Online EM Algorithm for Hidden Markov Models

Olivier Cappé

Online EM Algorithm for Hidden Markov Models

Olivier Cappé

LTCI, TELECOM ParisTech & CNRS

Abstract

This paper is about the estimation of fixed model parameters in hidden Markov models using an online (or recursive) version of the Expectation-Maximization (EM) algorithm. It is first shown that under suitable mixing assumptions, the large sample behavior of the traditional (batch) EM algorithm may be analyzed through the notion of a limiting EM recursion, which is deterministic. This observation generalizes results previously obtained for latent data model with independent observations. By using the recursive implementation of smoothing computations associated with sum functionals of the hidden state, it is then possible to propose an online EM algorithm that generalizes an approach recently proposed in the case of HMMs with finite-valued observations. The performance of the proposed algorithm is numerically evaluated through simulations in the case of a noisily observed Markov chain.

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

1 Introduction

Hidden Markov modelling certainly constitutes one the contributions of statistical time series analysis which has had the most profound practical impact in the latest forty years. Hidden Markov models (HMMs) in their simplest form (i.e. when the state variable is finite) are sufficiently simple to give rise to efficient inference procedures while allowing for useful modelling of a wide range of 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 application in various scenarios, especially in cases where the model parameters are constrained.

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 structure 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 the EM algorithm. 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 data 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 considered in 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 in Cappé and Moulines (2009) for online EM estimation in the case of independent observations. The framework of Cappé and Moulines (2009) is first extended to the case of HMMs by exhibiting a limiting, or population-based, EM algorithm, corresponding to the case of infinitely many observations. The existence of the limiting EM algorithm does provide fruitful insights on the behavior of EM for ergodic HMMs when the number of observations gets large. However, and in contrast to the case of independent observations considered in Cappé and Moulines (2009), approximating the limiting EM algorithm with an online sample-based stochastic approximation algorithm turns out to be a difficult task for HMMs. The second contribution of the paper consists in recognizing the recursion of Mongillo and Denève (2008) as an instance of the the recursive smoothing schemes for sum functionals described, among others, by Zeitouni and Dembo (1988), Elliott et al. (1995), Cappé et al. (2005). This observation makes it possible to propose a generic online EM framework for HMMs with finite state-space.

The paper opens with a brief review of online EM in the case of independent observations. The main results, that is, the existence of a limiting EM recursion in the case of HMMs (Theorem 1) and the online procedure (Algorithm 1) which generalizes the algorithm of Mongillo and Denève (2008) are exposed in Section 3. Finally, Section 4 is about the application of the proposed procedure in the specific example of a Markov chain observed in Gaussian white noise.

2 Online EM in the Independent Case

2.1 Fisher Relation and the Limiting EM Algorithm

Consider the case of an i.i.d. (independently and identically distributed) missing data model, where $(Y_t)_{t\in\mathbb{Z}}$ denote the observation sequence and $(X_t)_{t\in\mathbb{Z}}$ are the associated latent (or unobservable) variables, hereafter referred to as *states*. The joint probability density function (pdf) of X_t and Y_t is denoted by $p_\theta(x_t, y_t)$ and $\ell_\theta(y_t)$ is the marginal pdf, or *likelihood*, associated to the observation, where θ denotes the model parameter. In the following, it is assumed that the observation sequence is distributed under an actual unknown parameter value $\theta_\star$ (although the case where this assumption is relaxed has also been analyzed in Cappé and Moulines, 2009).

Under suitable regularity assumptions, it is well known that the normalized maximum-likelihood criterion $\frac{1}{n}\sum_{t=1}^n \log \ell_\theta(Y_t)$ tends, $\mathrm{P}_{\theta_\star}$ almost surely, to the limiting contrast $-K(\ell_{\theta_\star}\|\ell_\theta)$, where $K(q_1\|q_2) = \int \log \frac{q_1}{q_2}(y) q_1(y) dy$ denotes the Kullback-Leibler divergence between q_1 and q_2. Similarly, the intermediate quantity of EM, $\frac{1}{n}\sum_{t=1}^n \mathrm{E}_\theta[\log p_{\theta'}(X_t, Y_t)|Y_t]$ tends to the deterministic limit $\mathrm{E}_{\theta_\star}[\mathrm{E}_\theta(\log p_{\theta'}(X_t, Y_t)|Y_t)]$. The *limiting EM algorithm* thus consists of

$$\begin{aligned} &\textbf{E-step} \quad \text{compute } \mathrm{E}_{\theta_\star}\left[\mathrm{E}_{\theta_k}\left(\log p_\theta(X_0, Y_0)\middle| Y_0\right)\right] ; \\ &\textbf{M-sep} \quad \text{set } \theta_{k+1} = \arg\max_{\theta\in\Theta} \mathrm{E}_{\theta_\star}\left[\mathrm{E}_{\theta_k}\left(\log p_\theta(X_0, Y_0)\middle| Y_0\right)\right] . \end{aligned} \tag{1}$$

It is straightforward to show that, as in the usual EM algorithm, each iteration of the limiting EM algorithm decreases the target criterion, that is $K(\ell_{\theta_\star}\|\ell_{\theta_{k+1}}) \le K(\ell_{\theta_\star}\|\ell_{\theta_k})$. Convergence of the limiting EM recursion to the set the stationary points of the limiting contrast $K(\ell_{\theta_\star}\|\ell_\theta)$ can be proved using the so-called Fisher identity (see discussion of Dempster et al., 1977):

$$\nabla_\theta \log \ell_\theta(Y_0) = \mathrm{E}_\theta\left[\nabla_\theta \log p_\theta(X_0, Y_0)\middle| Y_0\right] , \tag{2}$$

where ∇ denotes the gradient operator. The Fisher identity implies that $\mathrm{E}_{\theta_\star}\left[\mathrm{E}_\theta\left(\nabla_\theta \log p_\theta(X_0, Y_0|\, Y_0)\right]\right.$ and $\mathrm{E}_{\theta_\star}[\nabla_\theta \log \ell_\theta(Y_0)]$ coincide and, hence, that stationary points of the limiting EM mapping are also such that $\nabla K(\ell_{\theta_\star} \| \ell_\theta) = 0$.

2.2 Exponential Families and the Sufficient Statistics Reparameterization

In general, the algorithm in (1) is not very explicit and is mostly useful in case where the joint pdf p_θ belongs to an exponential family:

$$p_\theta(x, y) = h(x, y) \exp\left(\langle \psi(\theta), s(x, y)\rangle - A(\theta)\right) ,$$

where $\langle\cdot\rangle$ denotes the scalar product, $s(x, y)$ are the (complete-data) sufficient statistics and $A(\theta)$ is the log-partition function. Furthermore assume that the equation $\langle \nabla_\theta \psi(\theta), s\rangle - \nabla_\theta A(\theta) = 0$ has a unique solution for all achievable values of s, which is denoted by $\theta = \bar{\theta}(s)$ (for a canonical exponential family, where $\psi(\theta) = \theta$, this requirement is equivalent to assuming that the Fisher information matrix is positive definite for all values of θ). Then, the limiting EM algorithm in (1) may be equivalently rewritten as

$$\theta_{k+1} = \bar{\theta}\left(\mathrm{E}_{\theta_\star}\left[\mathrm{E}_{\theta_k}\left(s(X_0, Y_0)|\, Y_0\right)\right]\right) . \tag{3}$$

The algorithm may also be equivalently written in terms of the sequence of associated sufficient statistics which are such that $\theta_k = \bar{\theta}(S_k)$. Under this reparameterization, the limiting EM algorithm obeys the simple recursion

$$S_{k+1} = \mathrm{E}_{\theta_\star}\left[\mathrm{E}_{\bar{\theta}(S_k)}\left(s(X_0, Y_0)|\, Y_0\right)\right] , \tag{4}$$

where the stationary points of $K(\ell_{\theta_\star} \| \ell_\theta)$ are now more explicitly identified as the roots of the equation $S = \mathrm{E}_{\theta_\star}\left[\mathrm{E}_{\bar{\theta}(S)}\left(s(X_0, Y_0)|\, Y_0\right)\right]$.

2.3 Additive Decomposition and the Stochastic Approximation Algorithm

Of course, the limiting EM algorithm discussed above is not a parameter estimation procedure as it requires the knowledge of $\theta_\star$. To obtain a practical estimation algorithm, one simply needs to observe that $\mathrm{E}_{\theta_\star}\left[\mathrm{E}_\theta\left(s(X_0, Y_0)|\, Y_0\right)\right]$ may be estimated consistently from the observations as $\frac{1}{n}\sum_{t=1}^n \mathrm{E}_\theta\left[s(X_t, Y_t)|\, Y_t\right]$. The online algorithm is then obtained by using the usual stochastic approximation (or Robbins-Monro) procedure

$$\hat{S}_{n+1} = \gamma_{n+1}\mathrm{E}_{\bar{\theta}(\hat{S}_n)}\left[s(X_{n+1}, Y_{n+1})|\, Y_{n+1}\right] + (1 - \gamma_{n+1})\hat{S}_n , \tag{5}$$

where (γ_n) is a decreasing sequence of step-sizes. The principle of (5) has been first exposed by Neal and Hinton (1999), Sato and Ishii (2000) and latter extended by Sato (2000), Cappé and Moulines (2009). Cappé and Moulines (2009) analyzed the recursion in (5) to show that, under suitable assumptions: (i) it is indeed consistent, converging to the stationary points of $K(\ell_{\theta_\star} \| \ell_\theta)$; (ii) by properly choosing the rate of decrease of the step-sizes γ_n and using Polyak-Ruppert averaging (see also Section 4.3 below), $\hat{\theta}_{n+1} = \bar{\theta}(\hat{S}_{n+1})$ is an asymptotically efficient estimator of $\theta_\star$. Compared to gradient algorithms, it is quite remarkable that the algorithm in (5) can achieve asymptotic efficiency without trying to explicitly estimate the Fisher information matrix. Furthermore, it can be observed that the choice of performing the stochastic approximation in the domain of the sufficient statistics rather than in the parameter domain, that is, using (4) rather than (3) is also most natural given that only (4) may be directly estimated by a running average of properly selected functions of the observations.

3 Online EM for HMMs

I now discuss the generalization of the ideas presented above to the case of Hidden Markov models. Quite surprisingly, there are, under suitable mixing assumptions, direct analogs of the ideas presented in Sections 2.1 and 2.2. The tricky part consists in finding a suitable replacement for the stochastic approximation procedure of Section 2.3 which does not apply exactly for HMMs due to the time dependence.

In this section, 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 parameter $\theta_\star$; $\ell_{\theta_\star}$, $p_{\theta_\star}$, $\mathrm{P}_{\theta_\star}$ and $\mathrm{E}_{\theta_\star}$ refer to, respectively, the likelihood, the joint density of the states and observations, the probability, and ,the expectation under this model. In practice, one observes the observation sequence $(Y_t)_{t\geq 0}$ starting from time 0 only and the postulated initial distribution ν will be be arbitrary; $\ell_{\nu,\theta}$, $p_{\nu,\theta}$, $\mathrm{P}_{\nu,\theta}$ and $\mathrm{E}_{\nu,\theta}$ refer to the same quantity as previously but computed under this second model. Note that ν is not considered as a model parameter as it cannot be estimated consistently from a single trajectory (see also Chapters 10 and 12 of Cappé et al., 2005 on this point). Finally, it is assumed that the state variable takes its values in the finite set $\mathcal{X}$ and the state transition matrix and state conditional pdf that characterize the HMM are denoted, respectively, by $q_\theta(x, x')$ and $g_\theta(x, y)$.

3.1 The Limiting EM Algorithm

Under suitable assumptions (see below), the normalized HMM log-likelihood $\frac{1}{n}\log \ell_{\nu,\theta}(Y_0, \ldots, Y_n)$ converges, $\mathrm{P}_{\theta_\star}$ almost surely and in L^1, to the limiting contrast

$$c_{\theta_\star}(\theta) = \mathrm{E}_{\theta_\star}[\log \ell_\theta(Y_0|Y_{-\infty:-1})] \,. \tag{6}$$

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 $\nabla c_{\theta_\star}(\theta)$. 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,

$$\frac{1}{n}\nabla_\theta \log \ell_{\nu,\theta}(Y_0, \ldots, Y_n) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1}) \middle| Y_{0:n}\right] + \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\nabla_\theta \log p_{\nu,\theta}(X_0, Y_0)\middle| Y_{0:n}\right] \,. \tag{7}$$

For simplicity, the last term on the right-hand, whose influence is clearly vanishing with increasing values of n, will not be considered in the following. Hence, the consistency result for the score function combined with (7) implies that $\frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1})\middle| Y_{0:n}\right]$ also converges $\mathrm{P}_{\theta_\star}$ almost surely to $\nabla_\theta c_{\theta_\star}(\theta)$, the gradient of the limiting contrast.

To obtain an alternative representation of this limit, assume that both q_θ and g_θ belongs to exponential families such that

$$\begin{aligned} q_\theta(x, x') &= h^q(x, x') \exp\left(\langle \psi^q(\theta), s^q(x, x')\rangle - A^q(\theta)\right) \,, \\ g_\theta(x, y) &= h^g(x, y) \exp\left(\langle \psi^g(\theta), s^g(x, y)\rangle - A^g(\theta)\right) \,. \end{aligned} \tag{8}$$

Note that under our assumption that $\mathcal{X}$ is finite, the first requirement is always satisfied.

For the sake of conciseness, I will adopt in the rest of Section 3 a condensed representation –which is also slightly more general than the HMM case– by assuming that the joint state and observation conditional density $p_\theta(x_t, y_t|x_{t-1})$ belongs to an exponential family such that

$$p_\theta(x_t, y_t|x_{t-1}) = h(x_t, y_t) \exp\left(\langle \psi(\theta), s(x_{t-1}, x_t, y_t)\rangle - A(\theta)\right) , \tag{9}$$

where $h(x, y) = h^q(x, x')h^g(x, y)$, $A(\theta) = A^q(\theta) + A^g(\theta)$ and

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

In this case, the non-vanishing term in the r.h.s. of (7) may rewritten as

$$\frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} \nabla_\theta \log p_\theta(X_t, Y_t|X_{t-1}) \middle| Y_{0:n}\right] = \left\langle \nabla_\theta\psi(\theta), \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n}\right]\right\rangle - \nabla_\theta A(\theta) . \tag{10}$$

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

Theorem 1. *Assume that* (i) $\mathcal{X}$ *is a finite set;* (ii) *the transition matrix is such that* $q_\theta(x, x') \geq \epsilon > 0$ *for all* $\theta \in \Theta$*;* (iii) $\sup_\theta \sup_y \bar{g}_\theta(y) < \infty$ *and* $\mathrm{E}_{\theta_\star}\left[|\log \inf_\theta \bar{g}_\theta(Y_0)|\right] < \infty$*, where* $\bar{g}_\theta(y) = \sum_x g_\theta(x, y)$*;* (iv) *the parameter space* Θ *is compact and* $\theta_\star \in \mathrm{interior}(\Theta)$*;* (v) ψ_q, A_q, ψ_g, A_g *in* (8) *are continuously differentiable functions on* $\mathrm{interior}(\Theta)$*; and,* (vi) *the equation* $\langle \nabla_\theta\psi(\theta), s\rangle - \nabla_\theta A(\theta) = 0$ *has a unique solution, which is denoted by* $\theta = \bar{\theta}(s)$*.*

Then,

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

and the stationary points of the limiting EM algorithm

$$\theta_{k+1} = \bar{\theta}\left\{\mathrm{E}_{\theta_\star}\left(\mathrm{E}_{\theta_k}\left[s(X_{-1}, X_0, Y_0) \middle| Y_{-\infty:\infty}\right]\right)\right\} \tag{11}$$

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

3.2 Online EM

Theorem 1 suggests a principle similar to the case of the i.i.d. mixture model of Section 2. To obtain an online algorithm however, one needs to be able to estimate consistently the limit $\mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0) \middle| Y_{-\infty:\infty}\right]\right)$. The normalized sum $\frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n}\right]$ is not directly a candidate as it is well known that it cannot be computed recursively when incorporating new observations. However, following the idea originally proposed by (Zeitouni and Dembo, 1988, Elliott et al., 1995), define

$$\phi_{n,\nu,\theta}(x) = \mathrm{P}_{\nu,\theta}\left(X_n = x \middle| Y_{0:n}\right) , \tag{12}$$

$$\rho_{n,\nu,\theta}(x) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \middle| Y_{0:n}, X_n = x\right] , \tag{13}$$

which are such that $\frac{1}{n}\sum_{x\in\mathcal{X}}\phi_{n,\nu,\theta}(x)\rho_{n,\nu,\theta}(x) = \mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n} s(X_{t-1},X_t,Y_t)|\, Y_{0:n}\right]$. The appeal of this new decomposition being that $\phi_{n,\nu,\theta}$ and $\rho_{n,\nu,\theta}$ can be updated recursively as shown by the following Proposition.

Proposition 1. *$\phi_{n,\nu,\theta}$ and $\rho_{n,\nu,\theta}$ may be computed according to the recursion*

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

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

Recursion *For $n \geq 0$, compute, for $x \in \mathcal{X}$,*

$$\phi_{n+1,\nu,\theta}(x) = \frac{\sum_{x'\in\mathcal{X}}\phi_{n,\nu,\theta}(x')q_\theta(x',x)g_\theta(x,Y_{n+1})}{\sum_{x',x''\in\mathcal{X}^2}\phi_{n,\nu,\theta}(x')q_\theta(x',x'')g_\theta(x'',Y_{n+1})} \tag{14}$$

$$\rho_{n+1,\nu,\theta}(x) = \sum_{x'\in\mathcal{X}}\left\{\frac{1}{n+1}s(x',x,Y_{n+1}) + \left(1-\frac{1}{n+1}\right)\rho_{n,\nu,\theta}(x')\right\}\frac{\phi_{n,\nu,\theta}(x')q_\theta(x',x)}{\sum_{x''\in\mathcal{X}}\phi_{n,\nu,\theta}(x'')q_\theta(x'',x)} \tag{15}$$

In Proposition 1 above, the rightmost term in (15) 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 (see Chapter 4 of Cappé et al., 2005 for a complete proof).

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

$$\hat{\phi}_{n+1}(x) = \frac{\sum_{x' \in \mathcal{X}} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x, Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} , \tag{16}$$

$$\hat{\rho}_{n+1}(x) = \sum_{x' \in \mathcal{X}} \left\{ \gamma_{n+1} s(x', x, Y_{n+1}) + (1 - \gamma_{n+1}) \hat{\rho}_n(x') \right\} \frac{\hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x)}{\sum_{x'' \in \mathcal{X}} \hat{\phi}_n(x'') q_{\hat{\theta}_n}(x'', x)} . \tag{17}$$

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

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

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

3.3 Discussion

Mongillo and Denève (2008) considered, the particular case of finite valued HMMs where 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 whatever the chosen parameterization of the model (Mongillo and Denève (2008) consider only parameterization by the sets of conditional probabilities), the complete-data sufficient statistics can be chosen to be $s(X_{t-1}, X_t, Y_t) = (\mathbb{1}\{X_{t-1} = i, X_t = j, Y_t = k\})_{(i,j,k) \in \mathcal{X}^2 \times \mathcal{Y}}$. 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 $\mathrm{E}_{\nu,\theta} \left[\left(\sum_{t=1}^{n} s(X_{t-1}, X_t, Y_t) \right) \mathbb{1}\{X_n = x\} \middle| Y_{0:n} \right]$. By multiplying (17) by $\hat{\phi}_{n+1}(x)$ and using (16) and , one obtains the following online update

$$\begin{aligned} \hat{\tau}_{n+1}(x) = \gamma_{n+1} \sum_{x' \in \mathcal{X}} s(x', x, Y_{n+1}) \frac{\hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x', Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} \\ + (1 - \gamma_{n+1}) \sum_{x' \in \mathcal{X}} \hat{\tau}_n(x') \frac{q_{\hat{\theta}_n}(x', x) g_{\hat{\theta}_n}(x', Y_{n+1})}{\sum_{x', x'' \in \mathcal{X}^2} \hat{\phi}_n(x') q_{\hat{\theta}_n}(x', x'') g_{\hat{\theta}_n}(x'', Y_{n+1})} , \end{aligned} \tag{18}$$

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 (17) or (18) is practically equivalent. The form of (17) is preferable from a conceptual point of view as it clearly shows that the new observation Y_{n+1} only plays a role in the filter update (16). The limiting behavior of $\hat{\rho}_n(x)$ is also expected to be simpler: From corollary 1 and proceeding as in the proof of Theorem 1, it is easily shown that, *for any* $x \in \mathcal{X}$, $\rho_{n,\nu,\theta}(x)$ indeed converges $\mathrm{P}_{\theta_\star}$ almost surely to the same fixed limit as that exhibited in Theorem 1, that is, $\mathrm{E}_{\theta_\star} \left(\mathrm{E}_{\theta} \left[s(X_{-1}, X_0, Y_0) \middle| Y_{-\infty:\infty} \right] \right)$. Hence, it is conjectured that $\hat{\rho}_n(x)$ will tends, as n increases, to a limit that is independent of x.

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, or when $\gamma_n = 1/n$, which is selected, following Neal and Hinton (1999), by analogy with the batch EM algorithm

and its recursive rewriting in Proposition 1. In general however, the choice $\gamma_n = 1/n$ is not very recommendable for a stochastic approximation procedure and step-sizes of the form $\gamma_n = 1/n^\gamma$ with γ in the range 05–0.6 combined with Polyak-Ruppert averaging are preferable (see discussion in Cappé and Moulines, 2009). This point is illustrated in the numerical simulations of Section 4 below. The role of $n_{\min}$ is only to guarantee that the M-step update is numerically well behaved (Cappé and Moulines, 2009) and for this purpose, a small value of $n_{\min}$ is usually sufficient (for instance, $n_{\min} = 20$ is used in the simulations of Section 4.3).

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 (18) 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 the next section 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. Although, the comparison is not directly meaningful as the batch EM algorithm does necessitate several iterations to converge (see numerical illustrations in Section 4.3), it is true that the scaling of the numerical complexity of the online-EM algorithm with $|\mathcal{X}|$ may 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 (17), one coordinate of the gradient for each of the couples $(x, x') \in \mathcal{X}^2$ (see, e.g., Le Gland and Mevel, 1997). Note that if the transition matrix is structured —i.e., parametered by a low dimensional parameter rather than by all its individual entries—, the numerical cost associated to the implementation of the approach will be reduced to an order of the number of parameters times $|\mathcal{X}|^2$.

4 Application to Gaussian HMMs

4.1 HMM with Product Parameterization

For the sake of concreteness, I consider in following the case where the state variables $\{X_t\}_{t\geq 0}$ 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 $\rho_{n,\nu,\theta}$:

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

$$\rho^g_{n,\nu,\theta}(i,k;\theta) = \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t=i\}s(Y_t)\middle| Y_{0:n}, X_n=k\right] , \tag{20}$$

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 (20) and (13), which is that (20) also incorporates the initial ($t=0$) conditional likelihood term, i.e., the contribution corresponding to the rightmost term on the r.h.s. of (7). As noted earlier, this difference is minor and does not modify the long-term behavior of the algorithm.

With these notations, Eq. (17) in Algorithm 1 is implemented as

$$\hat\rho^q_{n+1}(i,j,k) = \gamma_{n+1}\delta(j-k)\hat r_{n+1}(i|j) + (1-\gamma_{n+1})\sum_{k'=1}^{m}\hat\rho^q_n(i,j,k')\hat r_{n+1}(k'|k) , \tag{21}$$

$$\hat\rho^g_{n+1}(i,k) = \gamma_{n+1}\delta(i-k)s(Y_{n+1}) + (1-\gamma_{n+1})\sum_{k'=1}^{m}\hat\rho^g_n(i,k')\hat r_{n+1}(k'|k) , \tag{22}$$

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 :

$$\hat r_{n+1}(i|j) = \frac{\hat\phi_n(i)q_{\hat\theta_n}(i,j)}{\sum_{i'=1}^{m}\hat\phi_n(i')q_{\hat\theta_n}(i',j)} . \tag{23}$$

A complete iteration of the online algorithm involves the approximate filter update (16) and the stochastic approximation statistics updates (21) and (22) followed by an application of the M-step function $\bar\theta$ to $\hat S^q_{n+1}(i,j) = \sum_{k=1}^m \hat\rho^q_{n+1}(i,j,k)\hat\phi_{n+1}(k)$ and $\hat S^g_{n+1}(i) = \sum_{k=1}^m \hat\rho^g_{n+1}(i,k)\hat\phi_{n+1}(k)$. 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

$$q_{\hat\theta_n}(i,j) = \bar\theta\left(\hat S^q_n\right) = \frac{\hat S^q_n(i,j)}{\sum_{j=1}^m \hat S^q_n(i,j)} . \tag{24}$$

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

$$\mu_{\hat\theta_n}(i) = \bar\theta\left(\hat S^g_{n,0}, \hat S^g_{n,1}\right) = \frac{\hat S^g_{n,1}(i)}{\hat S^g_{n,0}(i)} , \tag{25}$$

$$\Sigma_{\hat\theta_n}(i) = \bar\theta\left(\hat S^g_{n,0}, \hat S^g_{n,1}, \hat S^g_{n,2}\right) = \frac{\hat S^g_{n,2}(i)}{\hat S^g_{n,0}(i)} - \mu_{\hat\theta_n}(i)\mu^{\mathrm{t}}_{\hat\theta_n}(i) . \tag{26}$$

The derivation of (24) and (25)–(26) is straightforward but some more details are provided in the next section for a particular case of Gaussian HMM.

4.2 Markov Chain Observed in Gaussian Noise

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

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

where (V_t) is a scalar additive Gaussian noise of variance υ 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 υ.

In this case, the intermediate quantity of the batch EM algorithm may be written, up to constants, as

$$\sum_{i=1}^{m}\sum_{j=1}^{m} S_n^q(i,j)\log q(i,j) - \frac{1}{2\upsilon}\sum_{i=1}^{m}\left(S_{n,2}^g(i;\theta) - 2\mu(i)S_{n,1}^g(i) + \mu^2(i)S_{n,0}^g(i)\right) \tag{27}$$

where

$$\begin{aligned}
S_n^q(i,j) &= \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=1}^{n}\mathbb{1}\{X_{t-1}=i, X_t=j\}\middle| Y_{0:n}\right] \,,\\
S_{n,0}^g(i) &= \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t=i\}\middle| Y_{0:n}\right] \,,\\
S_{n,1}^g(i) &= \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t=i\}Y_t\middle| Y_{0:n}\right] \,,\\
S_{n,2}^g(i) &= \frac{1}{n}\mathrm{E}_{\nu,\theta}\left[\sum_{t=0}^{n}\mathbb{1}\{X_t=i\}Y_t^2\middle| Y_{0:n}\right] \,.
\end{aligned}$$

Maximization of (27) with respect to q_θ, μ_θ and υ_θ directly yields (24) as well as

$$\mu(i) = \bar{\theta}\left(S_{n,0}^g, S_{n,1}^g\right) = \frac{S_{n,1}^g(i)}{S_{n,0}^g(i)} \,, \tag{28}$$

$$\upsilon = \bar{\theta}\left(S_{n,0}^g, S_{n,1}^g, S_{n,2}^g\right) = \frac{\sum_{i=1}^{m}\left(S_{n,2}^g(i) - \mu^2(i)S_{n,0}^g(i)\right)}{\sum_{i=1}^{m} S_{n,0}^g(i)} \,. \tag{29}$$

It is easily checked that, as usual, the M-step equations (24) and (29) do 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 $S_{n,2}^g$ could be avoided as it is only needed in the M-step under the form $\sum_{i=1}^{m} S_{n,2}^g(i)$, 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* $1 \leq i, j, k, \leq m$ *and* $0 \leq d \leq 2$,

$$\hat{\phi}_0(k) = \frac{\nu(k) g_{\hat{\theta}_0}(k, Y_0)}{\sum_{k'=1}^m g_{\hat{\theta}_0}(k, Y_0)} ,$$
$$\hat{\rho}_0^q(i, j, k) = 0 ,$$
$$\hat{\rho}_{0,d}^g(i, k) = \delta(i - k) Y_0^d .$$

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

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

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

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

where $\hat{r}_{n+1}(i|j) = \hat{\phi}_n(i) \hat{q}_n(i, j) / \sum_{i'=1}^m \hat{\phi}_n(i') \hat{q}_n(i', j)$.

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

$$\hat{S}_{n+1}^q(i, j) = \sum_{k'=1}^m \hat{\rho}_{n+1}^q(i, j, k') \hat{\phi}_{n+1}(k') ,$$
$$\hat{q}_{n+1}(i, j) = \frac{\hat{S}_{n+1}^q(i, j)}{\sum_{j'=1}^m \hat{S}_{n+1}^q(i, j')} ,$$
$$\hat{S}_{n+1,d}^g(i) = \sum_{k'=1}^m \hat{\rho}_{n+1,d}^g(i, k') \hat{\phi}_{n+1}(k') ,$$
$$\hat{\mu}_{n+1}(i) = \frac{S_{n+1,1}^g(i)}{S_{n+1,0}^g(i)} ,$$
$$\hat{\upsilon}_{n+1} = \frac{\sum_{i'=1}^m \left(S_{n+1,2}^g(i') - \hat{\mu}_{n+1}^2(i') S_{n+1,0}^g(i') \right)}{\sum_{i'=1}^m S_{n+1,0}^g(i')} .$$

4.3 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

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

With these parameters, state identification is a difficult task as the separation of the means corresponding to the two state 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

$$
\begin{aligned}
&q_\star(1,1) = 0.7\,,\ \mu_\star(1) = -0.5\,,\\
&q_\star(2,2) = 0.5\,,\ \mu_\star(2) = 0.5\,,\\
&\upsilon_\star = 2\,,
\end{aligned}
$$

and run on 100 independent trajectories simulated from the model.

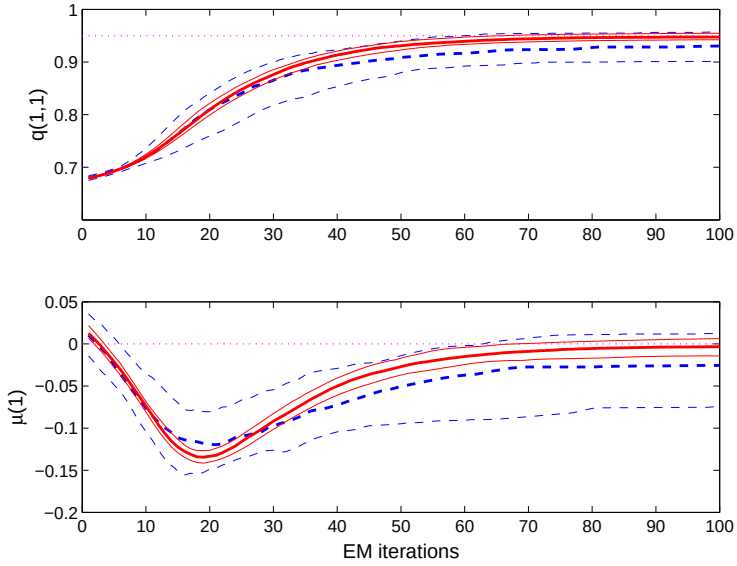

Figure 1: Estimated values of $q(1,1)$ (top) and $\mu(1)$ (bottom) as a function of the number of batch EM iterations for $n = 500$ (dotted lines) and $n = 8000$ (solid lines) observations. The plot is based on 100 independent realizations summarized by the median (bold central line) and the upper and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

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.

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

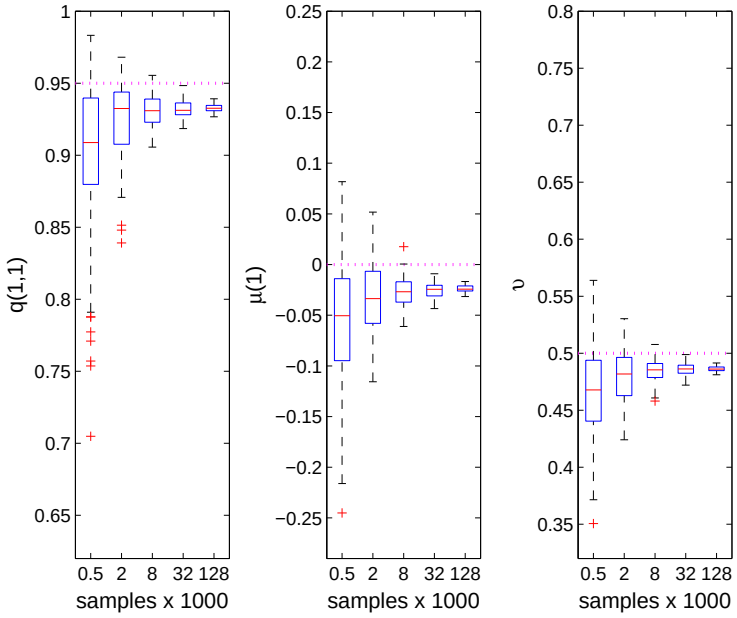

Figure 2: Estimation results when using 50 batch EM iterations. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. Box and whiskers plot based on 100 independent realizations.

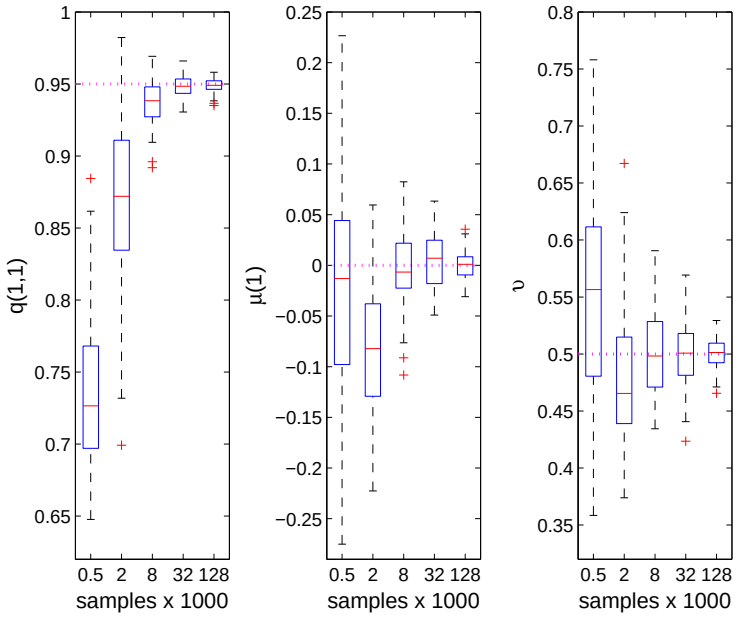

Figure 3: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. Box and whiskers plot based on 100 independent realizations.

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 known that the true maximum likelihood estimator does converge, as 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 box). The 100 realizations 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 outlined above as 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 still is 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 —$\gamma_n = n^{-0.6}$ was chosen following the recommendations of Cappé and Moulines (2009)— 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 3 as well as (Neal and Hinton, 1999)) provides estimates that converge much to slowly to be usable in any practical application. This observation is certainly a consequence of the temporal dependence between the observations and, correlatively, of the time taken by the filtering and smoothing relations to forget their initial state (as in the example under consideration the assumptions of Theorem 1 as satisfied for θ in a neighborhood of $\theta_\star$ with a constant $\epsilon = q_\star(1,2) = 0.05$, which is rather small).

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 (2.4 GHz processor).

Note that the comparison between Figure 2 and Figure 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 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

[1]The MATLAB/OCTAVE code used for the simulations is very simple and will be made available from the web. 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.

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 3.3, 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 performances of the algorithm, Figure 4 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 (27) so as to obtain

$$\begin{aligned}
\frac{1}{n}\nabla_{q(i,j)} \log \ell_\theta(Y_{0:n}) &= \frac{S^q_n(i,j)}{q(i,j)} - \frac{S^q_n(i,m)}{q(i,m)} \quad (\text{for } 1 \leq j < m)\,, \\
\frac{1}{n}\nabla_{\mu(i)} \log \ell_\theta(Y_{0:n}) &= \frac{S^g_{n,1} - \mu(i) S^g_{n,0}}{\upsilon}\,, \\
\frac{1}{n}\nabla_{\upsilon} \log \ell_\theta(Y_{0:n}) &= \sum_{i=1}^{m} \frac{1}{2\upsilon^2} \left(S^g_{n,2}(i;\theta) - 2\mu(i) S^g_{n,1}(i) + \mu^2(i) S^g_{n,0}(i) \right)\,.
\end{aligned}$$

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

Additionally, Figure 4 also displays results that have been post-processed using averaging (Polyak, 1990, Ruppert, 1988). In Figure 4, 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 $1/\sqrt{n}$ (see Cappé and Moulines, 2009, for an illustration in the case of the online EM for independent observations). At least for $\mu(1)$ and υ, Figure 4 suggest 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 disappear 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 4 is due to n_{avg} being too small (see corresponding panel in Figure 3). Although, this could be corrected 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.

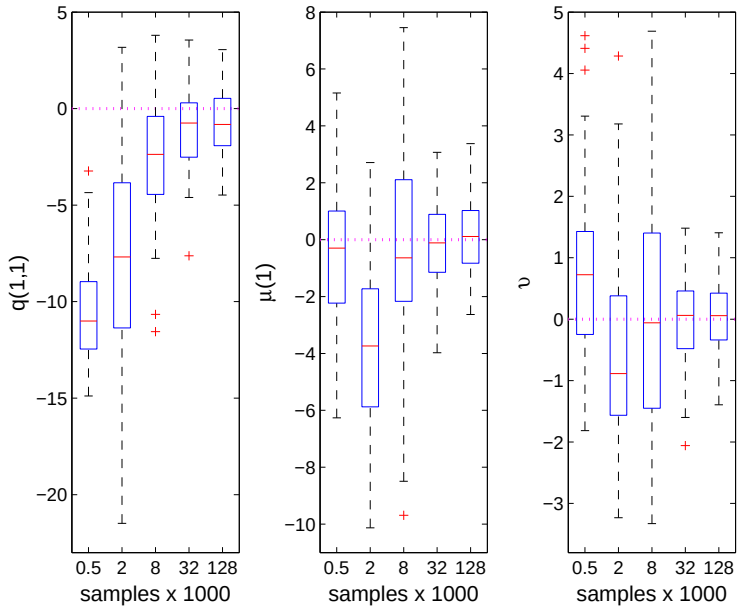

Figure 4: Estimation results when using the online EM algorithm with $\gamma_n = n^{-0.6}$ with Polyak-Ruppert started after $n = 8000$. From left to right, estimated values of $q(1,1)$, $\mu(1)$ and υ for values of n ranging from 0.5 to 128 thousands of samples. The estimated values are centered and scaled so as to be comparable with a unitary asymptotic standard deviation. Box and whiskers plot based on 100 independent realizations.

5 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), Cappé and Moulines (2009) consists in reparameterzing the model in the space of sufficient statistics and approximating the limiting EM recursion using a stochastic approximation procedure. Theorem 1 provides a first argument demonstrating that this idea is also 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 used in Algorithm 1. As discussed in Section 3, this possibility requires that the auxiliary quantity $\rho_{n,\nu,\theta}$ defined in (13) be approximated during the course of the algorithm.

Although the performance reported in Section 4 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 kernel convolution involved in (17). As discussed in Sections 3.3 and 4.3, the proposed algorithm may become less attractive, from a computational point of view, when used in models with many distinct state values. For such cases, it would be of interest to consider specific versions of the algorithm, using some form of approximation, perhaps based on Monte Carlo simulations.

A Proof of Theorem 1

Theorem 1 mainly relies on the use of a two-sided forgetting result which is first proved as Corollary 1 below. This result generalizes the one-sided forgetting bounds of Douc et al. (2004) and allow conditioning with respect to both past and future observations, which is required for studying the asymptotic behavior of (7) 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

$$\|J_{\alpha_1,q,\beta_1} - J_{\alpha_2,q,\beta_2}\|_1 \leq \frac{1}{\epsilon}\left(\|\alpha_1-\alpha_2\|_1 + \|\beta_1-\beta_2\|_1\right) , \tag{30}$$

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 (30) is obtained by noting that $|\alpha_1(x)\beta_1(x') - \alpha_2(x)\beta_2(x')| \leq |\alpha_1(x)-\alpha_2(x)|\beta_1(x') + |\beta_1(x')-\beta_2(x')|\alpha_2(x)$. □

Corollary 1. *Under the assumption 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 $1 \leq t \leq n$,*

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

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

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

observing that $\mathrm{P}_\theta(X_{t-1} = x, X_t = x' | Y_{0:t-1}, X_0 = x_{0,i}, X_n = x_{n,i}) = J_{\alpha_i,q_\theta,\beta_i}$. 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 1 follow by the general inequality $|\mu_1(g)-\mu_2(g)| \leq \frac{1}{2}\|\mu_1-\mu_2\|_1 \sup_{z_1,z_2\in\mathcal{Z}^2}|g(z_1)-g(z_2)|$. □

Note the fact that the backward function $\beta_i(x') = \mathrm{P}_\theta(Y_{t:n}, X_n = x_{n,i} | X_t = x')$ 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 (Briers et al., 2004, Cappé et al., 2005) but it is easily checked that Corollary 1 holds in greater generality under the "strong mixing conditions" discussed in Section 4.3 of Cappé et al. (2005).

Proof of Theorem 1. Corollary 1 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 $m(y)$ is an upper bound for $|s(X_{-1}, X_0, Y_0)|$, which 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, $\theta_\star \in \mathrm{interior}(\Theta)$, standard results on exponential family imply that $m(Y_0)$ admits finite moments of all orders under $\mathrm{P}_{\theta_\star}$. 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 $\mathrm{P}_{\theta_\star}$. 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 $\mathrm{E}_{\theta_\star}[m(Y_0)^p] < \infty$ for all p, standard applications of Markov inequality and Borel-Cantelli Lemma imply that the r.h.s. of the above expression tends 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 (7) converge a.s. to the gradient $\nabla_\theta c_{\theta_\star}(\theta)$ of the limiting contrast defined in (6). Eq. (10) combined with the previous result then shows that parameter values θ for which $\nabla_\theta c_{\theta_\star}(\theta)$ vanishes are also such that $\langle \nabla_\theta \psi(\theta), \mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty}\right]\right)\rangle - \nabla_\theta A(\theta) = 0$, that is, $\bar{\theta}\left\{\mathrm{E}_{\theta_\star}\left(\mathrm{E}_\theta\left[s(X_{-1}, X_0, Y_0) | Y_{-\infty:\infty}\right]\right)\right\} = \theta$. □

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.

[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 $\theta \in \Theta$.

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. (2004). Smoothing algorithms for state-space models. Technical Report TR-CUED-F-INFENG 498, University of Cambridge, Department of Engineering.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39(1):1–38 (with discussion).

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.

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.

Sato, M. and Ishii, S. (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Computation*, 12:407–432.

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).