the case of finite-valued observations and the actual proposal of Mongillo and Denève (2008). This difference is not significant from a practical point of view but is important for the understanding of the behavior of the algorithm.

Mongillo and Denève (2008) considered the particular case of finite valued HMMs in which the observations $(Y_t)_{t \geq 1}$ also take their values in a finite set $\mathcal{Y}$. In such a situation, it is easily checked that for any parameterization of the model, the complete-data sufficient statistics may be chosen as $s(X_{t-1}, X_t, Y_t) = (\mathbb{1}\{X_{t-1} = i, X_t = j, Y_t = k\})_{(i,j,k) \in \mathcal{X}^2 \times \mathcal{Y}}$. The recursion derived by Mongillo and Denève (2008) for this case is based on recursively updating the product $\tau_{n,\nu,\theta}(x) = \phi_{n,\nu,\theta}(x)\rho_{n,\nu,\theta}(x)$ rather than $\rho_{n,\nu,\theta}(x)$. The probabilistic interpretation of the new term $\tau_{n,\nu,\theta}(x)$ is $\mathrm{E}_{\nu,\theta}\left[\left(\sum_{t=1}^n s(X_{t-1}, X_t, Y_t)\right)\mathbb{1}\{X_n = x\}|Y_{0:n}\right]$. By multiplying (9) by $\hat{\phi}_{n+1}(x)$ and using (8), one

obtains the following online update

$$
\begin{aligned}
\hat{\tau}_{n+1}(x) = \gamma_{n+1} \sum_{x' \in \mathcal{X}} s(x', x, Y_{n+1}) & \frac{\hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x', Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} \\
+ (1 - \gamma_{n+1}) \sum_{x' \in \mathcal{X}} \hat{\tau}_n(x') & \frac{q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x', Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} , \quad (12)
\end{aligned}
$$

which coincides with Eqs. (15)-(16) of Mongillo and Denève (2008), for the particular choice of complete-data sufficient statistics discussed above.

Of course, using either (8) and (9) or (12) is practically equivalent. But the results of Section 4 indicate that $\hat{\phi}_n$ and $\hat{\rho}_n$ have very different limiting behaviors: the auxiliary recursion (9) should converge to a fixed deterministic limit while the same is not true for the approximate filtering recursion (8). Hence the use of decreasing step-sizes for (9) only is justifiable while its use in (12) is less natural (and/or potentially misleading) as $\hat{\tau}_n$ should not be expected to converge to a deterministic limit.

Regarding the choice of the step-size, Mongillo and Denève (2008) consider the cases where, either, the step-size γ_n is small but non-decreasing, which may be useful for tracking potential changes but is not sufficient to guarantee the consistency of the approach. The other option mentioned by Mongillo and Denève (2008) is to use $\gamma_n = n^{-1}$ by analogy with the work of Neal and Hinton (1999) and the case of the batch EM algorithm. The range of step-sizes mentioned in Algorithm 1 is chosen in reference to the theory of stochastic approximation and by analogy with Cappé and Moulines (2009). As will be illustrated below, in the numerical simulations of Section 5, the choice of $\gamma_n = n^{-1}$ should definitely be avoided for HMMs and we instead recommend using step-sizes of the form $\gamma_n = n^{-\alpha}$ with α in the interval (0.5,0.8), possibly combined with Polyak-Ruppert averaging (see Section 5.3 below).

3.2 Implementation and Numerical Complexity

Regarding the numerical complexity of Algorithm 1, observe that in the case considered by Mongillo and Denève (2008) where $s(X_{t-1}, X_t, Y_t) = (\mathbb{1}\{X_{t-1} = i, X_t = j, Y_t = k\})_{(i,j,k) \in \mathcal{X}^2 \times \mathcal{Y}}$, $s(X_{t-1}, X_t, Y_t)$ is a vector of dimension $|\mathcal{X}|^2 \times |\mathcal{Y}|$ (where $|\cdot|$ denotes the cardinal of the set). Thus, the numerical complexity of (12) is of order $|\mathcal{X}|^4 \times |\mathcal{Y}|$ per observation. For this case, it is indeed possible to bring down the numerical complexity to the order of $|\mathcal{X}|^4 + |\mathcal{X}|^3 \times |\mathcal{Y}|$ operations by updating separately the terms corresponding to the two statistics $(\mathbb{1}\{X_{t-1} = i, X_t = j\})_{(i,j) \in \mathcal{X}^2}$ and $(\mathbb{1}\{X_t = j, Y_t = k\})_{(j,k) \in \mathcal{X} \times \mathcal{Y}}$ (see the example considered in Section 5 for more details). Interestingly, the numerical complexity of the batch EM algorithm for this model, when implemented using traditional forward-backward smoothing (Rabiner, 1989), is of the order of $(|\mathcal{X}|^2 + |\mathcal{X}| \times |\mathcal{Y}|)$ per observation and per iteration of the EM algorithm. The comparison is not directly meaningful as the batch EM algorithm does necessitate several iterations to converge (see numerical illustrations in Section 5.3). The scaling of the numerical complexity of the online-EM algorithm with respect to $|\mathcal{X}|$ can constitute an hindrance in models with a large number of states. This being said, the complexity of online gradient-based approaches, is equivalent as the main burden comes from the necessity of updating, via a recursion related to (9), one coordinate of the gradient for each of the pairs $(x, x') \in \mathcal{X}^2$ of state values (see, e.g., Le Gland and Mevel, 1997). When the transition matrix is structured —i.e., parametered by a low dimensional parameter rather than by all its individual entries—, the numerical cost of implementing the online EM approach is reduced to an order of the number of parameters times $|\mathcal{X}|^2$.

7

4 Some Results on Convergence

In this section, we provide two important elements regarding Algorithm 1. The first is a generalization of the argument of Cappé and Moulines (2009) that makes it possible to identify a deterministic limit for the EM update as the number of observations tends to infinity. Interestingly, the form of this limit is non trivial and substantially different from the case of independent observations. The second result pertains to the limiting behavior of the auxiliary quantity $\hat{\rho}_n(x)$ of (10) that is instrumental in Algorithm 1. As discussed in the previous section, a complete analysis of the convergence of Algorithm 1 is still lacking but it is possible to show that when the parameter is frozen (i.e. when the M-step update of (10) is inhibited), $\hat{\rho}_n$ converges to a deterministic quantity. This limit *does not depend on x* (or, in other words, $|\hat{\rho}_n(x) - \hat{\rho}_n(x')| \to 0$) and is related to the limiting EM mapping. This result, although limited, is very important to understand the nature of the asymptotic attractors of Algorithm 1. In this section, we will work under the assumptions of Douc et al. (2004) which guarantee the asymptotic normality of the MLE, adapted to the (simplest) case of a finite state space $\mathcal{X}$.

We start by a closer inspection of the limiting behavior of the normalized score function (gradient of the log-likelihood). Under suitable assumptions (see below), the normalized HMM log-likelihood $\frac{1}{n} \log \ell_{\nu,\theta}(Y_0, \ldots, Y_n)$ converges, $P_{\theta_\star}$ almost surely and in L^1, to the limiting contrast

$$c_{\theta_\star}(\theta) = E_{\theta_\star}[\log \ell_\theta(Y_0|Y_{-\infty:-1})] . \tag{13}$$

The same is true for the normalized score $\frac{1}{n} \nabla_\theta \log \ell_{\nu,\theta}(Y_0, \ldots, Y_n)$ which converges to $\nabla_\theta c_{\theta_\star}(\theta)$. Such consistency results have been established, under various assumptions, by (among others) Baum and Petrie (1966), Bickel et al. (1998), Douc et al. (2004). Now, thanks to Fisher identity, for all n,

$$\frac{1}{n} \nabla_\theta \log \ell_{\nu,\theta}(Y_0, \ldots, Y_n) = \frac{1}{n} E_{\nu,\theta} \left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1}) \,\middle|\, Y_{0:n} \right]$$
$$+ \frac{1}{n} E_{\nu,\theta} \left[\nabla_\theta \log p_{\nu,\theta}(X_0, Y_0) \,\middle|\, Y_{0:n} \right] . \tag{14}$$

As already discussed, the last term on the r.h.s., whose influence is vanishing with increasing values of n, can be ignored. Hence, the consistency result for the score function combined with (14) implies that $\frac{1}{n} E_{\nu,\theta} \left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1}) \,\middle|\, Y_{0:n} \right]$ converges $P_{\theta_\star}$ almost surely to $\nabla_\theta c_{\theta_\star}(\theta)$, the gradient of the limiting contrast. Now using the exponential family representation in (1), the non-vanishing term in the r.h.s. of (14) may rewritten as

$$\frac{1}{n} E_{\nu,\theta} \left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1}) \,\middle|\, Y_{0:n} \right] =$$
$$\nabla_\theta \psi(\theta) \left\{ \frac{1}{n} E_{\nu,\theta} \left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \,\middle|\, Y_{0:n} \right] \right\} - \nabla_\theta A(\theta) . \tag{15}$$

The following theorem defines the limiting behavior of the r.h.s. of the above equation, and thus, the limiting behavior of the EM update for HMMs (see Appendix A for the corresponding proofs).

Theorem 1. *In addition to Assumption 1, assume that* (i) *$\mathcal{X}$ is a finite set;* (ii) *the parameter space Θ is compact and $\theta_\star$ lies in the interior of Θ;* (iii) *the transition matrix is such that $q_\theta(x, x') \geq$*

$\epsilon > 0$ *for all* $\theta \in \Theta$; *(iv)* $\sup_\theta \sup_y \bar{g}_\theta(y) < \infty$ *and* $\mathrm{E}_{\theta_\star}[|\log \inf_\theta \bar{g}_\theta(Y_0)|] < \infty$, *where* $\bar{g}_\theta(y) = \sum_x g_\theta(x, y)$; *and, (v)* ψ_q, A_q, ψ_g, A_g *in* (2) *are continuously differentiable functions on the interior of* Θ. *Then the following properties hold.*

(i)

$$\frac{1}{n} \mathrm{E}_{\nu,\theta} \left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n} \right] \longrightarrow \mathrm{E}_{\theta_\star} \left(\mathrm{E}_\theta \left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty} \right] \right), \quad \mathrm{P}_{\theta_\star} \; a.s. \quad (16)$$

(ii) The fixed points of the limiting EM algorithm

$$\theta_{k+1} = \bar{\theta} \left\{ \mathrm{E}_{\theta_\star} \left(\mathrm{E}_{\theta_k} \left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty} \right] \right) \right\} \quad (17)$$

are the stationary points of the limiting likelihood contrast $c_{\theta_\star}(\theta)$.

Theorem 1 provides an interpretation of the limiting form of the classical EM algorithm when used on very long sequences of observations. In the case of HMMs, this form is quite complicated and it is interesting to compare it to the case of independent observations investigated by Cappé and Moulines (2009). If the observations are assumed to be independent, the law of large number implies the convergence of the normalized log-likelihood to $c_{\theta_\star}(\theta) = \mathrm{E}_{\theta_\star}[\log \ell_\theta(Y_0)]$. In this case, it is obvious that maximizing $c_{\theta_\star}(\theta)$ is equivalent to minimizing the Kullback-Leibler divergence $D(\ell_{\theta_\star}|\ell_\theta) = \int \log \frac{\ell_{\theta_\star}(y)}{\ell_\theta(y)} \ell_{\theta_\star}(y) dy$. In the HMM case, $D(\ell_{\theta_\star}|\ell_\theta)$ needs to be replaced by the expression of $c_{\theta_\star}(\theta)$ given in (13), which is a consequence of the tower property of conditional expectation and of the forgetting property of the filter (Douc et al., 2004). However, in contrast to the case of independent observations, it is no more straightforward to provide an explicit expression for the gradient of $c_{\theta_\star}(\theta)$. The key idea here is the use of Fisher identity in (14) which yields the limiting term found in the r.h.s. of (16). The use of Fisher identity also explains the particular form of conditioning found in (16), which involves both the infinite future and past of the trajectory. This expression also suggests that being able to compute or approximate smoothed functionals of the state recursively is indeed a key requirement for online estimation in HMMs.

The next result shows that under parameter freeze, the online EM update equation (8) converges to a deterministic constant function, equal to the r.h.s. of (16).

Corollary 1. *Under the assumptions of Theorem 1, Algorithm 1 used without* (10) *—that is, with* $\hat{\theta}_n$ *equal to a fixed value* θ *— satisfies*

$$\hat{\rho}_n(x) \longrightarrow \mathrm{E}_{\theta_\star} \left(\mathrm{E}_\theta \left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty} \right] \right), \quad \mathrm{P}_{\theta_\star} \; a.s., \; for \; all \; x \in \mathcal{X}.$$

Corollary 1 shows that under parameter freeze the auxiliary quantity $\hat{\rho}_n(x)$ converges to a constant limit that does not depend on x and that is equal to the limit obtained in (16). Note that since $\hat{\phi}_n(x)$ is a probability on $\mathcal{X}$, this implies that $\bar{\theta}(\sum_{x \in \mathcal{X}} \hat{\rho}_{n+1}(x)\hat{\phi}_{n+1}(x))$ tends to $\bar{\theta}\{\mathrm{E}_{\theta_\star}(\mathrm{E}_\theta[s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}])\}$. This result together with Theorem 1–(ii) suggests that Algorithm 1 can only be stable at the stationary points of the limiting contrast $c_{\theta_\star}(\theta)$. The argument is only heuristic at this stage as Corollary 1 is obtained under the artificial assumption that the evolution of the parameter is frozen. Corollary 1 however highlights an essential characteristic of Algorithm 1: in contrast to $\hat{\phi}_n(x)$ which does approximate the infinite past filter and hence varies with each observation, $\hat{\rho}_n(x)$ converges to a deterministic limit that is independent of x. This property, which can be verified in simulations, justifies the form of Algorithm 1 and in particular the use of a stochastic approximation type of update for $\hat{\rho}_n(x)$ only.

5 Application to Gaussian HMMs

5.1 HMM with Product Parameterization

For the sake of concreteness, we consider in following the case where the state variables (X_t) take their values in the set $\{1, \ldots, m\}$. In addition, assume that, as is often the case in practise, the parameter θ may be split into two sub-components that correspond, respectively, to the state transition matrix q_θ and to the state-conditional densities $\{g_\theta(i, \cdot)\}_{1 \leq i \leq m}$. In the fully discrete case considered in Mongillo and Denève (2008) for instance, the parameter θ consist of the transition matrices q_θ and g_θ parametered by their respective entries, with the constraint that each line of a transition matrix must sum to one. In the case of Gaussian HMMs used in speech processing as well as many in other applications, the parameters are the state transition matrix q_θ and the mean vector and covariance matrix associated with each of the m state-conditional densities $\{g_\theta(i, \cdot)\}_{1 \leq i \leq m}$ (Rabiner, 1989).

In such a model, there are two distinct types of EM complete-data sufficient statistics which give rise to two separate forms of the auxiliary function $\rho_{n,\nu,\theta}$:

$$\rho_{n,\nu,\theta}^q(i, j, k; \theta) = \frac{1}{n} \mathrm{E}_{\nu,\theta} \left[\sum_{t=1}^{n} \mathbb{1}\{X_{t-1} = i, X_t = j\} \middle| Y_{0:n}, X_n = k \right] , \tag{18}$$

$$\rho_{n,\nu,\theta}^g(i, k; \theta) = \frac{1}{n} \mathrm{E}_{\nu,\theta} \left[\sum_{t=0}^{n} \mathbb{1}\{X_t = i\} s(Y_t) \middle| Y_{0:n}, X_n = k \right] , \tag{19}$$

where the form of s itself depend on the nature of the state-conditional distribution $g_\theta(x, \cdot)$ —see Gaussian example below. There's a slight difference between (19) and (5), which is that (19) also incorporates the initial ($t = 0$) conditional likelihood term, i.e., the contribution corresponding to the rightmost term on the r.h.s. of (14). As noted earlier, this difference is minor and does not modify the long-term behavior of the algorithm.

With these notations, Eq. (9) in Algorithm 1 is implemented as

$$\hat{\rho}_{n+1}^q(i, j, k) = \gamma_{n+1} \delta(j - k) \hat{r}_{n+1}(i|j) + (1 - \gamma_{n+1}) \sum_{k'=1}^{m} \hat{\rho}_n^q(i, j, k') \hat{r}_{n+1}(k'|k) , \tag{20}$$

$$\hat{\rho}_{n+1}^g(i, k) = \gamma_{n+1} \delta(i - k) s(Y_{n+1}) + (1 - \gamma_{n+1}) \sum_{k'=1}^{m} \hat{\rho}_n^g(i, k') \hat{r}_{n+1}(k'|k) , \tag{21}$$

where δ denotes the Kronecker delta (i.e., $\delta(i) = 0$ iff $i = 0$) and the notation $\hat{r}_{n+1}(i|j)$ refers to the approximate retrospective conditional probability :

$$\hat{r}_{n+1}(i|j) = \frac{\hat{\phi}_n(i) q_{\hat{\theta}_n}(i, j)}{\sum_{i'=1}^{m} \hat{\phi}_n(i') q_{\hat{\theta}_n}(i', j)} . \tag{22}$$

A complete iteration of the online algorithm involves the approximate filter update (8) and the statistics updates (20) and (21) followed by an application of the M-step function $\bar{\theta}$ to $\hat{S}_{n+1}^q(i, j) = \sum_{k=1}^{m} \hat{\rho}_{n+1}^q(i, j, k) \hat{\phi}_{n+1}(k)$ and $\hat{S}_{n+1}^g(i) = \sum_{k=1}^{m} \hat{\rho}_{n+1}^g(i, k) \hat{\phi}_{n+1}(k)$. The form of the M-step depends on the exact nature of q_θ and g_θ. If the transition matrix q_θ is parametered simply by its entries, the update is generic and is given by

$$q_{\hat{\theta}_n}(i, j) = \bar{\theta}\left(\hat{S}_n^q \right) = \frac{\hat{S}_n^q(i, j)}{\sum_{j=1}^{m} \hat{S}_n^q(i, j)} . \tag{23}$$

For the update of the state-dependent parameters, one needs to be more specific and the form of the equations depend on the choice of the state conditional density $g_\theta(x, \cdot)$. In the multivariate Gaussian case, the function s has to be chosen such that $s(y)$ consists of the three components $\{1, y, yy^{\mathrm{t}}\}$. The corresponding components of the approximated EM extended statistics are denoted, respectively, by $\hat{S}^g_{n,0}, \hat{S}^g_{n,1}, \hat{S}^g_{n,2}$. If the state conditional Gaussian densities are parametered by their mean vectors, $\mu_\theta(i)$, and covariances matrices, $\Sigma_\theta(i)$, the M-step update is defined as

$$\mu_{\hat{\theta}_n}(i) = \bar{\theta}\left(\hat{S}^g_{n,0}, \hat{S}^g_{n,1}\right) = \frac{\hat{S}^g_{n,1}(i)}{\hat{S}^g_{n,0}(i)} \,, \tag{24}$$

$$\Sigma_{\hat{\theta}_n}(i) = \bar{\theta}\left(\hat{S}^g_{n,0}, \hat{S}^g_{n,1}, \hat{S}^g_{n,2}\right) = \frac{\hat{S}^g_{n,2}(i)}{\hat{S}^g_{n,0}(i)} - \mu_{\hat{\theta}_n}(i)\mu^{\mathrm{t}}_{\hat{\theta}_n}(i) \,. \tag{25}$$

The derivation of (23) and (24)–(25) is straightforward but some more details are provided in the next section for a particular case of Gaussian HMM.

5.2 Markov Chain Observed in Gaussian Noise

In the numerical experiments described below, we consider the simple scalar model

$$Y_t = X_t + V_t \,,$$

where (V_t) is a scalar additive Gaussian noise of variance v and (X_t) is a Markov chain with transition matrix q, which takes its values in the set $\{\mu(1), \ldots, \mu(m)\}$. Although simple, this model is already statistically challenging and is of some importance in several applications, in particular, as a basic model for ion channels data (Chung et al., 1990) —see also, e.g., (Roberts and Ephraim, 2008) for discussion of the continuous time version of the model as well de Gunst et al. (2001), for an up-to-date account of models for ion channels. The parameter θ is comprised of the transition matrix q, the vector of means μ and the variance v.

In this case, the intermediate quantity of the batch EM algorithm may be written —with constant terms omitted, as

$$\sum_{i=1}^{m}\sum_{j=1}^{m} S^q_n(i,j) \log q(i,j) - \frac{1}{2v}\sum_{i=1}^{m}\left(S^g_{n,2}(i;\theta) - 2\mu(i)S^g_{n,1}(i) + \mu^2(i)S^g_{n,0}(i)\right) \,, \tag{26}$$

where

$$S^q_n(i,j) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n}\mathbb{1}\{X_{t-1} = i, X_t = j\}\,\middle|\, Y_{0:n}\right] \,,$$

$$S^g_{n,0}(i) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t = i\}\,\middle|\, Y_{0:n}\right] \,,$$

$$S^g_{n,1}(i) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t = i\}Y_t\,\middle|\, Y_{0:n}\right] \,,$$

$$S^g_{n,2}(i) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t = i\}Y_t^2\,\middle|\, Y_{0:n}\right] \,.$$

Maximization of (26) with respect to q_θ, μ_θ and υ_θ directly yields (23) as well as

$$\mu(i) = \bar{\theta}\left(S^g_{n,0}, S^g_{n,1}\right) = \frac{S^g_{n,1}(i)}{S^g_{n,0}(i)} \,, \tag{27}$$

$$\upsilon = \bar{\theta}\left(S^g_{n,0}, S^g_{n,1}, S^g_{n,2}\right) = \frac{\sum_{i=1}^m \left(S^g_{n,2}(i) - \mu^2(i)S^g_{n,0}(i)\right)}{\sum_{i=1}^m S^g_{n,0}(i)} \,. \tag{28}$$

It is easily checked that, as usual, the M-step equations (23) and (28) satisfy the constraints that q be a stochastic matrix and υ be non-negative. Note that for this particular model, the use of the statistic $S^g_{n,2}$ could be avoided as it is only needed in the M-step under the form $\sum_{i=1}^m S^g_{n,2}(i)$, which is equal to $\frac{1}{n}\sum_{t=0}^n Y_t^2$. Algorithm 2 below recaps the complete online EM algorithm pertaining to this example.

Algorithm 2 (Online EM algorithm for noisily observed m-state Markov chain)**.**

Initialization *Select $\hat{\theta}_0$ and compute, for all $1 \le i, j, k, \le m$ and $0 \le d \le 2$,*

$$\hat{\phi}_0(k) = \frac{\nu(k)g_{\hat{\theta}_0}(k, Y_0)}{\sum_{k'=1}^m g_{\hat{\theta}_0}(k, Y_0)} \,,$$

$$\hat{\rho}^q_0(i, j, k) = 0 \,,$$

$$\hat{\rho}^g_{0,d}(i, k) = \delta(i - k)Y_0^d \,.$$

Recursion *For $n \ge 0$, and $1 \le i, j, k, \le m$, $0 \le d \le 2$,*

> **Approx. Filter Update**
>
> $$\hat{\phi}_{n+1}(k) = \frac{\sum_{k'=1}^m \hat{\phi}_n(k')\hat{q}_n(k', k)g_{\hat{\theta}_n}(k, Y_{n+1})}{\sum_{k',k''=1}^m \hat{\phi}_n(k')\hat{q}_n(k', k'')g_{\hat{\theta}_n}(k'', Y_{n+1})} \,,$$
>
> *where $g_{\hat{\theta}_n}(k, y) = \exp\left[-(y - \hat{\mu}_n(k))^2/2\hat{\upsilon}_n\right]$.*
>
> **Stochastic Approximation E-step**
>
> $$\hat{\rho}^q_{n+1}(i, j, k) = \gamma_{n+1}\delta(j - k)\hat{r}_{n+1}(i|j) + (1 - \gamma_{n+1})\sum_{k'=1}^m \hat{\rho}^q_n(i, j, k')\hat{r}_{n+1}(k'|k) \,,$$
>
> $$\hat{\rho}^g_{n+1,d}(i, k) = \gamma_{n+1}\delta(i - k)Y^d_{n+1} + (1 - \gamma_{n+1})\sum_{k'=1}^m \hat{\rho}^g_{n,d}(i, k')\hat{r}_{n+1}(k'|k) \,,$$
>
> *where $\hat{r}_{n+1}(i|j) = \hat{\phi}_n(i)\hat{q}_n(i, j) \big/ \sum_{i'=1}^m \hat{\phi}_n(i')\hat{q}_n(i', j)$.*

M-step *If $n \geq n_{\min}$,*

$$\hat{S}_{n+1}^q(i,j) = \sum_{k'=1}^m \hat{\rho}_{n+1}^q(i,j,k')\hat{\phi}_{n+1}(k') \, ,$$

$$\hat{q}_{n+1}(i,j) = \frac{\hat{S}_{n+1}^q(i,j)}{\sum_{j'=1}^m \hat{S}_{n+1}^q(i,j')} \, ,$$

$$\hat{S}_{n+1,d}^g(i) = \sum_{k'=1}^m \hat{\rho}_{n+1,d}^g(i,k')\hat{\phi}_{n+1}(k') \, ,$$

$$\hat{\mu}_{n+1}(i) = \frac{S_{n+1,1}^g(i)}{S_{n+1,0}^g(i)} \, ,$$

$$\hat{v}_{n+1} = \frac{\sum_{i'=1}^m \left(S_{n+1,2}^g(i') - \hat{\mu}_{n+1}^2(i')S_{n+1,0}^g(i') \right)}{\sum_{i'=1}^m S_{n+1,0}^g(i')} \, .$$

5.3   Numerical Experiments

Algorithm 2 is considered in the case of a two-state ($m = 2$) model estimated from trajectories simulated from the model with parameters

$$q_\star(1,1) = 0.95 \, , \ \mu_\star(1) = 0 \, ,$$
$$q_\star(2,2) = 0.7 \, , \ \mu_\star(2) = 1 \, ,$$
$$v_\star = 0.5 \, .$$

With these parameters, state identification is a difficult task as the separation of the means corresponding to the two states is only 1.4 times the noise standard deviation. The optimal filter associated with the actual parameter does for instance misclassify the state (using Bayes' rule) in about 10.3% of the cases. As will be seen below, this is reflected in slow convergence of the EM algorithm.

All estimation algorithms are systematically started from the initial values

$$q_\star(1,1) = 0.7 \, , \ \mu_\star(1) = -0.5 \, ,$$
$$q_\star(2,2) = 0.5 \, , \ \mu_\star(2) = 0.5 \, ,$$
$$v_\star = 2 \, ,$$

and run on 100 independent trajectories simulated from the model.

Figure 1 illustrates the consequences of Theorem 1 by plotting the estimates of the parameters $q(1,1)$ and $\mu(1)$ obtained by the batch EM algorithm, as a function of the number of EM iterations, for two different sample sizes: $n = 500$ (dotted lines) and $n = 8000$ iterations. To give an idea of the variability of the estimates, Figure 1 feature the median estimate (bold line) as well as the lower and upper quartiles (lighter curves) for both sample sizes. The first striking observation is the slow convergence of EM in this case, which requires about 50 iterations or so to reach decent estimates of the parameters. When comparing the curves corresponding to the two samples sizes, it is also obvious that while the variability is greatly reduced for $n = 8000$ compared to $n = 500$, the median learning curve is very similar in both cases. Furthermore, the plots corresponding to $n = 8000$ provide a very clear picture of the deterministic limiting EM trajectory, whose existence is guaranteed by Theorem 1.
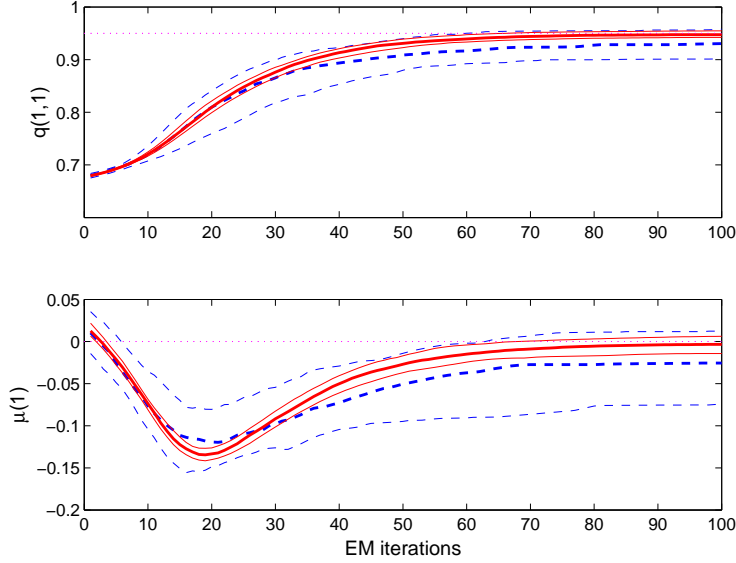
13

Figure 1: Estimated values of $q(1,1)$ (top) and $\mu(1)$ (bottom) as a function of the number of batch EM iterations for $n = 500$ (dotted lines) and $n = 8000$ (solid lines) observations. The plot is based on 100 independent runs summarized by the median (bold central line) and the upper and lower quartiles (lighter lines).

Indeed, the large sample behavior of the batch EM algorithm is rather disappointing as using a fixed number of iteration of EM does involve a computational cost that grows proportionally to n but will converge, as n grows, to a deterministic limit which only depends on the parameter initialization. This is all the more regrettable that from a statistical perspective, it is expected that the true maximum likelihood estimator converges, at rate $n^{-1/2}$ towards the actual value $\theta_\star$ of the parameter. This behavior of the batch EM algorithm is illustrated on Figure 2 which displays, from left to right, the estimation results for the parameters associated with the first component $(q(1,1), \mu(1))$, as in Figure 1, together with the noise variance υ (rightmost display). The results obtained on the 100 independent runs are here summarized as box and whiskers plots. Figure 2, which should be compared with Figure 3 below, shows that when using a potentially large (here, 50) but fixed number of iterations the variability of the batch EM estimates does decrease but the accuracy does not improve as n grows. Clearly, statistical consistency could only be achieved by using more batch EM iterations as the number n of observations grows.

In contrast, Figure 3 which corresponds to the online EM algorithm (Algorithm 2) used with $\gamma_n = n^{-0.6}$ does suggest that online EM estimation is consistent. For the smallest sample sizes ($n = 500$ or $n = 2000$), the estimation variance is still quite large and the online estimates are not as good as those obtained using 50 batch EM iterations. But for sample sizes of $n = 8000$ and larger, the online EM estimates are preferable despite their somewhat larger variance. In this application, the choice of a slowly decreasing step-size appears to be of utmost importance. In particular, the choice $\gamma_n = n^{-1}$, despite its strong analogy with the batch EM case (see Section 2 as well as Neal and Hinton, 1999) provides estimates that are not robust enough with respect to the choice of the initial parameter guess. Simple step-size schemes of the form $\gamma_n = n^{-\alpha}$ perform very poorly in this example when α is set to values higher than 0.8. Figure 4 features a more complex choice of step-sizes which was hand-tuned based on pilot runs. Even with this preliminary phase of step-size tuning —which would hardly be feasible in real-life applications of the method—
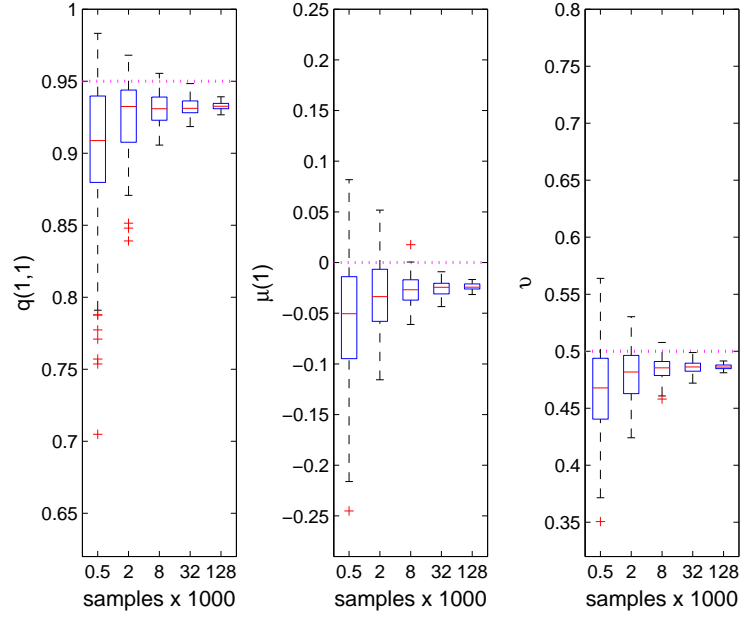
14

Figure 2: Estimation results when using 50 batch EM iterations. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. Box and whiskers plot based on 100 independent runs.
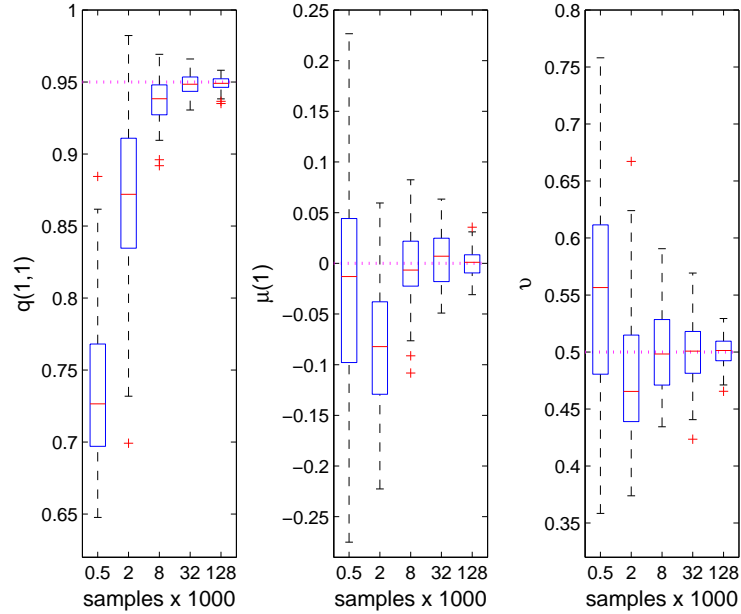

Figure 3: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. Box and whiskers plot based on 100 independent runs.
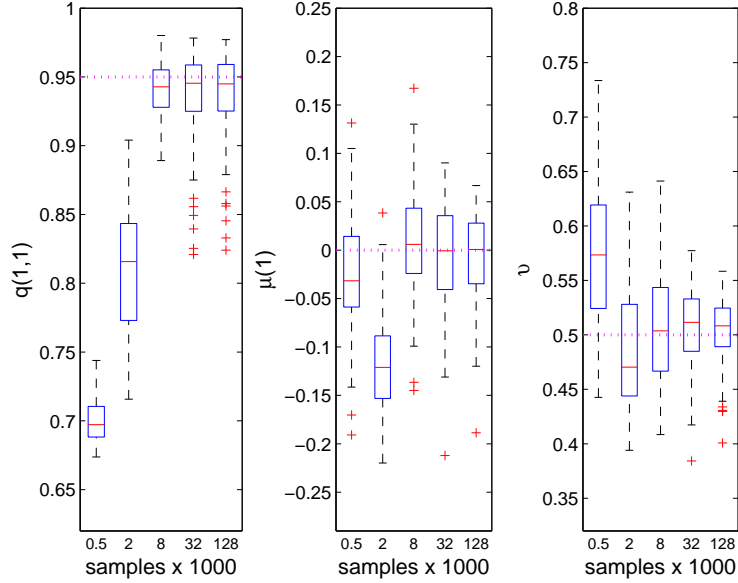
Figure 4: Estimation results when using the online EM algorithm with $\gamma_n = 0.01$ for $n \le n_0$ and $\gamma_n = 0.5(n - n_0)^{-1}$ for values of $n > n_0$, with $n_0 = 10000$. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. Box and whiskers plot based on 100 independent runs.

it is observed on Figure 4 that the results are not significantly improved compared to the simple choice of $\gamma_n = n^{-0.6}$ as in Figure 3. The comparatively higher variability observed for the largest sample sizes (32 and 128 thousands) on Figure 4 is caused by a very slow reduction of the bias of the parameter estimates when $\gamma_n = n^{-1}$. In contrast, parameter trajectories corresponding to the choice of $\gamma_n = n^{-0.6}$ look much less smooth but provide more reliable estimates, without requiring model-specific tuning of the step-sizes. This observation —which is also true, but to a lesser extent, in the case of independent observations— is certainly a consequence of the temporal dependence between the observations and, correlatively, of the time needed for the filtering and smoothing relations to forget their initial state.

Method	MATLAB 7.7	OCTAVE 3.0
Online EM	1.57	5.66
Batch EM (one iteration, recursive)	1.24	3.98
Batch EM (one iteration, forward-backward)	0.31	2.94

Table 1: Computing times in seconds for a record of length $n = 10000$ observations (Intel Core 2 E6600 2.4 GHz processor).

The comparison between Figures 2 and 3 is not meant to be fair in terms of computing time, as shown by Table 1[1]. The 50 batch EM iterations used to produce Figure 2 take about 10 to 40, depending on the implementation of batch EM, times longer than for the corresponding online estimates of Figure 3. Being fair in this respect would have mean using just five batch EM

---

[1]The MATLAB/OCTAVE code used for the simulations is very simple and is available as supplementary material. It is mostly vectorized, except for a loop through the observations, and hence it is expected that the differences in running times are indeed representative, despite the use of an interpreted programming language.

iterations which, as can be guessed from Figure 1, is not competitive with online EM, even for the smallest sample sizes. Note that a different option would have been to also consider running the online EM algorithm several times on the same batch data, following Neal and Hinton (1999). In the case of hidden Markov models however, this way of using the online EM algorithm for fixed-sample maximum likelihood estimation of the parameters appears to be less straightforward than in the case of i.i.d. data and has not been considered.

The two batch EM implementations featured in Table 1 correspond, respectively, to the use of the recursive form of smoothing based on Proposition 1 and to the usual forward-backward form of smoothing. The former implementation is obviously related to the online EM algorithm, which explains that both of them lead to rather similar running times. As discussed in Section 3, due to the fact that the whole $m \times m$ transition matrix q is here used as a parameter, the numerical complexity of the online EM algorithm and of the recursive implementation of batch EM scale as m^4, compared to m^2 only for the batch EM algorithm when implemented with forward-backward smoothing. Hence, it is to be expected that the forward-backward implementation of batch EM would be even more advisable for models with more than $m = 2$ states. On the other hand, when m is large it is usually not reasonable to parameterize the transition matrix q by its individual entries.

In order to provide a more detailed idea of the asymptotic performance of the algorithm, Figure 5 displays results similar to those of Figure 3 but centered and scaled as follows. Each parameter estimates, say $\hat{\theta}_n$ is represented as $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ and further scaled by the asymptotic standard deviation of θ deduced from the inverse of the Fisher information matrix. The Fisher information matrix has been estimated numerically by applying Fisher's identity to (26) so as to obtain

$$\frac{1}{n} \nabla_{q(i,j)} \log \ell_\theta(Y_{0:n}) = \frac{S_n^q(i,j)}{q(i,j)} - \frac{S_n^q(i,m)}{q(i,m)} \quad \text{(for } 1 \le j < m) \,,$$

$$\frac{1}{n} \nabla_{\mu(i)} \log \ell_\theta(Y_{0:n}) = \frac{S_{n,1}^g - \mu(i) S_{n,0}^g}{v} \,,$$

$$\frac{1}{n} \nabla_v \log \ell_\theta(Y_{0:n}) = \sum_{i=1}^m \frac{1}{2v^2} \left(S_{n,2}^g(i;\theta) - 2\mu(i) S_{n,1}^g(i) + \mu^2(i) S_{n,0}^g(i) \right) \,.$$

The information matrix has then been estimated by averaging the gradient computed in $\theta_\star$ for 100 independent sequences of length one million simulated under $\theta_\star$.

Additionally, Figure 5 also displays results that have been post-processed using Polyak-Ruppert averaging (Polyak, 1990, Ruppert, 1988). In Figure 5, Polyak-Ruppert averaging is used starting from $n_{\text{avg}} = 8000$. That is, for $n > 8000$, $\hat{\theta}_n$ is replaced by $1/(n-8000) \sum_{i=8001}^n \hat{\theta}_n$. For time indices n smaller than 8000, averaging is not performed and the estimates are thus as in Figure 3, except for the centering and the scaling. Under relatively mild assumptions, averaging has been shown to improve the asymptotic rate of convergence of stochastic approximation algorithm making it possible to recover the optimal rate of convergence of $n^{-1/2}$. At least for $\mu(1)$ and v, Figure 5 suggests that in this example the proposed algorithm does reach asymptotic efficiency, i.e., becomes asymptotically equivalent to the maximum likelihood estimator. For $q(1,1)$ the picture is less clear as the recentered and scaled estimates present a negative bias which disappears quite slowly. This effect is however typical of the practical trade-off involved in the choice of the index n_{avg} where averaging is started. To allow for a significant variance reduction, n_{avg} should not be too large. On the other hand, if averaging is started too early, forgetting of the initial guess of the parameters occurs quite slowly. In the present case, the negative bias visible on the left panel of Figure 5 is due to n_{avg} being too small (see corresponding panel in Figure 3). Although, this could be corrected
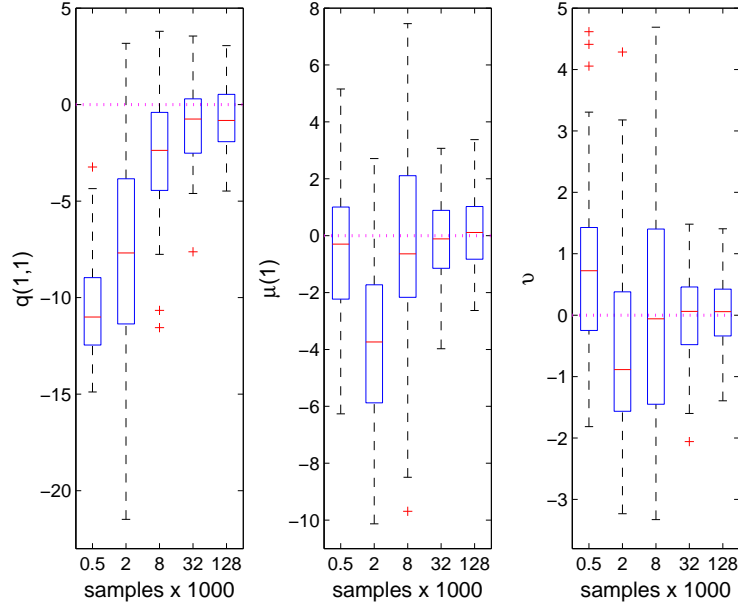
Figure 5: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$ with Polyak-Ruppert averaging started after $n = 8000$. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. The estimated values are centered and scaled so as to be comparable with a unitary asymptotic standard deviation. Box and whiskers plot based on 100 independent runs.

here by setting n_{avg} to twenty thousands or more, it is important to underline that optimally setting n_{avg} is usually not feasible in practice.

6   Conclusions

The algorithm proposed in this paper for online estimation of HMM parameters is based on two ideas. The first, which is inspired by Sato (2000) and Cappé and Moulines (2009), consists in reparameterizing the model in the space of sufficient statistics and approximating the limiting EM recursion by a procedure resembling stochastic approximation. Theorem 1 provides a first argument demonstrating that this idea can also be fruitful in the case of HMMs. The second element is more specific to HMMs and relies on the recursive implementation of smoothing computations for sum functionals of the hidden state which is provided by Proposition 1. As discussed in Section 2, this possibility requires that the auxiliary quantity $\rho_{n,\nu,\theta}$ defined in (5) be approximated during the course of the algorithm.

Although the performance reported in Section 5 is encouraging, there are several questions raised by this approach. The first is of course the theoretical analysis of the convergence of Algorithm 1, which is still missing. Although originally inspired by stochastic approximation ideas, it seems that Algorithm 1 would be difficult to analyze using currently available stochastic approximation results due to the backward kernel operator $\hat{r}_{n+1}(x'|x)$ involved in (9). As discussed in Sections 3 and 5.3, the proposed algorithm may become less attractive, from a computational point of view, when used in models with many distinct state values. In such cases and, more generally, in cases where the state-space $\mathcal{X}$ of the hidden chain is no longer finite, a promising

approach consists in using some form of Monte Carlo computation to approximate (8) and (9). Cappé (2009) and Del Moral et al. (2009) report encouraging first results in that direction.

A Proofs

Theorem 1 mainly relies on the use of a two-sided forgetting result which is first proved in Corollary 2 below. This result generalizes the one-sided forgetting bounds of Douc et al. (2004) and allows conditioning with respect to both past and future observations, which is required for studying the asymptotic behavior of (14) and related quantities. The proof of Theorem 1 then mostly relies on the results of Douc et al. (2004).

Lemma 1. *Given q a transition matrix on the finite set $\mathcal{X}$ such that $q(x, x') \geq \epsilon > 0$ and α and β probabilities on $\mathcal{X}$, define*

$$J_{\alpha,q,\beta}(x, x') = \frac{\alpha(x)q(x, x')\beta(x')}{\sum_{x,x'\in\mathcal{X}^2} \alpha(x)q(x, x')\beta(x')} \ .$$

Then

$$\|J_{\alpha_1,q,\beta_1} - J_{\alpha_2,q,\beta_2}\|_1 \leq \frac{1}{\epsilon} \left(\|\alpha_1 - \alpha_2\|_1 + \|\beta_1 - \beta_2\|_1 \right) \ , \tag{29}$$

where $\|\mu\|_1 = \sum_x |\mu(x)|$ denotes the L^1 or total variation norm.

Proof. Lemma 1 is obviously related to the application of Bayes' formula. Hence, one may apply Lemma 3.6 of Künsch (2001) to obtain $\|J_{\alpha_1,q,\beta_1} - J_{\alpha_2,q,\beta_2}\|_1 \leq \frac{1}{\epsilon}\|\alpha_1 \otimes \beta_1 - \alpha_2 \otimes \beta_2\|_1$. The r.h.s. of (29) is obtained by noting that $|\alpha_1(x)\beta_1(x') - \alpha_2(x)\beta_2(x')| \leq |\alpha_1(x) - \alpha_2(x)|\beta_1(x') + |\beta_1(x') - \beta_2(x')|\alpha_2(x)$. $\square$

Corollary 2. *Under the assumptions of Theorem 1, for any function f such that $0 < f < \|f\|_\infty$ and probabilities μ_1 and μ_2 on $\mathcal{X}^2$, and any index $1 \leq t \leq n$,*

$$\left| \sum_{x,x'\in\mathcal{X}^2} \mathrm{E}_\theta \left[f(X_{t-1}, X_t)| Y_{0:n}, X_0 = x, X_n = x' \right] \mu_1(x, x') \right.$$
$$\left. - \sum_{x,x'\in\mathcal{X}^2} \mathrm{E}_\theta \left[f(X_{t-1}, X_t)| Y_{0:n}, X_0 = x, X_n = x \right] \mu_2(x, x') \right| \leq \frac{\|f\|_\infty}{\epsilon} \left(\rho^{t-1} + \rho^{n-t} \right) \ ,$$

where $\rho = (1 - \epsilon)$.

Proof. First apply Lemma 1 to the familiar forward-backward decomposition

$$\alpha_i(x) = \mathrm{P}_\theta(X_{t-1} = x|Y_{0:t-1}, X_0 = x_{0,i}) \ ,$$
$$\beta_i(x') \propto \mathrm{P}_\theta(Y_{t:n}, X_n = x_{n,i}|X_t = x') \ ,$$

for $i = 1, 2$ (where the normalization factor in the second equation is determined by the constraint $\sum_{x\in\mathcal{X}} \beta_i(x) = 1$) to obtain

$$\left| \mathrm{E}_\theta \left[f(X_{t-1}, X_t)| Y_{0:n}, X_0 = x_{0,1}, X_n = x_{n,1} \right] \right.$$
$$\left. - \mathrm{E}_\theta \left[f(X_{t-1}, X_t)| Y_{0:n}, X_0 = x_{0,2}, X_n = x_{n,2} \right] \right| \leq \frac{\|f\|_\infty}{\epsilon} \left(\|\alpha_1 - \alpha_2\|_1 + \|\beta_1 - \beta_2\|_1 \right) \ ,$$

19

observing that $P_\theta(X_{t-1} = x, X_t = x'|Y_{0:t-1}, X_0 = x_{0,i}, X_n = x_{n,i}) = J_{\alpha_i, q_\theta, \beta_i}$. Next use, the one-sided forgetting bounds of Douc et al. (2004) (Corollary 1 and Eq. (20)) to obtain $\|\alpha_1 - \alpha_2\|_1 \leq \rho^{t-1}$ and $\|\beta_1 - \beta_2\|_1 \leq \rho^{n-t}$. The result of Corollary 2 follow by the general inequality $|\mu_1(g) - \mu_2(g)| \leq \frac{1}{2}\|\mu_1 - \mu_2\|_1 \sup_{z_1, z_2 \in \mathcal{Z}^2} |g(z_1) - g(z_2)|$. $\qquad\square$

Note that the fact that the backward function $\beta_i(x') = P_\theta(Y_{t:n}, X_n = x_{n,i}|X_t = x')$ may be normalized to a pseudo-probability, has been used in the above proof. This is generally not the case outside of the context where $\mathcal{X}$ is finite (Cappé et al., 2005, Briers et al., 2010) but it is easily checked that Corollary 2 holds in greater generality under the "strong mixing conditions" discussed in Section 4.3 of Cappé et al. (2005).

Proof of Theorem 1. Corollary 2 implies that

$$|\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-n:n}\right] - \mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-m:m}\right]| \leq \frac{2}{\epsilon}\rho^n M(Y_0),$$

for $m \geq n$, where $M(y) \geq \sup_{x,x'} |s(x,x',y)|$. $M(y)$ may be chosen as $M(y) = \sum_x 1 + |s^g(x,y)|$ due to the fact that $s^q(x,x')$ is a vector of indicator functions when $\mathcal{X}$ is finite. As, $\theta_\star$ lies in the interior of Θ, standard results on exponential family imply that $M(Y_0)$ has finite first and second order moments under $P_{\theta_\star}$. Hence, the a.s. limit of $\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-m:m}\right]$ as $m \to \infty$, which is denoted by $\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}\right]$, exists and has finite expectation under $P_{\theta_\star}$. Similarly,

$$\frac{1}{n}\sum_{t=1}^n \left(\mathrm{E}_{\nu,\theta}\left[s(X_{t-1}, X_t, Y_t)|Y_{0:n}\right] - \mathrm{E}_\theta\left[s(X_{t-1}, X_t, Y_t)|Y_{-\infty:\infty}\right]\right) \leq \frac{1}{n\epsilon}\sum_{t=1}^n \left(\rho^{t-1} + \rho^{n-t}\right)M(Y_t).$$

As $\mathrm{E}_{\theta_\star}[M(Y_0)^2] < \infty$, standard applications of Markov inequality and Borel-Cantelli Lemma imply that the r.h.s. of the above expression tends $P_{\theta_\star}$–a.s. to zero. Hence, the quantities $\frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^n s(X_{t-1}, X_t, Y_t)|Y_{0:n}\right]$ and $\frac{1}{n}\sum_{t=1}^n \mathrm{E}_\theta\left[\sum_{t=1}^n s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}\right]$ have the same limit, where the latter expression converges to $\mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}\right]\right)$ by the ergodic theorem. This proves the first assertion of Theorem 1.

For the second statement, one can check that the assumptions of Theorem 1 imply (A1)–(A3) of Douc et al. (2004) as well as a form of (A6)–(A8)[2]. Hence, proceeding as in proof of Theorem 3 of Douc et al. (2004), shows that (14) converge a.s. to the gradient $\nabla_\theta c_{\theta_\star}(\theta)$ of the limiting contrast defined in (13). Eq. (15) combined with the previous result then shows that parameter values θ for which $\nabla_\theta c_{\theta_\star}(\theta)$ vanishes are also such that $\nabla_\theta \psi(\theta)\left\{\mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}\right]\right)\right\} - \nabla_\theta A(\theta) = 0$, that is, $\bar{\theta}\left\{\mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0)|Y_{-\infty:\infty}\right]\right)\right\} = \theta$. $\qquad\square$

In order to prove Corollary 1, one first needs the following simple lemma.

Lemma 2. *Let $(e_k)_{k\geq 0}$ denote a deterministic sequence and $(\gamma_k)_{k\geq 1}$ a sequence of step-sizes such that $\gamma_k \in [0,1]$, $\sum_{k\geq 1} \gamma_k = \infty$, and, $\sum_{k\geq 1} \gamma_k^2 < \infty$; define $r_k = (1 - \gamma_k)r_{k-1} + \gamma_k e_k$ for $k \geq 1$. Then*

(i) $r_n = \sum_{k=1}^n \omega_k^n e_k + \omega_0^n r_0$, where $\omega_k^n = \gamma_k \prod_{j=1}^n (1 - \gamma_j)$.

(ii) r_n and $\frac{1}{n}\sum_{k=1}^n e_k$ have the same limit.

---

[2]Theorem 3 of Douc et al. (2004) deals with the Hessian of the normalized log-likelihood. As we are only concerned with the gradient here, one can drop the second order conditions in (A6)–(A7). Furthermore, as the assumption of Theorem 1 are supposed to hold uniformly on Θ, the set G can be dropped in (A6)–(A7), which provides a law of large number for the score that holds for all values of $\theta \in \Theta$.

Lemma 2-(ii) can be proved using Abel transformation noticing that $\sum_{k=1}^{n} \omega_k^n$ is by definition equal to one and that for the range of step-sizes considered here, $(\omega_k^n)_{1 \leq k \leq n}$ is a non-decreasing sequence (it is equal to n^{-1} when $\gamma_k = k^{-1}$ and strictly increasing for sequences of step-sizes that decrease more slowly). Lemma 2 shows that we may concentrate on the case where $\gamma_k = k^{-1}$ to prove Corollary 1. Applying Proposition 1, $\hat{\rho}_n$ is then equal under parameter freeze to $\rho_{n,\nu,\theta}$ defined in (5). In light of Corollary 2, $\rho_{n,\nu,\theta}$ has the same behavior as the term $\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}_{\nu,\theta} \left[s(X_{t-1}, X_t, Y_t) \mid Y_{0:n} \right]$ that was analyzed above when proving Theorem 1.

References

Baum, L. E. and Eagon, J. A. (1967). An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology. *Bull. Am. Math. Soc.*, 73:360–363.

Baum, L. E. and Petrie, T. P. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37:1554–1563.

Baum, L. E., Petrie, T. P., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41(1):164–171.

Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26:1614–1635.

Briers, M., Doucet, A., and Maskell, S. (2010). Smoothing algorithms for state-space models. *Annals Institute Statistical Mathematics*, 62(1):61–89.

Cappé, O. (2009). Online sequential Monte Carlo EM algorithm. In *IEEE Workshop Statist. Signal Process. (SSP)*, Cardiff, Wales, UK.

Cappé, O. and Moulines, E. (2009). On-line expectation-maximization algorithm for latent data models. *J. Roy. Statist. Soc. B*, 71(3):593–613.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer.

Chung, S. H., Moore, J. B., Xia, L., Premkumar, L. S., and Gage, P. W. (1990). Characterization of single channel currrents using digital signal processing techniques based on hidden Markov models. *Phil. Trans. Roy. Soc. London B*, 329:265–285.

de Gunst, M. C. M., Künsch, H. R., and Schouten, J. G. (2001). Statistical analysis of ion channel data using hidden Markov models with correlated state-dependent noise and filtering. *J. Am. Statist. Assoc.*, 96(455):805–815.

Del Moral, P., Doucet, A., and Singh, S. S. (2009). Forward smoothing using sequential Monte Carlo. Technical Report CUED/F-INFENG/TR 638, Cambridge University Engineering Department.

Douc, R., Moulines, E., and Rydén, T. (2004). Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime. *Ann. Statist.*, 32(5):2254–2304.

Elliott, R. J., Aggoun, L., and Moore, J. B. (1995). *Hidden Markov Models: Estimation and Control*. Springer, New York.

Elliott, R. J., Ford, J. J., and Moore, J. B. (2002). On-line almost-sure parameter estimation for partially observed discrete-time linear systems with known noise characteristics. *Int. J. Adaptive Control Signal Process.*, 16(6):435–453.

Ford, J. (1998). *Adaptive hidden Markov model estimation and applications.* PhD thesis, Australian National University.

Kantas, N., Doucet, A., Singh, S., and Maciejowski, J. (2009). An overview of sequential Monte Carlo methods for parameter estimation in general state-space models. In *Proc. IFAC Symposium on System Identification (SYSID).*

Krishnamurthy, V. and Moore, J. B. (1993). On-line estimation of hidden Markov model parameters based on the kullback-leibler information measure. *IEEE Trans. Signal Process.*, 41(8):2557–2573.

Künsch, H. R. (2001). State space and hidden Markov models. In Barndorff-Nielsen, O. E., Cox, D. R., and Klueppelberg, C., editors, *Complex Stochastic Systems*, pages 109–173. CRC Publisher, Boca raton.

Le Gland, F. and Mevel, L. (1997). Recursive estimation in HMMs. In *Proc. IEEE Conf. Decis. Control*, pages 3468–3473.

Mongillo, G. and Denève, S. (2008). Online learning with hidden Markov models. *Neural Computation*, 20(7):1706–1716.

Neal, R. M. and Hinton, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in graphical models*, pages 355–368. MIT Press, Cambridge, MA, USA.

Polyak, B. T. (1990). A new method of stochastic approximation type. *Autom. Remote Control*, 51:98–107.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–285.

Roberts, W. J. J. and Ephraim, Y. (2008). An EM algorithm for ion-channel current estimation. *IEEE Trans. Signal Process.*, 56:26–33.

Ruppert, D. (1988). Efficient estimation from a slowly convergent Robbins-Monro process. Technical Report 781, Cornell University, School of Operations Research and Industrial Engineering.

Rydén, T. (1997). On recursive estimation for hidden Markov models. *Stochastic Process. Appl.*, 66(1):79–96.

Sato, M. (2000). Convergence of on-line EM algorithm. In *prooceedings of the International Conference on Neural Information Processing*, volume 1, pages 476–481.

Zeitouni, O. and Dembo, A. (1988). Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes. *IEEE Trans. Inform. Theory*, 34(4).