HAL open science

Online EM Algorithm for Latent Data Models

Olivier Cappé, Eric Moulines

To cite this version:

Olivier Cappé, Eric Moulines. Online EM Algorithm for Latent Data Models. Journal of the Royal Statistical Society: Series B, 2009, 71 (3), pp.593-613. 10.1111/j.1467-9868.2009.00698.x . hal-00201327v4

HAL Id: hal-00201327

https://hal.science/hal-00201327v4

Submitted on 1 Mar 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Online EM Algorithm for Latent Data Models

Olivier Cappé & Eric Moulines

LTCI, TELECOM ParisTech, CNRS.
46 rue Barrault, 75013 Paris, France.

Abstract

In this contribution, we propose a generic online (also sometimes called adaptive or recursive) version of the Expectation-Maximisation (EM) algorithm applicable to latent variable models of independent observations. Compared to the algorithm of Titterington (1984), this approach is more directly connected to the usual EM algorithm and does not rely on integration with respect to the complete data distribution. The resulting algorithm is usually simpler and is shown to achieve convergence to the stationary points of the Kullback-Leibler divergence between the marginal distribution of the observation and the model distribution at the optimal rate, i.e., that of the maximum likelihood estimator. In addition, the proposed approach is also suitable for conditional (or regression) models, as illustrated in the case of the mixture of linear regressions model.

Keywords: Latent data models, Expectation-Maximisation, adaptive algorithms, online estimation, stochastic approximation, Polyak-Ruppert averaging, mixture of regressions.

1 Introduction

The EM (Expectation-Maximisation) algorithm (Dempster et al., 1977) is a popular tool for maximum-likelihood (or maximum a posteriori) estimation. The common strand to problems where this approach is applicable is a notion of *incomplete data*, which includes the conventional sense of missing data but is much broader than that. The EM algorithm demonstrates its strength in situations where some hypothetical experiments yields *complete* data that are related to the parameters more conveniently than the measurements are. Problems where the EM algorithm has proven to be useful include, among many others, mixture of densities (Titterington et al., 1985), censored data models (Tanner, 1993), etc. The EM algorithm has several appealing properties. Because it relies on complete data computations, it is generally simple to implement: at each iteration, *(i)* the so-called *E-step* only involves taking expectation over the conditional distribution of the latent data given the observations and *(ii)* the *M-step* is analogous to complete data weighted maximum-likelihood estimation. Moreover, *(iii)* the EM algorithm naturally is an ascent algorithm, in the sense that it increases the (observed) likelihood at each iteration. Finally under some mild additional conditions, *(iv)* the EM algorithm may be shown to converge to a *stationary point* (i.e., a point where the gradient vanishes) of the log-likelihood (Wu, 1983). Note that convergence to the maximum likelihood estimator cannot in general be guaranteed due to possible presence of multiple stationary points.

When processing large data sets or data streams however, the EM algorithm becomes impractical due to the requirement that the whole data be available at each iteration of the algorithm. For this reason, there has been a strong interest for online variants of the EM which make it possible to estimate the parameters of a latent data model without storing the data. In this work, we consider online algorithms for latent data models with independent observations. The dominant approach (see also Section 2.2 below) to online EM-like estimation follows the method proposed by Titterington (1984) which consists in using a stochastic approximation algorithm, where the parameters are updated after each new observation using the gradient of the incomplete data likelihood weighted by the complete

data Fisher information matrix. This approach has been used, with some variations, in many different applications (see, e.g., Chung and Böhme, 2005; Liu et al., 2006); a proof of convergence was given by Wang and Zhao (2006).

In this contribution, we propose a new online EM algorithm that sticks more closely to the principles of the original (batch-mode) EM algorithm. In particular, each iteration of the proposed algorithm is decomposed into two steps, where the first one is a stochastic approximation version of the E-step aimed at incorporating the information brought by the newly available observation, and, the second step consists in the maximisation program that appears in the M-step of the traditional EM algorithm. In addition, the proposed algorithm does not rely on the complete data information matrix, which has two important consequences: firstly, from a practical point of view, the evaluation and inversion of the information matrix is no longer required, secondly, the convergence of the procedure does not rely on the implicit assumption that the model is *well-specified*, that is, that the data under consideration is actually generated by the model, for some unknown value of the parameter. As a consequence, and in contrast to previous work, we provide an analysis of the proposed algorithm also for the case where the observations are not assumed to follow the fitted statistical model. This consideration is particularly relevant in the case of conditional missing data models, a simple case of which is used as an illustration of the proposed online EM algorithm. Finally, it is shown that, with the additional use of Polyak-Ruppert averaging, the proposed approach converges to the stationary points of the limiting normalised log-likelihood criterion (i.e., the Kullback-Leibler divergence between the marginal density of the observations and the model pdf) at a rate which is optimal.

The paper is organised as follows: In Section 2, we review the basics of the EM and associated algorithms and introduce the proposed approach. The connections with other existing methods are discussed at the end of Section 2.3 and a simple example of application is described in Section 2.4. Convergence results are stated in Section 3, first in term of consistency (Section 3.1) and then of convergence rate (Section 3.2), with the corresponding proofs given in Appendix A. Finally in Section 4, the performance of this approach is illustrated in the context of mixture of linear regressions.

2 Algorithm Derivation

2.1 EM Basics

In this section, we review the key properties of the EM algorithm as introduced by Dempster et al. (1977). The latent variable statistical model postulates the existence of a non-observable or latent variable X distributed under $f(x;\theta)$ where $\{f(x;\theta), \theta \in \Theta\}$ denotes a parametric family of probability density functions indexed by a parameter value $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$. The observation Y is then viewed as a deterministic function of X which takes its values in the set $\mathcal{Y}$. This latent variable mechanism provides a unified framework for situations which includes missing data, censored observations, noisily observed data, ... (Dempster et al., 1977).

We will denote by $g(y;\theta)$ the (observed) likelihood function induced by the latent data model. In addition, the notations $\mathbb{E}_\theta[\cdot]$ and $\mathbb{E}_\theta[\cdot|Y]$ will be used to denote, respectively, the expectation and conditional expectation under the model parameterised by θ. Likewise, let π denote the probability density function of the observation Y, where we stress again that we do not restrict ourselves to the case where $\pi(\cdot) = g(\cdot;\theta^\star)$, for an unknown value $\theta^\star$ of the parameter. The notations $\mathbb{P}_\pi$ and $\mathbb{E}_\pi$ will be used to denote probability and expectation under the actual distribution of the observation.

Given n independent and identically distributed observations $Y_{1:n} \stackrel{\text{def}}{=} (Y_1, \ldots, Y_n)$, the maximum likelihood estimator is defined as $\hat{\theta}_n \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta\in\Theta} n^{-1} \log g(Y_{1:n};\theta)$, where

$$\log g(Y_{1:n};\theta) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \log g(Y_i;\theta) \ . \tag{1}$$

Note that we normalise the log-likelihood (by n) to ease the transition to the online setting where n increases when new observations become available.

The EM algorithm is an iterative optimisation algorithm that maximises the above (normalised) log-likelihood function despite the possibly complicated form of g resulting from the latent data model. Traditionally, each EM iteration is decomposed in two steps. The E-step consists in evaluating the conditional expectation

$$Q_{\hat{\theta}_k}(Y_{1:n};\theta) \stackrel{\text{def}}{=} n^{-1}\sum_{i=1}^{n}\mathbb{E}_{\hat{\theta}_k}\left[\log f(X_i;\theta)\middle|Y_i\right] \tag{2}$$

where $\hat{\theta}_k$ is the current estimate of θ, after k iterations of the algorithm. In the M-step, the value of θ maximising $Q_{\hat{\theta}_k}(Y_{1:n};\theta)$ is found. This yields the new parameter estimate $\hat{\theta}_{k+1}$. This two step process is repeated until convergence. The essence of the EM algorithm is that increasing $Q_{\hat{\theta}_k}(Y_{1:n};\theta)$ forces an increase of the log-likelihood $\log g(Y_{1:n};\theta)$ (Dempster et al., 1977).

For $\mu : \Theta \to \mathbb{R}$ a differentiable function, denote by $\nabla_\theta \mu = (\partial\mu/\partial\theta_1, \ldots, \partial\mu/\partial\theta_{d_\theta})^T$ the *gradient* of μ. If μ is twice differentiable, we denote by $\nabla^2_\theta \mu$ the *Hessian matrix* which is a $d_\theta \times d_\theta$ matrix whose components are given by $[\nabla^2_\theta \mu]_{i,j} = \frac{\partial^2 \mu}{\partial\theta_i \partial\theta_j}$, $1 \leq i, j \leq d_\theta$. Following Lange (1995), if the M-step of the EM algorithm is replaced by a Newton update, one obtains, assuming some regularity, the following recursion

$$\hat{\theta}_{k+1} = \hat{\theta}_k + \gamma_{k+1}\left[J(Y_{1:n};\hat{\theta}_k)\right]^{-1}\sum_{i=1}^{n}\mathbb{E}_{\hat{\theta}_k}\left[\nabla_\theta \log f(X_i;\hat{\theta}_k)\middle|Y_i\right] \ , \tag{3}$$

where γ_{k+1} is a step size ($\gamma_{k+1} = 1$ correspond to the actual Newton update) and $J(Y_{1:n};\theta) = n^{-1}\sum_{i=1}^{n} J(Y_i;\theta)$ with $J(y;\theta) = -\mathbb{E}_\theta\left[\nabla^2_\theta \log f(X;\theta)\middle|Y = y\right]$. Note that due to the so-called Fisher's identity (see discussion of Dempster et al., 1977), the gradient term indeed coincides with the (observed data) score function as

$$\mathbb{E}_\theta\left[\nabla_\theta \log f(X;\theta)\middle|Y\right] = \nabla_\theta \log g(Y;\theta) \ . \tag{4}$$

The algorithm in (3) can be shown to be locally equivalent to the EM algorithm at convergence (Lange, 1995). In practise, the step-size γ_{k+1} is often adjusted using line searches to ensure that the likelihood is indeed increased at each iteration. In addition, $J(Y_{1:n};\theta)$ is not necessarily a positive definite matrix or could be badly conditioned; therefore, some adjustment of the weight matrix $J(Y_{1:n};\theta)$ may be necessary to avoid numerical problems.

A well-known modification of the Newton recursion consists in replacing $J(Y_{1:n};\theta)$ in (3) by the Fisher Information Matrix (FIM) associated to a *complete* observation,

$$I(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_\theta\left[\nabla^2_\theta \log f(X;\theta)\right] \ . \tag{5}$$

Under the mild assumption that the complete data model is regular, $I(\theta)$ is guaranteed to be positive definite. This modified recursion, which is more robust, may be seen as an approximation of the *scoring method* (McLachlan and Krishnan, 1997), where the complete data FIM is used in place of the actual (*observed*) FIM

$$I_{\text{obs}}(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_\theta\left[\nabla^2_\theta \log g(Y;\theta)\right] \ , \tag{6}$$

despite the fact that, in general, $I_{\text{obs}}(\theta)$ and $I(\theta)$ are different. $I(\theta)$ usually also differs from $J(Y_{1:n};\theta)$, as $J(Y_{1:n};\theta)$ converges, as n tends to infinity, to

$$I_\pi(\theta) \stackrel{\text{def}}{=} -\mathbb{E}_\pi\left[\mathbb{E}_\theta\left[\nabla^2_\theta \log f(X;\theta)\middle|Y\right]\right] \ , \tag{7}$$

which doesn't correspond to a Fisher information matrix in the complete data model, except when π coincides with $f(\cdot;\theta)$. In the particular case, however, where the complete data model belongs to a

canonical (or naturally parameterised) exponential family of distributions, $J(Y_{1:n};\theta)$ coincides with $I(\theta)$ and thus does not depend on π anymore. Hence, except in some specific cases or if one assumes that the model is well-specified (i.e., $\pi = g(\cdot;\theta^\star)$), the convergence behaviour of the recursion in (3) will be different when $I(\theta)$ is used instead of $J(Y_{1:n};\theta)$.

2.2 Stochastic Gradient EM Algorithms

Being able to perform online estimation means that the data must be run through only once, which is obviously not possible with the vanilla EM algorithm. To overcome this difficulty we consider in the sequel online algorithms which produce, at a fixed computational cost, an updated parameter estimate $\hat{\theta}_n$ for each new observation Y_n. Note that in the online setting, the iteration index (which was previously denoted by k) is identical to the observation index n and we will use the latter when describing the algorithms. To our best knowledge, the first online parameter estimation procedure for latent data models is due to Titterington (1984) who proposed to use a stochastic approximation version of the modified gradient recursion:

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} I^{-1}(\hat{\theta}_n) \nabla_\theta \log g(Y_{n+1};\hat{\theta}_n) \ , \tag{8}$$

where $\{\gamma_n\}$ is a decreasing sequence of positive step sizes. One may also consider using a stochastic approximation version of the original (Newton) recursion in (3):

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} I_\pi^{-1}(\hat{\theta}_n) \nabla_\theta \log g(Y_{n+1};\hat{\theta}_n) \ . \tag{9}$$

Note that (9) does not correspond to a practical algorithm as $I_\pi(\hat{\theta}_n)$ is usually unknown, although it can be estimated, for instance, by recursively averaging over the values of $J(Y_n;\hat{\theta}_n)$. As discussed above however, this algorithm may be less robust than (8) because $J(Y_n;\hat{\theta}_n)$ is (usually) not guaranteed to be positive definite. In the following, we will refer to (8) as *Titterington's online algorithm* and to (9) as the *online gradient algorithm* (in reference to the title of the paper by Lange, 1995). Note that both of these algorithms are based on the stochastic gradient approach and bear very little resemblance with the original EM algorithm.

2.3 The Proposed Online EM Algorithm

We now consider an online approach which is more directly related to the principle underpinning the EM algorithm. The basic idea is to replace the expectation step by a stochastic approximation step, while keeping the maximisation step unchanged. More precisely, at iteration n, consider the function

$$\hat{Q}_{n+1}(\theta) = \hat{Q}_n(\theta) + \gamma_{n+1} \left(\mathbb{E}_{\hat{\theta}_n} \left[\log f(X_{n+1};\theta) \middle| Y_{n+1} \right] - \hat{Q}_n(\theta) \right) \ , \tag{10}$$

and set $\hat{\theta}_{n+1}$ as the maximum of the function $\theta \mapsto \hat{Q}_{n+1}(\theta)$ over the feasible set Θ. One important advantage of (10) compared to (8) is that it automatically satisfies the parameter constraints without requiring any further modification. In addition, (10) does not explicitly require the inversion of a $(d_\theta \times d_\theta)$ matrix. For further comparisons between both approaches, both practical and in terms of rate of convergence, we refer to the example of Section 2.4 and to the analysis of Section 3.

Of course, this algorithm is of practical interest only if it is possible to compute and maximise $\hat{Q}_n(\theta)$ efficiently. In the following we focus on the case where the complete data likelihood belongs to an *exponential family* satisfying the following assumptions. Let $\langle \cdot, \cdot \rangle$ denotes the scalar product between two vectors of $\mathbb{R}^d$ and $|\cdot|$ the associated norm.

Assumption 1. (a) *The complete data likelihood is of the form*

$$f(x;\theta) = h(x) \exp\left\{ -\psi(\theta) + \langle S(x), \phi(\theta) \rangle \right\} \ . \tag{11}$$

(b) *The function*

$$\bar{s}(y;\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta \left[S(X) \middle| Y = y \right] \ , \tag{12}$$

is well defined for all $(y,\theta) \in \mathcal{Y} \times \Theta$.

(c) *There exists a convex open subset* $\mathcal{S} \subset \mathbb{R}^d$*, which is such that*

- *for all* $s \in \mathcal{S}$, $(y,\theta) \in \mathcal{Y} \times \Theta$ *and* $\gamma \in [0,1)$, $(1-\gamma)s + \gamma \bar{s}(y;\theta) \in \mathcal{S}$.
- *for any* $s \in \mathcal{S}$*, the function* $\theta \mapsto \ell(s;\theta) \stackrel{\text{def}}{=} -\psi(\theta) + \langle s, \phi(\theta) \rangle$ *has a unique global maximum over* Θ *denoted* $\bar{\theta}(s)$*,i.e.*

$$\bar{\theta}(s) \stackrel{\text{def}}{=} \text{argmax}_{\theta \in \Theta} \ell(s;\theta) \ . \tag{13}$$

Assumption 1 implies that the evaluation of $\mathbb{E}_\theta \left[\log f(X;\theta) \middle| Y \right]$, and hence the E-step of the EM algorithm, reduces to the computation of the expected value $\mathbb{E}_\theta \left[S(X) \middle| Y \right]$ of the *complete data sufficient statistic* $S(X)$. Indeed, the EM reestimation functional $Q_{\theta'}(Y_{1:n};\theta)$ is then defined by

$$Q_{\theta'}(Y_{1:n};\theta) = \ell \left(n^{-1} \sum_{i=1}^{n} \bar{s}(Y_i;\theta');\theta \right) \ .$$

The $(k+1)$-th iteration of the (batch mode) EM algorithm may thus be expressed as

$$\hat{\theta}_{k+1} = \bar{\theta} \left(n^{-1} \sum_{i=1}^{n} \bar{s}(Y_i;\hat{\theta}_k) \right) \ , \tag{14}$$

where the M-step corresponds to the application of the function $\bar{\theta}$. Note that the construction of the set $\mathcal{S}$ in Assumption 1(c) reflects the fact that in most applications of EM, the M-step is unambiguous only when a sufficient number of observations have been gathered. This point will be illustrated in the example to be considered in Section 4 below. Assumption 1(c) takes care of this issue in the case of the online EM algorithm. As an additional comment about Assumption 1, note that we do not require that ϕ be a one to one mapping and hence the complete data model may also correspond to a *curved* exponential family, where typically θ is of much lower dimension than $\psi(\theta)$ (see, for instance, Chung and Böhme (2005); Cappé et al. (2006) for an example involving Gaussian densities with structured covariance matrices).

In this setting, the proposed online EM algorithm takes the following form

$$\begin{aligned} \hat{s}_{n+1} &= \hat{s}_n + \gamma_{n+1}(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n) \ , \\ \hat{\theta}_{n+1} &= \bar{\theta}(\hat{s}_{n+1}) \ . \end{aligned} \tag{15}$$

Algorithm of that kind have a rather long history in the machine learning community. The idea of sequentially updating the vector of sufficient statistics has apparently been first proposed by Nowlan (1991), using a fixed step size (or learning rate) $\gamma_n = \gamma$ (see also Jordan and Jacobs, 1994). The online EM algorithm (15) is also closely related to the "incremental" version of the EM algorithm derived by Neal and Hinton (1999). The incremental setting is more general than the recursive setting considered here, because the observations are not necessarily processed sequentially in time and might be used several times. The incremental EM algorithm of Neal and Hinton (1999) defines the $(k+1)$-th parameter estimate as

$$\hat{\theta}_{k+1} = \bar{\theta} \left([\min(k+1,n)]^{-1} \sum_{i=1}^{\min(k+1,n)} \hat{s}_{k+1,i} \right) \ , \tag{16}$$

where $\hat{s}_{k+1,i} = \hat{s}_{k,i}$ if $i \neq I_{k+1}$ and $\hat{s}_{k+1,I_{k+1}} = \bar{s}(Y_{I_{k+1}}; \hat{\theta}_k)$. The index I_{k+1} is typically chosen as $k+1$ while $k \leq n-1$ and runs through the data set, that is, $I_k \in \{1, \ldots, n\}$, in a fixed or pseudo random scanning order for subsequent iterations. When used in batch mode (that is when $k > n$) it is seen that it mostly differs from the traditional EM strategy in (14) by the fact that the parameters are updated after each computation of the conditional expectation of the complete data sufficient statistic corresponding to one observation. When used in online mode ($k \leq n$), the algorithm of Neal and Hinton (1999) coincides with the proposed online EM algorithm with a step-size of $\gamma_k = 1/k$ (see Section 3 for further discussion of this particular choice of step sizes). A specific instance of the proposed online EM algorithm has been derived by Sato and Ishii (2000) for maximum likelihood estimation in the so-called normalised Gaussian network; this algorithm was later extended by Sato (2000) to a canonical exponential family ($\phi(\theta) = \theta$ in (11)) and a sketch of the proof of convergence, based on stochastic approximation results, was given. The online EM algorithm defined in (15) may be seen as a generalisation of this scheme.

2.4 An Example: Poisson Mixture

Before analysing the convergence of the above algorithm, we first consider a simple example of application borrowed from Liu et al. (2006): consider the case of a mixture of m Poisson distributions

$$g(y; \theta) = \sum_{j=1}^{m} \omega_j \frac{\lambda_j^y}{y!} \mathrm{e}^{-\lambda_j} \ , \quad \text{for } y = 0, 1, 2, \ldots \ , \tag{17}$$

where the unknown parameters $\theta = (\omega_1, \ldots, \omega_m, \lambda_1, \ldots, \lambda_m)$ satisfies the constraints $\omega_j > 0$, $\sum_{i=1}^m \omega_j = 1$ and $\lambda_j > 0$. In the mixture problem, the incompleteness is caused by the ignorance of the component of the mixture. Let W be a random variable taking value in $\{1, \ldots, m\}$ with probabilities $\{\omega_1, \ldots, \omega_m\}$. The random variable W is called the regime or state and is not observable. The probability density defined in (17) corresponds to the assumption that Y is distributed, given that $W = j$, according to a Poisson random variable with parameter λ_j. Note that in this case, as in all examples which involve the simpler missing data mechanism rather than the general latent data model introduced in Section 2.1, the complete data X simply consists of the couple (Y, W) and hence conditional expectations of X given Y really boils down to expectations of W given Y.

For the Poisson mixture, the complete data log-likelihood is given by

$$\log f(y, w; \theta) = -\log(y!) + \sum_{j=1}^{m} \left[\log(\omega_j) - \lambda_j\right] \delta_{w,j} + \sum_{j=1}^{m} \log(\lambda_j) y \delta_{w,j} \ , \tag{18}$$

where $\delta_{i,l}$ is the Kronecker delta symbol: $\delta_{i,l} = 1$ if $i = l$ and $\delta_{i,l} = 0$ otherwise. The complete data likelihood may be rewritten as in (11) with $h(y, w) = -\log(y!)$, $S(y, w) = (S_1(y, w), \ldots, S_m(y, w))$ and $\phi(\theta) = (\phi_1(\theta), \ldots, \phi_m(\theta))$, where

$$S_j(y, w) \stackrel{\text{def}}{=} \begin{pmatrix} \delta_{w,j} \\ y\delta_{w,j} \end{pmatrix} \ , \qquad \text{and} \quad \phi_j(\theta) \stackrel{\text{def}}{=} \begin{pmatrix} \log(\omega_j) - \lambda_j \\ \log(\lambda_j) \end{pmatrix} \ .$$

In this case, the conditional expectation of the complete data sufficient statistics is fully determined by the posterior probabilities of the mixture components defined by

$$\bar{w}_j(y; \theta) \stackrel{\text{def}}{=} \mathbb{P}_\theta[W = j | Y = y] = \frac{\omega_j \lambda_j^y \mathrm{e}^{-\lambda_j}}{\sum_{l=1}^m \omega_l \lambda_l^y \mathrm{e}^{-\lambda_l}} \ , \quad \text{for } j = 1, \ldots, m \ .$$

The $(n+1)$-th step of the online EM algorithm consists in computing, for $j = 1, \ldots, m$,

$$\hat{s}_{j,n+1} = \hat{s}_{j,n} + \gamma_{n+1} \left\{ \begin{pmatrix} \bar{w}_j(Y_{n+1}; \hat{\theta}_n) \\ \bar{w}_j(Y_{n+1}; \hat{\theta}_n) Y_{n+1} \end{pmatrix} - \hat{s}_{j,n} \right\} ,$$
$$\hat{\omega}_{j,n+1} = \hat{s}_{j,n+1}(1) , \qquad \hat{\lambda}_{j,n+1} = \frac{\hat{s}_{j,n+1}(2)}{\hat{s}_{j,n+1}(1)} . \tag{19}$$

Comparing with the generic update equations in (15), one recognises the stochastic approximation version of the E-step, in the first line of (19), followed by the application of $\bar{\theta}$.

To compare with Titterington's online algorithm in (8), one first need to evaluate the complete Fisher information matrix $I(\theta)$. To deal with the equality constraint $\sum_{j=1}^{m} \omega_j = 1$, only the first $m-1$ weights are used as parameters and the remaining one is represented as $\omega_m = 1 - \sum_{j=1}^{m-1} \omega_j$ as in Titterington (1984). The complete data Fisher information matrix defined in (5) is then given by

$$I(\theta) = \begin{pmatrix} \mathrm{diag}(\omega_1^{-1}, \ldots, \omega_{m-1}^{-1}) + \omega_m^{-1} \mathbf{1}_{m-1} \mathbf{1}_{m-1}^T & \mathbf{0}_{(m-1)\times m} \\ \mathbf{0}_{m\times(m-1)} & \mathrm{diag}(\omega_1/\lambda_1, \ldots, \omega_m/\lambda_m) \end{pmatrix} ,$$

where the superscript T denotes transposition, $\mathbf{1}$ and $\mathbf{0}$ respectively denote a vector of ones and a matrix of zeros, whose dimensions are specified as subscript. Upon inverting $I(\theta)$, the following expression for the $(n+1)$-th step of Titterington's online algorithm is obtained :

$$\hat{\omega}_{j,n+1} = \hat{\omega}_{j,n} + \gamma_{n+1} \left(\bar{w}_j(Y_{n+1}, \hat{\theta}_n) - \hat{\omega}_{j,n} \right) ,$$
$$\hat{\lambda}_{j,n+1} = \hat{\lambda}_{j,n} + \gamma_{n+1} \frac{\bar{w}_j(Y_{n+1}; \hat{\theta}_n)}{\hat{\omega}_{j,n}} \left(Y_{n+1} - \hat{\lambda}_{j,n} \right) . \tag{20}$$

To make the connection more explicit with the update of the online EM algorithm, note that due to the fact that, in this simple case, there is an identification between some components of the vector of sufficient statistics and the weight parameters (i.e., $\bar{\theta}(s_{j,n}) = \omega_j$), it is possible to rewrite (19) in terms of the latter only:

$$\hat{\omega}_{j,n+1} = \hat{\omega}_{j,n} + \gamma_{n+1} \left(\bar{w}_j(Y_{n+1}; \hat{\theta}_n) - \hat{\omega}_{j,n} \right) ,$$
$$\hat{\lambda}_{j,n+1} = \frac{\hat{\lambda}_{j,n}\hat{\omega}_{j,n} + \gamma_{n+1} \left(\bar{w}_j(Y_{n+1}; \hat{\theta}_n) Y_{n+1} - \hat{\lambda}_{j,n}\hat{\omega}_{j,n} \right)}{\hat{\omega}_{j,n+1}} .$$

In the Poisson mixture example, the two algorithms differ only in the way the intensities of the Poisson components are updated. Whereas the online EM algorithm in (19) does ensure that all parameter constraints are satisfied, it may happen, in contrast, that (20) produces negatives values for the intensities. Near convergence however, the two algorithms behave very similarly in this simple case (see Proposition 6 below).

2.5 Extensions

As previously mentioned, Neal and Hinton (1999) advocate the use of online algorithms also in the case of batch training with large sample sizes. The online algorithm then operates by repeatedly scanning through the available sample. In our setting, this use of the proposed online EM algorithm may be analysed by letting π denote the empirical measure associated with the fixed sample $X_1, \ldots, X_n$. The results to follow thus also apply in this context, at least when the data scanning order is random.

In semi-parametric regression models, each observation Y comes with a vector of covariates Z whose distribution is usually unspecified and treated as a nuisance parameter. To handle latent data versions of regression models (mixture of regressions, mixture of experts, etc.—see Grün and Leisch,

2007; Jordan and Jacobs, 1994, as well as the example of Section 4) in our framework, one only needs to assume that the model consists of a parametric family $\{f(x|z;\theta), \theta \in \Theta)\}$ of *conditional* pdfs. In this setting however, it is not possible anymore to compute expectations under the complete data distribution and the model can never be well-specified, as the distribution of Z is left unspecified. Thus Titterington's algorithm in (8) does not directly apply in this setting. In contrast, the proposed algorithm straightforwardly extends to this case by considering covariate-dependent expectations of the sufficient statistics of $f(x|z;\theta)$, of the form $\bar{s}(y,z;\theta) = \mathbb{E}_\theta[S(X)|Y=y, Z=z]$, instead of (12). For notational simplicity, we state our results in the following section without assuming the presence of covariates but extension to the case where there are covariates is straightforward; the example of Section 4 corresponds to a case where covariates are available.

3 Convergence Issues

3.1 Consistency

In this section, we establish the convergence of the proposed algorithm towards the set of stationary points of the Kullback-Leibler divergence between the actual observation density and the model likelihood. These results are the analogues of those given by Wang and Zhao (2006) for Titterington's online algorithm, with a somewhat broader scope since we do not assume that the model is well-specified. The proofs corresponding to this section are given in Appendix A. In addition to the conditions listed in Assumption 1, we will require the following additional regularity assumptions.

Assumption 2. **(a)** *The parameter space Θ is a convex open subset of $\mathbb{R}^{d_\theta}$ and ψ and ϕ in (11) are twice continuously differentiable on Θ.*

(b) *The function $s \mapsto \bar{\theta}(s)$, defined in* (13)*, is continuously differentiable on $\mathcal{S}$,*

(c) *For some $p > 2$, and all compact subsets $\mathcal{K} \subset \mathcal{S}$,*

$$\sup_{s\in\mathcal{K}} \mathbb{E}_\pi \left(\left| \bar{s}(Y;\bar{\theta}(s)) \right|^p \right) < \infty \ .$$

To analyse the recursion (15), the first step consists in expressing it as a standard Robbins-Monro stochastic approximation procedure operating on the complete data sufficient statistics:

$$\hat{s}_{n+1} = \hat{s}_n + \gamma_{n+1} \left(\mathrm{h}(\hat{s}_n) + \xi_{n+1} \right) \ , \tag{21}$$

where $\mathrm{h} : \mathcal{S} \to \mathbb{R}^{d_\theta}$ is the so-called *mean field* given by

$$\mathrm{h}(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\bar{s}(Y;\bar{\theta}(s)) \right] - s \ , \tag{22}$$

and $\{\xi_n\}_{n\geq 1}$ is a sequence of random variables representing stochastic perturbations defined by

$$\xi_{n+1} \stackrel{\text{def}}{=} \bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)) - \mathbb{E}\left[\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)) \,\middle|\, \mathcal{F}_n \right] \ , \tag{23}$$

where $\mathcal{F}_n$ is the σ-field generated by $(\hat{s}_0, \{Y_i\}_{i=1}^n)$. The aim of the Robbins-Monro procedure (21) is to solve the equation $\mathrm{h}(s) = 0$. As a preliminary step, we first characterise the set of roots of the mean field h. The following proposition shows that, if $s^\star$ belongs to

$$\Gamma \stackrel{\text{def}}{=} \{ s \in \mathcal{S} : \mathrm{h}(s) = 0 \} \ , \tag{24}$$

then $\theta^\star = \bar{\theta}(s^\star)$ is a stationary point of the Kullback-Leibler divergence between π and g_θ,

$$\mathrm{K}(\pi \,\|\, g_\theta) \stackrel{\text{def}}{=} \mathbb{E}_\pi \left[\log \left(\frac{\pi(Y)}{g(Y;\theta)} \right) \right] \ . \tag{25}$$

Proposition 3. *Under Assumptions 1–2, if $s^\star \in \mathcal{S}$ is a root of* h*, i.e.,* $\mathrm{h}(s^\star) = 0$*, then* $\theta^\star = \bar{\theta}(s^\star)$ *is a stationary point of the function* $\theta \mapsto \mathrm{K}(\pi \| g_\theta)$*, i.e.,* $\nabla_\theta \mathrm{K}(\pi \| g_\theta)|_{\theta=\theta^\star} = 0$*. Conversely, if* $\theta^\star$ *is a stationary point of* $\theta \mapsto \mathrm{K}(\pi \| g_\theta)$*, then* $s^\star = \mathbb{E}_\pi[\bar{s}(Y;\theta^\star)]$ *is a root of* h*.*

We then show that the function $\mathrm{w} : \mathcal{S} \to [0, \infty)$ defined by

$$\mathrm{w}(s) \stackrel{\text{def}}{=} \mathrm{K}\left(\pi \left\| g_{\bar{\theta}(s)}\right.\right) , \tag{26}$$

is a *Lyapunov function* for the mean field h and the set Γ, i.e. for any $s \in \mathcal{S}$, $\langle \nabla_s \mathrm{w}(s), \mathrm{h}(s)\rangle \leq 0$ and $\langle \nabla_s \mathrm{w}(s), \mathrm{h}(s)\rangle = 0$ if and only if $\mathrm{h}(s) = 0$. The existence of a Lyapunov function is a standard argument to prove the global asymptotic stability of the solutions of the Robbins-Monro procedure. This property can be seen as an analog of the monotonicity property of the EM algorithm: each unperturbed iteration $\bar{s}_{k+1} = \bar{s}_k + \gamma_{k+1}\mathrm{h}(\bar{s}_k)$ decreases the Kullback-Leibler divergence to the target distribution π, provided that γ_{k+1} is small enough.

Proposition 4. *Under Assumptions 1–2,*

- $\mathrm{w}(s)$ *is continuously differentiable on* $\mathcal{S}$*,*
- *for any compact subset* $\mathcal{K} \subset \mathcal{S} \setminus \Gamma$*,*

$$\sup_{s\in\mathcal{K}} \langle \nabla_s \mathrm{w}(s), \mathrm{h}(s)\rangle < 0 .$$

Using this result, we may now prove the convergence of the sequence $\{\hat{s}_k\}$. Denote by $\mathcal{L} = \{\theta \in \Theta : \nabla_\theta \mathrm{K}(\pi \| g_\theta) = 0\}$ the set of stationary points of the Kullback-Leibler divergence, and, for $x \in \mathbb{R}^m$ and $A \subset \mathbb{R}^m$, let $d(x, A) = \inf\{y \in A, |x - y|\}$.

Theorem 5. *Assume 1–2 and that, in addition,*

1. $0 < \gamma_i < 1$, $\sum_{i=1}^\infty \gamma_i = \infty$ *and* $\sum_{i=1}^\infty \gamma_i^2 < \infty$*,*
2. $\hat{s}_0 \in \mathcal{S}$ *and with probability one,* $\limsup |\hat{s}_n| < \infty$ *and* $\liminf d(\hat{s}_n, \mathcal{S}^c) > 0$*.*
3. *The set* $\mathrm{w}(\Gamma)$ *is nowhere dense.*

Then, $\lim_{n\to\infty} d(\hat{s}_n, \Gamma) = 0$ *and* $\lim_{n\to\infty} d(\hat{\theta}_n, \mathcal{L}) = 0$*, with probability one.*

The first condition of Theorem 5 is standard for decreasing step-size stochastic approximation procedures (Kushner and Yin, 1997). It is satisfied for instance by setting $\gamma_i = \gamma_0 i^{-\alpha}$, with $\alpha \in (1/2, 1]$. The additional requirements that γ_i be less than 1 and $\hat{s}_0$ be chosen in $\mathcal{S}$ is just meant to ensure that the whole sequence $\{\hat{s}_k\}$ stays in $\mathcal{S}$ (see Assumption 1(c)). The rest of the second assumption of Theorem 5 correspond to a stability assumption which is not trivial. In general settings, the stability of the algorithm can be enforced by truncating the algorithm updates, either on a fixed set (see, e.g., Kushner and Yin, 2003, chapter 2) or on an expanding sequence of sets (see, e.g., Chen, 2002, chapter 2, or Andrieu et al., 2005). We do not explicitly carry out these constructions here to keep the exposition concise.

3.2 Rate of Convergence

In this section, we show that when approaching convergence, the online EM algorithm is comparable to the online gradient algorithm in (9). The existence of such links is hardly surprising, in view of the discussions in Section 4 of Titterington (1984) and Section 3 of Lange (1995), and may be seen as a counterpart, for stochastic approximation, of the asymptotic equivalence of the gradient EM algorithm of Lange (1995) and the EM algorithm. To highlight these relations, we first express the online EM algorithm as a stochastic approximation procedure on θ.

Proposition 6. *Under the assumptions of Theorem 5, the online EM sequence* $\{\hat{\theta}_n\}_{n\geq 0}$ *given by* (15) *follows the recursion*

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \gamma_{n+1} I_\pi^{-1}(\hat{\theta}_n)\nabla_\theta \log g(Y_{n+1};\hat{\theta}_n) + \gamma_{n+1}\rho_{n+1} \tag{27}$$

where $\lim_{n\to\infty}\rho_n = 0$ *a.s. and* $I_\pi(\theta)$ *is defined in* (7).

Hence, the online EM algorithm is equivalent, when approaching convergence, to the online gradient algorithm defined in (9) which coincides with Titterington's online algorithm with $I_\pi(\hat{\theta}_n)$ substituted for $I(\hat{\theta}_n)$. It is remarkable that the online EM algorithm can achieve a convergence performance similar to that of the online gradient algorithm without explicit matrix approximation nor inversion. Note that, as previously discussed in Section 2.1, in the particular case of canonical exponential families, $I_\pi(\theta)$ and $I(\theta)$ coincide and the proposed online EM algorithm is thus also equivalent (near convergence) to Titterington's online algorithm.

Although the recursion (27) will not lead to asymptotic efficiency, we can, under appropriate additional conditions, guarantee $\gamma_n^{-1/2}$-consistency and asymptotic normality. We use the weak convergence result presented in Pelletier, 1998, Theorem 1.

Theorem 7. *Under the assumptions of Theorem 5, let* $\theta^\star$ *be a (possibly local) minimum of the Kullback-Leibler divergence:* $\theta \mapsto \mathrm{K}(\pi \,\|g_\theta)$. *Denote by*

$$H(\theta^\star) \stackrel{\text{def}}{=} I_\pi^{-1}(\theta^\star)\left[-\nabla_\theta^2 \mathrm{K}(\pi\,\|g_\theta)|_{\theta=\theta^\star}\right] ,$$
$$\Gamma(\theta^\star) \stackrel{\text{def}}{=} I_\pi^{-1}(\theta^\star)\,\mathbb{E}_\pi\left(\nabla_\theta \log g(Y;\theta^\star)\left\{\nabla_\theta \log g(Y;\theta^\star)\right\}^T\right) I_\pi^{-1}(\theta^\star) .$$

Then,

1. $H(\theta^\star)$ *is a stable matrix whose eigenvalues have their real part upper bounded by* $-\lambda(\theta^\star)$, *where* $\lambda(\theta^\star) > 0$.

2. *Let* $\gamma_n = \gamma_0 n^{-\alpha}$, *where* γ_0 *may be chosen freely in* $(0,1)$ *when* $\alpha \in (1/2,1)$ *but must satisfy* $\gamma_0 > \lambda(\theta^\star)$ *when* $\alpha = 1$; *then, on the event* $\Omega(\theta^\star) = \{\lim_{n\to\infty}\hat{\theta}_n = \theta^\star\}$, *the sequence* $\gamma_n^{-1/2}\left(\hat{\theta}_n - \theta^\star\right)$ *converges in distribution to a zero mean Gaussian distribution with covariance* $\Sigma(\theta^\star)$, *where* $\Sigma(\theta^\star)$ *is the solution of the Lyapunov equation*

$$\left(H(\theta^\star) + \zeta \mathrm{Id}\right)\Sigma(\theta^\star) + \Sigma(\theta^\star)\left(H^T(\theta^\star) + \zeta\mathrm{Id}\right) = -\Gamma(\theta^\star) , \tag{28}$$

where $\zeta = 0$ *if* $\alpha \in (1/2,1)$ *and* $\zeta = \gamma_0^{-1}$ *if* $\alpha = 1$, *and,* Id *denotes the identity matrix.*

Solving equation (28) is easy for a well-specified model, i.e., when $\pi = g_{\theta^\star}$, as the FIM $I_{\text{obs}}(\theta^\star)$ that is associated with the (observed) data model then satisfies

$$\begin{aligned} I_{\text{obs}}(\theta^\star) &= -\mathbb{E}_{\theta^\star}\left[\nabla_\theta^2 \log g(Y;\theta^\star)\right] \\ &= \nabla_\theta^2 \mathrm{K}(g_{\theta^\star}\,\|g_\theta)|_{\theta=\theta^\star} = \mathbb{E}_\pi\left(\nabla_\theta \log g(Y;\theta^\star)\left\{\nabla_\theta \log g(Y;\theta^\star)\right\}^T\right) . \end{aligned}$$

When $\zeta = 0$, the solution of the Lyapunov equation is thus given by $\Sigma(\theta^\star) = I_\pi^{-1}(\theta^\star)/2$. The model being well-specified also implies that $I_\pi(\theta^\star) = I(\theta^\star)$, and, hence, the asymptotic covariance matrix is given by half the inverse of the complete data FIM in this case. When $\zeta \neq 0$, the Lyapunov equation cannot be solved in explicitly, except when the parameter is scalar (the result then coincides with Titterington, 1984, Theorem 1). Note that using $\gamma_n = \gamma_0 n^{-\alpha}$ with $\alpha = 1$ provides the optimal convergence rate of $1/\sqrt{n}$ but only at the price of a constraint on the scale γ_0, which is usually impossible to check in practice. On the other hand, using $\alpha \in (1/2,1)$ results in a slower convergence

rate but without constraint on the scale γ_0 of the step-size (except for the fact that it has to be smaller than 1).

To circumvent this difficulty, we recommend to use the so-called Polyak-Ruppert averaging technique (Polyak, 1990; Ruppert, 1988) as a post-processing step. Following Polyak (1990) —see also Polyak and Juditsky (1992); Mokkadem and Pelletier (2006)—, if $\gamma = \gamma_0 n^{-\alpha}$, with $\alpha \in (1/2, 1)$, then the running average

$$\tilde{\theta}_n \stackrel{\text{def}}{=} n^{-1} \sum_{j=n_0}^{n} \hat{\theta}_j \ , \qquad n \geq 1 \tag{29}$$

converges at rate $1/\sqrt{n}$, for all values of γ_0. Furthermore, on the event $\Omega(\theta^\star)$ defined in Theorem 7 above, $\sqrt{n}(\tilde{\theta}_n - \theta^\star)$ is asymptotically normal, with asymptotic covariance matrix

$$\overline{\Sigma}(\theta^\star) = H^{-1}(\theta^*)\Gamma(\theta^\star)H^{-1}(\theta^\star) = \\ \left[-\nabla_\theta^2 \, \mathrm{K}\left(\pi \, \|g_\theta\right)|_{\theta=\theta^\star}\right]^{-1} \pi \left(\nabla_\theta \log g_{\theta^\star} \left\{\nabla_\theta \log g_{\theta^\star}\right\}^T\right) \left[-\nabla_\theta^2 \, \mathrm{K}\left(\pi \, \|g_\theta\right)|_{\theta=\theta^\star}\right]^{-1} \ , \tag{30}$$

which is known to be optimal (Kushner and Yin, 1997). If $\pi = g_{\theta^\star}$, the previous result shows that the averaged sequence $\tilde{\theta}_n$ is an asymptotically efficient sequence of estimates of $\theta^\star$, i.e. the asymptotic covariance of $\sqrt{n}(\tilde{\theta}_n - \theta^\star)$ is equal to the inverse of the (observed data) FIM $I_{\text{obs}}(\theta^\star)$.

4 Application to Mixtures of Gaussian Regressions

To illustrate the performance of the proposed method, we consider a regression model which, as discussed in Section 2.5, corresponds to a case where the complete data FIM is not available. In contrast, we illustrate below the fact that the proposed algorithm, without explicitly requesting the determination of a weighting matrix does provide asymptotically efficient parameter estimates when Polyak-Ruppert averaging is used.

The model we consider is a finite mixture of Gaussian linear regressions, where the complete data consists of the response variable R, here assumed to be scalar for simplicity, the d_z-dimensional vector Z that contains the explanatory variables, and, W which corresponds, as in the example of Section 2.4, to a latent mixture indicator taking its value in the finite set $\{1, \ldots, m\}$. We assume that given $W = j$ and Z, R is distributed as a $\mathcal{N}(\beta_j^T Z, \sigma_j^2)$ Gaussian variable, while W is independent of Z and such that $\mathbb{P}_\theta(W = j) = \omega_j$. Thus the parameters of the model are the mixture weights ω_j and the regression vectors β_j and variances σ_j^2, for $j = 1, \ldots, m$. As is usually the case in conditional regression models, we specify only the part of the complete data likelihood that depends on the parameters, without explicitly modelling the marginal distribution of the vector of regressors Z. In terms of our general notations, the complete data X is the triple (R, Z, W), the observed data is the couple (R, Z) and the model is not well-specified, in the sense that the distribution of the observation (R, Z) is not fully determined by the model. We refer to Hurn et al. (2003) or Grün and Leisch (2007) and references therein for more information on mixture of regression models and their practical use.

In the mixture of Gaussian regressions model, the part of the complete data log-likelihood that depends on the parameters may be written as

$$\log f(r, w, z; \theta) = \sum_{j=1}^{m} \left\{ \log(\omega_j) - \frac{1}{2} \left[\log \sigma_j^2 + \frac{\left(r - \beta_j^T z\right)^2}{\sigma_j^2} \right] \right\} \delta_{w,j} \ , \tag{31}$$

where δ denotes, as before, the Kronecker delta. To put (31) in the form given in (11), one needs to

define the statistics $S = (S_{1,j}, S_{2,j}, S_{3,j}, S_{4,j})_{1\leq j\leq m}$ where

$$\begin{aligned}
S_{1,j}(r,w,z) &= \delta_{w,j} && \text{(scalar)} \ , \\
S_{2,j}(r,w,z) &= \delta_{w,j} r z && (d_z \times 1) \ , \\
S_{3,j}(r,w,z) &= \delta_{w,j} z z^T && (d_z \times d_z) \ , \\
S_{4,j}(r,w,z) &= \delta_{w,j} r^2 && \text{(scalar)} \ .
\end{aligned} \tag{32}$$

As in the simple Poisson mixture example of Section 2.4, the E-step statistics only depend on the conditional expectation of the indicator variable W through

$$\begin{aligned}
\bar{s}_{1,j}(r,z;\theta) &= \bar{w}_j(r,z;\theta) \ , \\
\bar{s}_{2,j}(r,z;\theta) &= \bar{w}_j(r,z;\theta) r z \ , \\
\bar{s}_{3,j}(r,z;\theta) &= \bar{w}_j(r,z;\theta) z z^T \ , \\
\bar{s}_{4,j}(r,z;\theta) &= \bar{w}_j(r,z;\theta) r^2 \ ,
\end{aligned} \tag{33}$$

where

$$\bar{w}_j(r,z;\theta) \stackrel{\text{def}}{=} \mathbb{E}_\theta[W = j | R = r, Z = z] = \frac{\frac{\omega_j}{\sigma_j} \exp\left[-\frac{1}{2}\frac{(r-\beta_j^T z)^2}{\sigma_j^2}\right]}{\sum_{l=1}^m \frac{\omega_l}{\sigma_l} \exp\left[-\frac{1}{2}\frac{(r-\beta_l^T z)^2}{\sigma_l^2}\right]} \ .$$

Finally, it is easily checked that that the M-step is equivalent to an application of the function $\bar{\theta} : s \mapsto \left(\bar{\omega}_j(s), \bar{\beta}_j(s), \bar{\sigma}_j(s)\right)_{1\leq j\leq m}$ where

$$\begin{aligned}
\bar{\omega}_j(s) &= s_{1,j} \ , \\
\bar{\beta}_j(s) &= s_{3,j}^{-1} s_{2,j} \ , \\
\bar{\sigma}_j^2(s) &= \left(s_{4,j} - \bar{\beta}_j^T(s) s_{2,j}\right) / s_{1,j} \ .
\end{aligned} \tag{34}$$

In this example, the role played by the set $\mathcal{S}$ in Assumption 1(c) is important: In order to apply (34), its required that the scalars $s_{1,j}$ belong to the open set $(0,1)$ and that the (d_z+1)-dimensional matrices block-defined by

$$M_j = \begin{pmatrix} s_{3,j} & s_{2,j} \\ s_{2,j}^T & s_{4,j} \end{pmatrix} \ ,$$

be positive definite, since $\bar{\sigma}_j^2(s)$ is, up to normalisation by $s_{1,j}$, the Schur complement of M_j. These constraints, for $j = 1, \ldots, m$ define the set $\mathcal{S}$ which is indeed open and convex. The function $\bar{s}$ defined in (33) however *never* produces values of s which are in $\mathcal{S}$. In particular, $\bar{s}_{3,j}(r,z;\theta)$ is a rank one matrix which is not invertible (unless $d_z = 1$). Hence the importance of using an initialisation s_0 which is chosen in $\mathcal{S}$. For the simulations below, we took care of this issue by inhibiting the parameter re-estimation step in (34) for the first twenty observations of each run. In other words, the first twenty observations are used only to build a up a value of $\hat{s}_{20}$, using the first line of (15), which is in $\mathcal{S}$ with great probability.

For illustration purpose, we consider a variation of a simple simulation example used in the `flexmix` `R` package (Leisch, 2004), where $m = 2$, $\omega_1 = \omega_2 = 0.5$, and

$$R = \begin{cases} 5U + V & \text{(when } W = 1) \\ 15 + 10\,U - U^2 + V & \text{(when } W = 2) \end{cases} \ ,$$

where $U \sim \text{Unif}(0,10)$ and $V \sim \mathcal{N}(0, 9^2)$. In order to balance the asymptotic variances of the regression parameters (see below) we used $Z^T = (1, U, U^2/10)$ as the vector of regressors, hence the actual

value of the regression parameter is $\beta_1^T = (0, 5, 0)$ for the first component and $\beta_2^T = (15, 10, -10)$. The corresponding data is shown in Figure 1 where the points corresponding to both classes are plotted differently for illustration purpose, despite the fact that only unsupervised estimation is considered here. The labelling is indeed rather ambiguous in this case as the posterior probability of belonging to one of the two classes is between 0.25 and 0.75 for about 40% of the points.

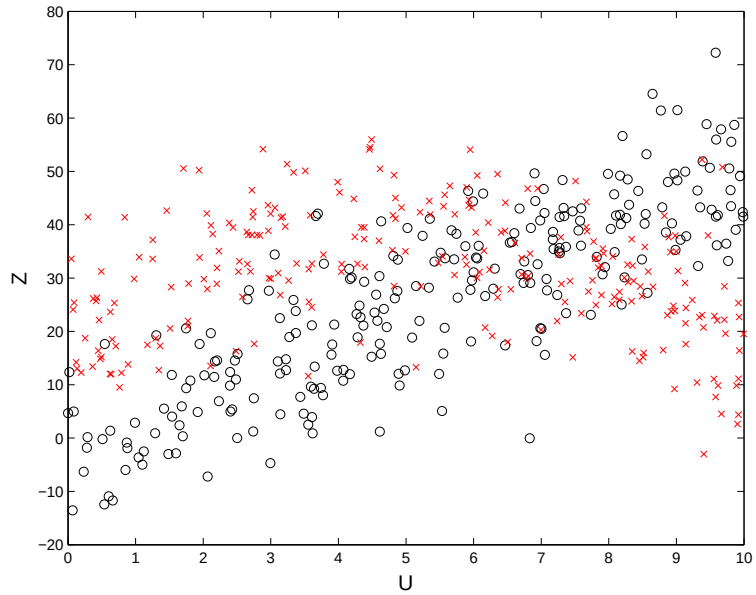

Figure 1: 500 points drawn from the model: circles, points drawn from the first class, and, crosses, points drawn from the second class (the algorithms discussed here ignore the class labels).

Clearly, the mixture of regressions model is such that the associated complete data sufficient likelihood has the form given in Assumption 1(a), where the marginal density of the explanatory variables in Z appears in the term $h(x)$ since it does not depend on the parameters. Hence the previous theory applies straightforwardly and the online EM algorithm may be used to maximise the conditional likelihood function of the responses R given the regressors Z. However, the explicit evaluation of the complete data FIM $I(\theta)$ defined in (5) is not an option here because the model does not specify the marginal distribution of Z. Titterington's online algorithm may thus not be used directly. Applying the recursion in (8) without a weighting matrix is not recommended here as the regression parameters are greatly correlated due to the non-orthogonality of the regressors.

In order to determine a suitable weighting matrix, one can use Fisher's relation (4) which gives, for the regression parameters,

$$\begin{aligned}\nabla_{\beta_j} \log g(r|z;\theta) &= \mathbb{E}_\theta \left[\nabla_{\beta_j} \log f(R, W|Z;\theta) \middle| R = r, Z = z; \theta \right] \\ &= \mathbb{E}_\theta \left[\delta_{W,j} \frac{(R - \beta_j^T Z) Z}{\sigma_j^2} \middle| R = r, Z = z; \theta \right] = \bar{w}_j(r, z; \theta) \frac{(r - \beta_j^T z) z}{\sigma_j^2} \ .\end{aligned}$$

Hence, if we assume that the model is well-specified, the (observed) FIM $I_{\text{obs}}(\theta)$ may be approximated, near convergence, by computing empirical averages of the form

$$1/n \sum_{i=1}^{n} \nabla_{\beta_j} \log g(R_i|Z_i;\theta) \left\{ \nabla_{\beta_j} \log g(R_i|Z_i;\theta) \right\}^T \ .$$

As the online EM algorithm does not require such computations, this estimate has been used only to determine the FIM at the actual parameter value for comparison purpose. It is easily checked that

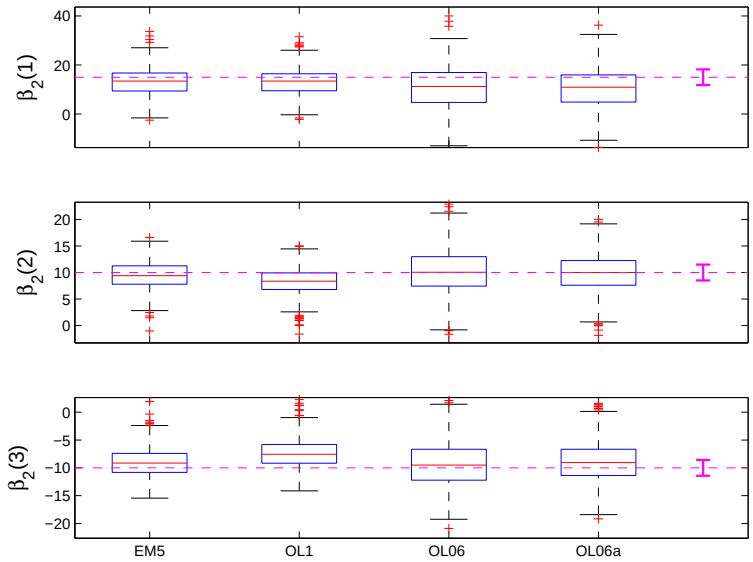

Figure 2: Box-and-whisker plots of the three components of β_2 (from top to bottom) estimated from 500 independent runs of length $n = 100$ for, EM5: five iterations of the batch EM algorithm, OL1: online EM algorithm with $\gamma_i = 1/i$, OL06: online EM algorithm with $\gamma_i = 1/i^{0.6}$, OL06a: online EM algorithm with $\gamma_i = 1/i^{0.6}$ and averaging started from the 50th iteration. The horizontal dashed line corresponds to the actual parameter value and the interval in bold at the right of each plot to the interquartile range deduced from the asymptotic normal approximation.

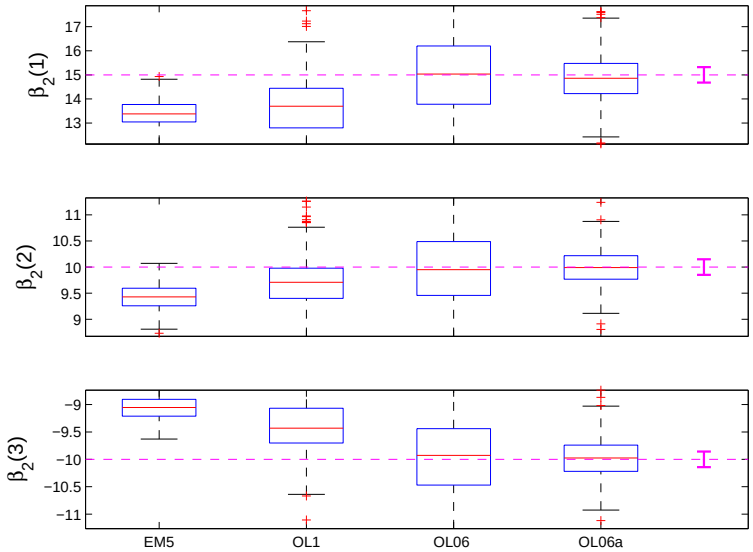

Figure 3: Same plots as in Figure 2 for signals of length $n = 10,000$ (OL06a uses averaging started from the $5,000$th iteration).

due to the linearity of the model and the fact that both components have equal weights and variance, the covariance matrices for β_1 and β_2 are the same. The numerical approximation determined from a million simulated observations yields asymptotic standard deviations of $(47.8, 22.1, 21.1)$ for the coordinates of β_j, with an associated correlation matrix of

$$\begin{pmatrix} 1 & -0.87 & 0.75 \\ -0.87 & 1 & -0.97 \\ 0.75 & -0.97 & 1 \end{pmatrix} .$$

As noted above, the coordinates of the regression vector are very correlated which would make the unweighted parameter-space stochastic approximation algorithm (i.e., (8) with an identity matrix instead of $I^{-1}(\hat{\theta}_n)$) very inefficient.

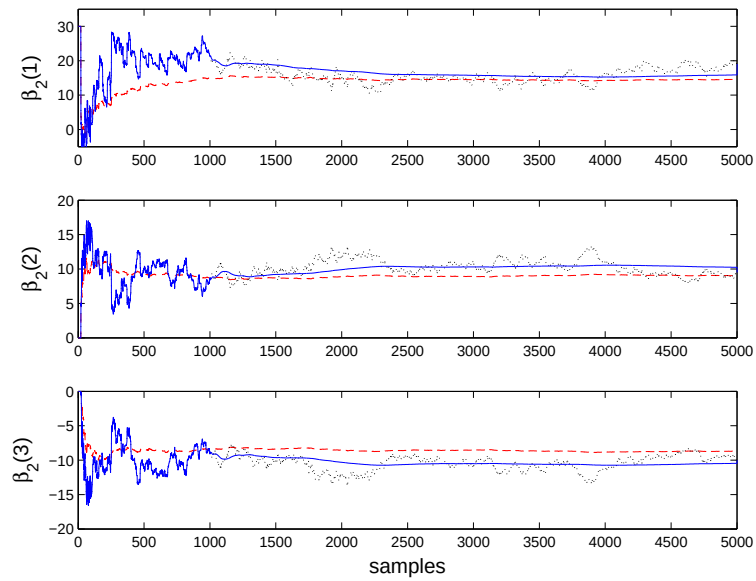

Figure 4: Example of parameters trajectories for the three components of β_2 (from top to bottom) for a signal of length $n = 5,000$: OL1, dashed line, OL06 doted line, OL06a solid line (with averaging started after $1,000$ iterations.

For run-lengths of $n = 100$ and $n = 10,000$ observations, we illustrate the performance of the following four algorithmic options :

EM5 Five iterations of the batch EM algorithm, using the whole data.

OL1 The online EM algorithm with step size $\gamma_n = 1/n$.

OL06 The online EM algorithm with step size $\gamma_n = n^{-0.6}$.

OL06a The online EM algorithm with step size $\gamma_n = n^{-0.6}$ and averaging started from $n_0 = n/2$ according to (29).

Note that whereas OL1 and OL06a have the same computational complexity (as the averaging post-processing has a negligible cost), EM5 is significantly more costly requiring five times as many E-step computations; it is also non-recursive. All algorithm are started from the same point and run for 500 independent simulated replicas of the data. The results (for β_2) are summarised as box-and-whisker plots in Figure 2, for $n = 100$, and Figure 3 for $n = 10,000$. Comparing both figures, one observes that OL06a is the only approach which appears to be consistent with a variance compatible with the asymptotic interquartile range shown on the right of each plot. EM5 (five iterations of batch

EM) is clearly the method which has the less variability but Figure 3 suggests that it is not $1/\sqrt{n}$ consistent, which was indeed confirmed using longer runs not shown here. This observation supports the claim of Neal and Hinton (1999) that, for large sample sizes, online EM approaches are more efficient, from a computational point of view, than the batch EM algorithm which requires several iterations to converge properly. The online EM with step size $\gamma_n = 1/n$ (OL1) presents a bias which becomes very significant when n increases. According to Theorem 7, this problem could be avoided (asymptotically) by choosing a sufficiently small value of γ_0. For fixed n however, lowering γ_0 can only reduce the perceived speed of convergence, which is already very slow, as illustrated by Figure 4. In contrast, the online EM algorithm with Polyak-Ruppert averaging (OL06a) appears to be very efficient: averaging significantly reduces the variability of the OL06 estimate, reducing it to a level which is consistent with the asymptotic interquartile range, while maintaining a systematic bias which vanishes as n increases, as expected.

5 Conclusion

Compared to other alternatives, the main advantages of the proposed approach to online parameter estimation in latent data models are its analogy with the standard batch EM algorithm, which makes the online algorithm easy to implement, and its provably optimal convergence behaviour. In addition, the combination of a slow parameter decrease ($\gamma_n = n^{-0.5+\epsilon}$ being a typical choice) with Polyak-Ruppert averaging appears to be very robust.

A limitation is the fact that the function $\bar{\theta}(s)$ has to be explicit, which, for instance, would not be the case for mixture of regression models with generalised link functions. Another extension of interest concerns non independent models and in particular hidden Markov models or Markov random fields.

References

C. Andrieu, É. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44(1):283–312 (electronic), 2005.

O. Cappé, M. Charbit, and E. Moulines. Recursive EM algorithm with applications to DOA estimation. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Toulouse, France, may 2006.

H.-F. Chen. *Stochastic approximation and its applications*, volume 64 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 2002.

P.-J. Chung and F. B. Böhme. Recursive EM and SAGE-inspired algorithms with application to DOA estimation. *IEEE Trans. Signal Process.*, 53(8):2664–2677, 2005.

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

B. Grün and F. Leisch. Fitting finite mixtures of generalized linear regressions in R. *Comput. Statist. Data Anal.*, 51(11):5247–5252, July 2007. doi: 10.1016/j.csda.2006.08.014.

M. Hurn, A. Justel, and C. P. Robert. Estimating mixtures of regressions. *Comput. Statist. Data Anal.*, 12(1):55–79, 2003. ISSN 1061-8600.

M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6:181–214, 1994.

H. J. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, 1997.

H. J. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer, New York, 2nd edition, 2003.

K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 57(2):425–437, 1995.

F. Leisch. FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8):1–18, 2004.

Z. Liu, J. Almhana, V. Choulakian, and R. McGorman. Online EM algorithm for mixture with application to internet traffic modeling. *Comput. Statist. Data Anal.*, 50(4):1052–1071, 2006.

G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.

A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(3):1671–1702, 2006. ISSN 1050-5164.

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

S. Nowlan. *Soft competitive adaptation: neural network learning algorithms based on fitting statistical mixtures.* PhD thesis, School of Computer Science, Carnegie Mellon University, 1991.

M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab.*, 8(1):10–44, 1998. ISSN 1050-5164.

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

B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control Optim.*, 30(4):838–855, 1992.

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

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

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

M. A. Tanner. *Tools for Statistical Inference*. Springer, New York, 2nd edition, 1993.

D. M. Titterington. Recursive parameter estimation using incomplete data. *J. Roy. Statist. Soc. Ser. B*, 46(2):257–267, 1984.

D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, Chichester, 1985.

S. Wang and Y. Zhao. Almost sure convergence of Titterington's recursive estimator for mixture models. *Statist. Probab. Lett.*, 76:2001–2006, 2006.

C. F. J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11:95–103, 1983.

A Proofs

For $\mu = (\mu_1, \ldots, \mu_m)^T$ a differentiable function from Θ to $\mathbb{R}^m$, we define by $\nabla_\theta \mu^T$ the $d_\theta \times m$ matrix whose columns are the gradients, $\nabla_\theta \mu^T = [\nabla_\theta \mu_1, \ldots, \nabla_\theta \mu_m]$. Thus the symbol ∇_θ denotes either a vector or a matrix, depending on whether the function to which it is applied is scalar or vector-valued. With this convention, the matrix $\nabla_\theta \mu^T$ is the transpose of the usual *Jacobian matrix*.

Proof of Proposition 3. Let $s^\star$ be a root of h, and set $\theta^\star = \bar{\theta}(s^\star)$. For any $s \in \mathcal{S}$, the function $\theta \mapsto \ell(s;\theta)$ has a unique stationary point at $\bar{\theta}(s)$. Therefore,

$$-\nabla_\theta \psi(\theta^\star) + \nabla_\theta \phi^T(\theta^\star) s^\star = 0 \,. \tag{35}$$

Note that, from *Fisher's identity*,

$$\nabla_\theta \log g(y;\theta) = \mathbb{E}_\theta \left[\nabla_\theta \log f(X;\theta) | \, Y = y \right] = -\nabla_\theta \psi(\theta) + \nabla_\theta \phi^T(\theta) \bar{s}(y;\theta) \,. \tag{36}$$

As $\nabla_\theta \mathrm{K}(\pi \| g_\theta) = \mathbb{E}_\pi \left[\nabla_\theta \log g(Y;\theta) \right]$, the latter relation implies

$$\nabla_\theta \mathrm{K}(\pi \| g_\theta) = -\nabla_\theta \psi(\theta) + \nabla_\theta \phi^T(\theta) \mathbb{E}_\pi \left[\bar{s}(Y;\theta) \right] \,. \tag{37}$$

Since $\mathrm{h}(s^\star) = 0$, $s^\star = \mathbb{E}_\pi[\bar{s}(Y;\theta^\star)]$, and thus (35) and (37) imply

$$\nabla_\theta \mathrm{K}(\pi \| g_\theta)|_{\theta=\theta^\star} = -\nabla_\theta \psi(\theta^\star) + \nabla_\theta \phi^T(\theta^\star) s^\star = 0 \,,$$

establishing the first assertion. Conversely, suppose that $\nabla_\theta \mathrm{K}(\pi \| g_\theta)|_{\theta=\theta^\star} = 0$ and set $s^\star = \mathbb{E}_\pi[\bar{s}(Y;\theta^\star)]$. By (37),

$$-\nabla_\theta \psi(\theta^\star) + \nabla_\theta \phi^T(\theta^\star) s^\star = 0 \,.$$

Under Assumption 1(c), the function $\theta \mapsto -\psi(\theta) + \langle \phi(\theta), s^\star \rangle$ has a unique stationary point at $\bar{\theta}(s^\star)$, which is a maximum. Hence, $\theta^\star = \bar{\theta}(s^\star)$ establishing the second assertion. □

Proof of Proposition 4. Using (37) and the chain rule of differentiation,

$$\nabla_s \mathrm{w}(s) = -\nabla_s \bar{\theta}^T(s) \left\{ -\nabla_\theta \psi(\bar{\theta}(s)) + \nabla_\theta \phi^T(\bar{\theta}(s)) \mathbb{E}_\pi \left[\bar{s}(Y;\bar{\theta}(s)) \right] \right\} \,. \tag{38}$$

For any $s \in \mathcal{S}$, $\bar{\theta}(s)$ is the maximum of $\theta \mapsto \ell(s;\theta)$, thus,

$$\nabla_\theta \ell(s;\theta)|_{\theta=\bar{\theta}(s)} = 0 = -\nabla_\theta \psi(\bar{\theta}(s)) + \nabla_\theta \phi^T(\bar{\theta}(s)) s \,. \tag{39}$$

Plugging this relation into (38) and using the definition (22), we obtain

$$\nabla_s \mathrm{w}(s) = -\nabla_s \bar{\theta}^T(s) \, \nabla_\theta \phi^T(\bar{\theta}(s)) \mathrm{h}(s) \,. \tag{40}$$

By differentiating the function $s \mapsto \Phi(s;\bar{\theta}(s))$ where $\Phi(s,\theta) \stackrel{\text{def}}{=} \nabla_\theta \ell(s;\theta)$, we obtain

$$\nabla_s \Phi^T(s;\bar{\theta}(s)) = \nabla_s \Phi^T(s;\bar{\theta}(s)) + \nabla_s \bar{\theta}^T(s) \, \nabla_\theta \Phi^T(s,\bar{\theta}(s)) \,.$$

Since $\nabla_s \Phi^T(s;\theta) = \nabla_s \left[\nabla_\theta \ell(s;\theta) \right]^T = \left[\nabla_\theta \phi^T(\theta) \right]^T$ and $\nabla_\theta \Phi^T(s,\theta) = \nabla_\theta^2 \ell(s,\theta)$, the latter relation may be alternatively written as

$$\nabla_s \left[\nabla_\theta \ell(s;\theta)|_{\theta=\bar{\theta}(s)} \right]^T = \left[\nabla_\theta \phi^T(\bar{\theta}(s)) \right]^T + \nabla_s \bar{\theta}^T(s) \;\; \nabla_\theta^2 \ell(s;\theta)\big|_{\theta=\bar{\theta}(s)} \,.$$

Because the function $s \mapsto \nabla_\theta \ell(s;\theta)|_{\theta=\bar{\theta}(s)}$ is identically equal to 0, $\nabla_s [\nabla_\theta \ell(s;\theta)|_{\theta=\bar{\theta}(s)}]^T = 0$, and thus,

$$\nabla_\theta \phi^T(\bar{\theta}(s)) = -\, \nabla_\theta^2 \ell(s;\theta)\big|_{\theta=\bar{\theta}(s)} \left[\nabla_s \bar{\theta}^T(s) \right]^T \,. \tag{41}$$

Plugging this relation into (40) yields

$$\langle \nabla_s \mathrm{w}(s), \mathrm{h}(s) \rangle = \mathrm{h}^T(s) \left[\nabla_\theta \phi^T(\bar{\theta}(s)) \right]^T \left\{ \nabla_\theta^2 \ell(s;\theta)\big|_{\theta=\bar{\theta}(s)} \right\}^{-1} \nabla_\theta \phi^T(\bar{\theta}(s)) \mathrm{h}(s) \ , \tag{42}$$

where assumption 1(c) implies that, for any $s \in \mathcal{S}$, the matrix $\nabla_\theta^2 \ell(s;\theta)\big|_{\theta=\bar{\theta}(s)}$ is negative definite (and hence invertible). Hence, for any $s \in \mathcal{S}$, $\langle \nabla_s \mathrm{w}(s), \mathrm{h}(s) \rangle \leq 0$ with equality if and only if $\nabla_\theta \phi^T(\bar{\theta}(s)) \mathrm{h}(s) = 0$.

To deal with the equality case, assume that $s^\star$ is such that $\nabla_\theta \phi^T(\bar{\theta}(s^\star)) \mathrm{h}(s^\star) = 0$, or equivalently that

$$\nabla_\theta \phi^T(\bar{\theta}(s^\star)) \mathbb{E}_\pi \left[\bar{s}(Y; \bar{\theta}(s^\star)) \right] = \nabla_\theta \phi^T(\bar{\theta}(s^\star)) s^\star \ .$$

Under Assumptions 1(a) and 1(c), $\theta^\star = \bar{\theta}(s^\star)$ is the unique solution to the score equation (39), that is, such that $\nabla_\theta \psi(\theta^\star) = \nabla_\theta \phi^T(\theta^\star) s^\star$. Hence, $\nabla_\theta \phi^T(\theta^\star) \mathbb{E}_\pi [\bar{s}(Y;\theta^\star)] = \nabla_\theta \psi(\theta^\star)$, which implies that $\bar{\theta}(\mathbb{E}_\pi [\bar{s}(Y;\theta^\star)]) = \theta^\star$ showing that $\theta^\star \in \mathcal{L}$ and thus, from Proposition 3, that $s^\star \in \mathcal{S}$. The proof then follows upon noting that $s \mapsto \langle \nabla_s \mathrm{w}(s), \mathrm{h}(s) \rangle$ is continuous. □

Proof of Theorem 5. Under the stated assumptions, for any $\epsilon > 0$, there exists a compact $\mathcal{K} \subset \mathcal{S}$ and n, such that, $\mathbb{P}\left(\bigcap_{k \geq n} \{\hat{s}_n \in \mathcal{K}\} \right) \geq 1 - \epsilon$. Therefore, for any $\eta > 0$,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{k \geq n} \left| \sum_{i=n}^k \gamma_i \xi_i \right| \geq \eta \right) \\
&\quad \leq \mathbb{P}\left(\sup_{k \geq n} \left| \sum_{i=n}^k \gamma_i \xi_i 1_{\mathcal{K}}(\hat{s}_i) \right| \geq \eta, \bigcap_{i \geq n} \{\hat{s}_i \in \mathcal{K}\} \right) + \mathbb{P}\left(\bigcup_{i \geq n} \{\hat{s}_i \notin \mathcal{K}\} \right) \ , \\
&\quad \leq \epsilon + \mathbb{P}\left(\sup_{k \geq n} \left| \sum_{i=n}^k \gamma_i \xi_i 1_{\mathcal{K}}(\hat{s}_i) \right| \geq \eta \right) \ .
\end{aligned}$$

Note that $M_{n,k} = \sum_{i=n}^k \gamma_i \xi_i 1_{\mathcal{K}}(\hat{s}_i)$ is a L_2-martingale, and that its angle-bracket is bounded by $\sup_{s \in \mathcal{K}} \mathbb{E}_\pi [|\bar{s}(Y;\bar{\theta}(s))|^2] \sum_{i=n}^k \gamma_i^2 < \infty$. Using Chebyshev's inequality associated to the Doob martingale inequality, we conclude that

$$\mathbb{P}\left(\sup_{k \geq n} |M_{n,k}| \geq \eta \right) \leq 2\eta^{-2} \sup_{s \in \mathcal{K}} \mathbb{E}_\pi [|\bar{s}(Y;\bar{\theta}(s))|^2] \sum_{i=n}^\infty \gamma_i^2 \ ,$$

finally showing that $\limsup_n \sup_{k \geq n} |M_{n,k}| = 0$ with probability one. The proof is concluded by applying Theorem 2.3 of Andrieu et al. (2005) which states that the sequence of sufficient statistics defined by (21) is then such that $\limsup_n d(\hat{s}_n, \Gamma) = 0$ with probability one; the result on the sequence of parameter estimates $\hat{\theta}_n$ follows by continuity of $\bar{\theta}$. □

To prove Proposition 6, we will make use of the following stability lemma.

Lemma 8. *Let $p \geq 1$. Assume that for any compact subset $\mathcal{K} \subset \mathcal{S}$, $\sup_{s \in \mathcal{K}} \mathbb{E}_\pi \left[|\bar{s}(Y;\bar{\theta}(s))|^p \right] < \infty$ for some $p > 0$ and that $\mathbb{P}_\pi$-a.s. $\lim \hat{s}_n$ exists and belongs to $\mathcal{S}$. Then, the sequence $\{\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n))\}_{n \geq 0}$ is bounded in probability, i.e.*

$$\lim_{M \to \infty} \limsup_{n \to \infty} \mathbb{P}\left(|\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)|)| \geq M \right) = 0 \ .$$

Proof. Let $\mathcal{K}$ be a compact subset of $\mathcal{S}$. We may decompose $\mathbb{P}\left(|\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)|)| \geq M \right)$ as follows

$$\begin{aligned}
\mathbb{P}\left(|\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)|)| \geq M \right) &\leq \mathbb{P}(\hat{s}_n \notin \mathcal{K}) + \mathbb{P}\left(|\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)|)| \geq M, \hat{s}_n \in \mathcal{K} \right) \ , \\
&\leq \mathbb{P}(\hat{s}_n \notin \mathcal{K}) + M^{-p} \sup_{s \in \mathcal{K}} \mathbb{E}_\pi \left[|S(Y,\bar{\theta}(s))|^p \right] \mathbb{P}(\hat{s}_n \in \mathcal{K}) \ ,
\end{aligned}$$

which implies that

$$\limsup_{n\to\infty} \mathbb{P}\left(|\bar{s}(Y_{n+1};\bar{\theta}\left(\hat{s}_n\right)|)| \geq M\right) \leq \mathbb{P}(\lim_{n\to\infty} \hat{s}_n \notin \mathcal{K}) + M^{-p} \sup_{s\in\mathcal{K}} \mathbb{E}_\pi \left[|S(Y,\bar{\theta}\left(s\right))|^p\right] \ .$$

□

Proof of Proposition 6. A Taylor expansion with integral remainder shows that

$$\begin{aligned}\hat{\theta}_{n+1} &= \bar{\theta}\left(\hat{s}_n + \gamma_{n+1}\left(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right)\right) \\ &= \hat{\theta}_n + \gamma_{n+1}\left[\nabla_s\bar{\theta}^T\left(\hat{s}_n\right)\right]^T\left(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right) + \gamma_{n+1} r_{n+1} \ , \end{aligned} \tag{43}$$

where the remainder r_{n+1} is given by

$$\begin{aligned} r_{n+1}^T \stackrel{\text{def}}{=} &\left(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right)^T \\ &\times \int_0^1 \left[\nabla_s\bar{\theta}^T\left(\hat{s}_n + t\gamma_{n+1}\left\{\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right\}\right) - \nabla_s\bar{\theta}^T\left(\hat{s}_n\right)\right] \mathrm{d}t \ . \end{aligned} \tag{44}$$

We will first show that $\lim_{n\to\infty} r_n = 0$ a.s. Lemma 8 shows that $\{\bar{s}(Y_{n+1};\hat{\theta}_n)\}_{n\geq 0}$ is bounded in probability, which we denote by $\bar{s}(Y_{n+1};\hat{\theta}_n) = O_{\mathbb{P}}(1)$. Under the assumption of Theorem 5, $\hat{s}_n = O_{\mathbb{P}}(1)$, which implies that $\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n = O_{\mathbb{P}}(1)$. Choose $\epsilon > 0$ and then a compact subset $\mathcal{K}$ and a constant M large enough such that

$$\limsup_{n\to\infty} \mathbb{P}\left(\hat{s}_n \notin \mathcal{K}\right) + \limsup_{n\to\infty} \mathbb{P}\left(\left|\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right| \geq M\right) \leq \epsilon \ . \tag{45}$$

Since the function $\nabla_s\bar{\theta}\left(\cdot\right)$ is assumed continuous, it is uniformly continuous over every compact subset, i.e., there exists a constant δ_0 such that $\mathcal{K}_{\delta_0} \stackrel{\text{def}}{=} \{s \in \mathcal{S}, d(s,\mathcal{K}) \leq \delta_0\} \subset \mathcal{S}$ and

$$\sup_{|h|\leq\delta_0, s\in\mathcal{K}} |\nabla_s\bar{\theta}\left(s+h\right) - \nabla_s\bar{\theta}\left(s\right)| \leq \epsilon \ . \tag{46}$$

Since $\lim_{n\to\infty}\gamma_n = 0$ and $(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n)$ is bounded in probability, $\gamma_{n+1}(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n)$ converges to zero in probability, which we denote by $\gamma_{n+1}(\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n) = o_{\mathbb{P}}(1)$. For $\delta_0 > 0$ satisfying (46), this implies in particular that

$$\lim_{n\to\infty} \mathbb{P}\left(\gamma_{n+1}\left|\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right| \geq \delta_0\right) = 0 \ .$$

Therefore,

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}\left(|r_{n+1}| \geq M\epsilon\right) \leq &\limsup_{n\to\infty} \mathbb{P}\left(\gamma_{n+1}\left|\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right| \geq \delta_0\right) \\ &+ \limsup_{n\to\infty} \mathbb{P}\left(\hat{s}_n \notin \mathcal{K}\right) + \limsup_{n\to\infty} \mathbb{P}\left(\left|\bar{s}(Y_{n+1};\hat{\theta}_n) - \hat{s}_n\right| \geq M\right) \leq \epsilon \ , \end{aligned}$$

showing that $r_n = o_{\mathbb{P}}(1)$.

We now proceed with the first order term in (43). From (41) we have,

$$\nabla_s\bar{\theta}^T\left(s\right) = \left[\nabla_\theta\phi^T(\bar{\theta}\left(s\right))\right]^T \left[-\nabla_\theta^2\ell(s;\theta)\big|_{\theta=\bar{\theta}(s)}\right]^{-1} \ . \tag{47}$$

In addition, (36) and (39) show that

$$\nabla_\theta\phi^T\left(\bar{\theta}\left(s\right)\right)\left\{\bar{s}(y;\bar{\theta}\left(s\right)) - s\right\} = \nabla_\theta \log g(y;\bar{\theta}\left(s\right)) \ . \tag{48}$$

Therefore, by combining (47) and (48), the first order term in (43) may be rewritten as

$$\left[\nabla_s \bar{\theta}^T(\hat{s}_n)\right]^T \left(\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n)) - \hat{s}_n\right) = I_\pi^{-1}(\hat{\theta}_n)\nabla_\theta \log g(Y_{n+1};\hat{\theta}_n) + \left\{\left[-\nabla_\theta^2 \ell(\hat{s}_n;\theta)\big|_{\theta=\hat{\theta}_n}\right]^{-1} - I_\pi^{-1}(\hat{\theta}_n)\right\}\nabla_\theta \log g(Y_{n+1};\hat{\theta}_n) \ . \quad (49)$$

By definition, the complete data FIM may be rewritten as $I_\pi(\theta) = \nabla_\theta^2 \ell(s;\theta)\big|_{s=\mathbb{E}_\pi[\bar{s}(Y;\theta)]}$. Note also that $\hat{\theta}_n$ converges a.s. to $\theta^\star$ and that $I_\pi(\theta^\star)$ is assumed to be positive definite. Hence, to ensure that the term between braces in (49) may be neglected, we need to show that:

$$\nabla_\theta^2 \ell(\hat{s}_n;\theta)\big|_{\theta=\hat{\theta}_n} - \nabla_\theta^2 \ell(s;\theta)\big|_{(s,\theta)=(\mathbb{E}[\bar{s}(Y_{n+1};\hat{\theta}_n)|\mathcal{F}_n],\hat{\theta}_n)} = o_{\mathbb{P}}(1) \ . \quad (50)$$

Since the function $(s,\theta) \mapsto \nabla_\theta^2 \ell(s;\theta)$ is continuous, there exists $\delta_1 > 0$ such that, $\mathcal{K}_{\delta_1} = \{s \in \mathcal{S}, d(s,\mathcal{K}) \leq \delta_1\} \subset \mathcal{S}$ and

$$\sup_{|h|\leq\delta_1, s\in\mathcal{K}} \left|\nabla_\theta^2 \ell(s+h;\bar{\theta}(s)) - \nabla_\theta^2 \ell(s;\bar{\theta}(s))\right| \leq \epsilon \ ,$$

where the set $\mathcal{K}$ is defined in (45). Under the stated assumption $\lim_{n\to\infty} d(\hat{s}_n,\mathcal{L}) = 0$ a.s., which implies that

$$\lim_{n\to\infty} \mathrm{h}(\hat{s}_n) = \lim_{n\to\infty} \{\mathbb{E}[\bar{s}(Y_{n+1};\bar{\theta}(\hat{s}_n))|\mathcal{F}_n] - \hat{s}_n\} = 0 \ , \quad \text{a.s.}$$

By combining the latter two equations, we therefore obtain

$$\limsup_{n\to\infty} \mathbb{P}\left(\left|\nabla_\theta^2 \ell(\hat{s}_n;\theta)\big|_{\theta=\hat{\theta}_n} - \nabla_\theta^2 \ell(s;\theta)\big|_{(s,\theta)=(\mathbb{E}[\bar{s}(Y_{n+1};\hat{\theta}_n)|\mathcal{F}_n],\hat{\theta}_n)}\right| \geq \epsilon\right) \leq \limsup_{n\to\infty} \mathbb{P}(\hat{s}_n \notin \mathcal{K}) + \limsup_{n\to\infty} \mathbb{P}(|\mathrm{h}(\hat{s}_n)| \geq \delta_0) \leq \epsilon \ ,$$

showing (50). □

Proof of Theorem 7. We use the definition of the recursive EM sequence given by Proposition 6. The mean field associated to this sequence is given by $\mathrm{h}_\theta(\theta) \stackrel{\text{def}}{=} -I_\pi^{-1}(\theta)\nabla_\theta \mathrm{K}(\pi \| g_\theta)$. The Jacobian of this vector field at $\theta^\star$ is equal to $H(\theta^\star)$. Since $\theta^\star$ is a (possibly local) minimum of the Kullback-Leibler divergence, the matrix $\nabla_\theta^2 \mathrm{K}(\pi \| g_\theta)\big|_{\theta=\theta^\star}$ is positive definite. Because the two matrices $I_\pi^{-1}(\theta^\star)\, \nabla_\theta^2 \mathrm{K}(\pi \| g_\theta)\big|_{\theta=\theta^\star}$ and $I_\pi^{-1/2}(\theta^\star)\, \nabla_\theta^2 \mathrm{K}(\pi \| g_\theta)\big|_{\theta=\theta^\star}\, I_\pi^{-1/2}(\theta^\star)$ have the same eigenvalues, counting multiplicities, the eigenvalues of the matrix $H(\theta^\star) = I_\pi^{-1}(\theta^\star)\, \nabla_\theta^2 \mathrm{K}(\pi \| g_\theta)\big|_{\theta=\theta^\star}$ are all real and strictly positive. This shows the first assertion of the theorem; the proof of the second assertion follows directly from Pelletier, 1998, Theorem 1. □

Acknowledgement

The authors thank Maurice Charbit and Christophe Andrieu for precious feedback during the early stages of this work.