On the Convergence Properties of the Mini-Batch EM and MCEM Algorithms

Belhal Karimi, Marc Lavielle, Eric Moulines
CMAP, Ecole Polytechnique, Universite Paris-Saclay, 91128 Palaiseau, France
belhal.karimi@polytechnique.edu

July 30, 2018

Abstract

The EM algorithm is one of the most popular algorithm for inference in latent data models. For large datasets, each iteration of the algorithm can be numerically involved. To alleviate this problem, (Neal and Hinton, 1998) has proposed an incremental version in which the conditional expectation of the latent data (E-step) is computed on mini-batch of observations. In this paper, we propose and analyse the Monte Carlo version of the incremental EM in which the conditional expectation is evaluated by a Markov Chain Monte Carlo (MCMC). We establish the almost-sure convergence of these algorithms, covering both the mini-batch EM and its stochastic version. Various numerical applications are introduced in this article to illustrate our findings.

1 Introduction

Many problems in computational statistics reduce to maximising a function, defined on a feasible set Θ, of the following form:

$$g(\theta) \triangleq \int_{\mathsf{Z}} f(z,\theta)\mu(\mathrm{d}z) \ , \tag{1}$$

where $f : \mathsf{Z} \times \Theta \to \mathbb{R}^+$ is a positive function and μ is a σ-finite measure. In the incomplete data framework, the function g is the incomplete data likelihood, z is the missing data vector and f stands for the complete data likelihood, that is the joint likelihood of the observations and the missing data.

When the direct optimisation of the function g is difficult, the EM algorithm may be an option. The EM algorithm iteratively computes a sequence of estimates $\{\theta^k, k \in \mathbb{N}\}$ starting from some initial parameter θ^0. Each iteration of the EM algorithm may be decomposed into two steps. In the E-step, a surrogate function

$$\theta \mapsto Q(\theta,\theta^{k-1}) \triangleq \int_{\mathsf{Z}} \log f(z,\theta) p(z,\theta^{k-1})\mu(\mathrm{d}z)$$

is computed where $p(z, \theta^{k-1}) \triangleq f(z, \theta^{k-1})/g(\theta^{k-1})$ is the probability density of the latent variables at the current fit of the parameter θ^{k-1}. In the M-step, this surrogate function is maximised yielding to a new fit of the parameter $\theta^k = \operatorname{argmax}_\theta Q(\theta, \theta^{k-1})$. The EM algorithm has been the object of considerable interest since its introduction in (Dempster et al., 1977); the scope of the algorithm and many applications are presented in the reference book (McLachlan and Krishnan, 2008). The EM algorithm has a number of interesting features: it is monotone - at each iteration, the algorithm improves the objective function or leaves it unchanged if a local maximum has been achieved - , it is invariant in the choice of the parametrisation, it is numerically very stable - when the optimisation set is well defined - and easy to implement on a large class of models.

Many possible improvements have been proposed. In a landmark paper, (Neal and Hinton, 1998) has proposed an incremental version of the algorithm. In many applications, $\log f(z, \theta)$ can be written as a large sum terms: it is therefore possible to update at each iteration only a subsample of the terms in this sum and then to perform the M-step. As this algorithm makes use of the new information immediately, it is expected that it might improves the convergence of the EM algorithm in this context. This algorithm has had an enormous impact in applied statistics and machine learning; see among many others (Thiesson et al., 2001) for maximum likelihood estimation with missing data in large datasets, (Hsiao et al., 2006) for PET tomographic reconstruction, (Vlassis and Likas, 2002; Ng and McLachlan, 2003) for Gaussian mixture learning, (Likas and Galatsanos, 2004) for blind image deconvolution, (Ng and McLachlan, 2004) for segmentation of magnetic resonance images, (Blei et al., 2017) for variational inference and (Ablin et al., 2018) for Independent Component Analysis. A closely related version of the EM has been introduced in (Cappé and Moulines, 2009; Cappé, 2011); the objective is therefore slightly different, since in this case the observations are processed online.

In a recent paper, closing the gap between the practical use of EM and its theoretical understanding, (Balakrishnan et al., 2017) developed a "sample-splitting EM" algorithm where the parameters are obtained, at each iteration, using a subset of the observations. The authors give quantitative characterisation of the region of attraction around the global optimum for both the idealized limit of infinite samples and finite sample sets. Even though this comprehensive work gives strong theoretical guarantees of the EM algorithm and its subsample-based variant, it does not deal with the incremental version of the EM algorithm we are studying here since the update of this "sample-splitting" EM algorithm does not include any terms of previous iterations.

The convergence of the incremental version of the EM algorithm was first tackled by (Gunawardana and Byrne, 2005) exploiting the interpretation of the EM algorithm as an alternating minimisation procedure under the information geometric framework developed in (Csiszár and Tusnády, 1984). Nevertheless, (Gunawardana and Byrne, 2005) assume that the latent variables take only a finite number of values and the order in which the observations are processed remains the same from one pass to the other. There is no obvious way to extend this analysis to more general latent variables and sampling schemes.

In the high-dimensional setting, exact maximisation based M-step reveals to be very time consuming, or even ill-posed, thus gradient EM (Wang et al., 2014) can be very attractive. In (Zhu et al., 2017), the authors propose a novel high-dimensional EM algorithm by incorporating variance reduction into the stochastic gradient method for EM. In this algorithm, the full gradient is calculated incrementally, updating only a mini-batch of components at each iteration. This method is widely inspired by recent advances in stochastic optimisation (Roux et al., 2012; Defazio et al., 2014) and can be appealing when the dimension is much larger than the sample size. Their algorithm has an improved overall computational complexity over (Wang et al., 2014) gradient EM and converges at a linear rate to a local optimum (see the results and proofs therein).

The Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990) has been proposed, when the quantity computed at the E-step involves infeasible computations. It lies in splitting the E-step in a first step where the latent variables are simulated and then computing a Monte Carlo integration of the intractable expectation of the complete log likelihood. The M-step remains unchanged. The MCEM algorithm has been successfully applied in mixed effects models (McCulloch, 1997; Hughes, 1999; Baey et al., 2016) or to do inference for joint modelling of time to event data coming from clinical trials in (Chakraborty and Das, 2010).

This algorithm has been initially studied in (Chan and Ledolter, 1995) followed by many authors such as (Sherman et al., 1999) showing the convergence of the MCEM when the Monte Carlo integration is done using independent Markov chains generated by a Gibbs sampler, (Levine and Casella, 2001; Booth et al., 2001) giving details on the implementation of the MCEM, (Fort et al., 2003) generalizing the results of (Sherman et al., 1999) for a wide class of MCMC simulation techniques, and then in (McLachlan and Krishnan, 2008) and (Neath et al., 2013).

In this contribution, we propose the stochastic version of the mini-batch EM algorithm, called the mini-batch MCEM (MBMCEM). The mini-batch of surrogate functions, computed at each iteration, are no longer determnistic but are rather approximated using Monte Carlo integration. The incremental framework developed in (Mairal, 2015), called MISO (Minimisation by Incremental Surrogate Optimisation), used to analyse the MBEM, can not be applied in this context and we provide, in this article, its extension so that convergence guarantees of the objective function and similar stationary point condition are established.

We summarise the major contributions of this paper as follows. Using the MISO framework (Mairal, 2015) we establish, under mild assumptions on the incomplete model and the auxiliary Q-function, the almost-sure convergence of the mini-batch EM algorithm by constructing suitable surrogate functions. We then establish the almost-sure convergence of the mini-batch version of the MCEM using an extension of the MISO framework when the surrogate functions are stochastic.

The remainder of this paper is organised as follows. Section 2 provides the assumptions on the model, the MBEM algorithm and sets out the convergence of the objective function. Section 3 introduces the MBMCEM algorithm and its convergence theorem. Each section also provides the executed algorithm when the complete model belongs to the curved exponential family. We investigate Section 4, through a simulation study on a mixed effect model and a logistic regression, how these algorithms converge with respect to the mini-batch size. Section 5 is devoted to the technical proofs of our results.

2 Convergence of the mini-batch EM algorithm

2.1 Model assumptions and notations

M 1. *The parameter set Θ is a closed convex subset of $\mathbb{R}^p$.*

Let N be an integer and for $i \in [\![1, N]\!]$, Z_i be a subset of $\mathbb{R}^{m_i}$, μ_i be a σ-finite measure on the Borel σ-algebra $\mathcal{Z}_i = \mathcal{B}(\mathsf{Z}_i)$ and $\{f_i(z_i, \theta), \theta \in \Theta\}$ be a family of positive μ_i-integrable Borel functions on Z_i. Set $z = (z_i \in \mathsf{Z}_i, i \in [\![1, N]\!]) \in \mathsf{Z}$ where $\mathsf{Z} = \times_{n=1}^{N} \mathsf{Z}_i$ and μ is the product of the measures $(\mu_i, i \in [\![1, N]\!])$.
Define, for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$:

$$g_i(\theta) \triangleq \int_{\mathsf{Z}_i} f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \quad \text{and} \quad p_i(z_i, \theta) \triangleq \begin{cases} \frac{f_i(z_i,\theta)}{g_i(\theta)} & \text{if } g_i(\theta) \neq 0 \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

Note that $p_i(z_i, \theta)$ defines a probability density function with respect to μ_i. Thus $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ is a family of probability density. We denote by $\{\mathbb{P}_{i,\theta}; \theta \in \Theta\}$ the associated family of probability measures. For all $\theta \in \Theta$, we set

$$f(z, \theta) = \prod_{i=1}^{N} f_i(z_i, \theta), \quad g(\theta) = \prod_{i=1}^{N} g_i(\theta) \quad \text{and} \quad p(z, \theta) = \prod_{i=1}^{N} p_i(z_i, \theta) . \tag{3}$$

Remark 1. *An example of this setting is the incomplete data framework. In this case, we consider N independent observations $(y_i \in \mathsf{Y}_i, i \in [\![1, N]\!])$ where Y_i is a subset of $\mathbb{R}^{\ell_i}$ and missing data $(z_i \in \mathsf{Z}_i, i \in [\![1, N]\!])$. In this framework, we get*

- *$f_i(z_i, \theta)$ is the complete data likelihood that is the likelihood of the observed data y_i augmented with the missing data z_i.*
- *$g_i(\theta)$ is the incomplete data likelihood that is the likelihood of the observed data y_i.*
- *$p_i(z_i, \theta)$ is the posterior distribution of the missing data z_i given the observed data y_i.*

Our objective is to maximise the function $\theta \to \log g(\theta)$ or equivalently to minimise the objective function $\ell : \Theta \mapsto \mathbb{R}$ defined as:

$$\ell(\theta) \triangleq -\log g(\theta) = \sum_{i=1}^{N} \ell_i(\theta) \quad \text{for all } \theta \in \Theta , \tag{4}$$

where $\ell_i(\theta) \triangleq -\log g_i(\theta)$. The EM algorithm is an iterative optimisation algorithm that minimises the function $\theta \to \ell(\theta)$ when its direct minimisation is difficult. Denote by θ^{k-1} the current fit of the parameter at iteration k. The k-th step of the EM algorithm might be decomposed into two steps. The E-step consists in computing the surrogate function defined for all $\theta \in \Theta$ as :

$$Q(\theta, \theta^{k-1}) \triangleq -\int_{\mathsf{Z}} p(z, \theta^{k-1}) \log f(z, \theta) \mu(\mathrm{d}z) \tag{5}$$

$$= -\sum_{i=1}^{N} \int_{\mathsf{Z}_i} p_i(z_i, \theta^{k-1}) \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) = \sum_{i=1}^{N} Q_i(\theta, \theta^{k-1}) \;, \tag{6}$$

where:

$$Q_i(\theta, \theta^{k-1}) \triangleq -\int_{\mathsf{Z}_i} p_i(z_i, \theta^{k-1}) \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \;. \tag{7}$$

In the M-step, the value of θ minimising $Q(\theta, \theta^{k-1})$ is calculated. This yields the new parameter estimate θ^k. These two steps are repeated until convergence. The essence of the EM algorithm is that decreasing $Q(\theta, \theta^{k-1})$ forces a decrease of the function $\theta \to \ell(\theta)$ (Dempster et al., 1977). The mini-batch version of the EM algorithm is described as follows:

Algorithm 1 mini-batch EM algorithm

Initialisation: given an initial parameter estimate θ^0, for all $i \in [\![1, N]\!]$ compute a surrogate function $\vartheta \to R_i^0(\vartheta) = Q_i(\vartheta, \theta^0)$ defined by (7).
Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$.
2. For all $i \in I_k$, compute $\vartheta \to Q_i(\vartheta, \theta^{k-1})$ defined by (7).
3. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{N} R_i^k(\vartheta)$ where $R_i^k(\vartheta)$ are defined recursively as follows:

$$R_i^k(\vartheta) = \begin{cases} Q_i(\vartheta, \theta^{k-1}) & \text{if } i \in I_k \\ R_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{8}$$

We remark that, for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$:

$$R_i^k(\theta) = Q_i(\theta, \theta^{\tau_{i,k}}) \;, \tag{9}$$

where for all $i \in [\![1, N]\!]$, $\tau_{i,0} = 0$ and $k \geq 1$ the indices $\tau_{i,k}$ are defined recursively as follows:

$$\tau_{i,k} = \begin{cases} k-1 & \text{if } i \in I_k \\ \tau_{i,k-1} & \text{otherwise} \end{cases} \tag{10}$$

As noted in (Gunawardana and Byrne, 2005) and (Neal and Hinton, 1998), there is no guarantee, unlike the EM algorithm, that the objective function $\theta \to \ell(\theta)$ decreases at each iteration. We also remark that we recover the full EM algorithm when the mini-batch size p is set to be equal to N. Let $\mathcal{T}(\Theta)$ be a neighborhood of Θ. To study the convergence of the MBEM algorithm we consider the following assumptions:

M 2. *For all $i \in [\![1, N]\!]$, assume that:*

a. *For all $\theta \in \Theta$ and $z_i \in \mathsf{Z}_i$, $f_i(z_i, \theta)$ is strictly positive, the function $\theta \to f_i(z_i, \theta)$ is two-times differentiable on $\mathcal{T}(\Theta)$ for μ_i almost every z_i and for all $\vartheta \in \Theta$:*

$$\int_{\mathsf{Z}_i} |\nabla f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty \quad \textit{and} \quad \int_{\mathsf{Z}_i} p_i(z_i, \vartheta) |\log f_i(z_i, \theta)| \mu_i(\mathrm{d}z_i) < \infty \,. \tag{11}$$

b. *For all $\theta \in \Theta$, there exist $\delta > 0$ and a measurable function $\psi_\theta : \mathsf{Z}_i \to \mathbb{R}$ such that*

$$\sup_{\|\vartheta - \theta\| \leq \delta} |\nabla^2 f_i(z_i, \vartheta)| \leq \psi_\theta(z_i)$$

for μ_i almost every z_i with $\int_{\mathsf{Z}_i} \psi_\theta(z_i) \mu_i(\mathrm{d}z_i) < \infty$.

c. *There exist a measurable function $\phi_i : \mathsf{Z}_i \to \mathbb{R}$ and $L_i < \infty$ such that*

$$\sup_{\theta \in \Theta} |\nabla^2 \log f_i(z_i, \theta)| \leq \phi_i(z_i)$$

for μ_i almost every z_i with $\sup_{\theta \in \Theta} \int_{\mathsf{Z}_i} p_i(z_i, \theta) \phi_i(z_i) \mu_i(\mathrm{d}z_i) \leq L_i$.

d. *For all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, $\sup_{\theta \in \Theta} |\nabla^2 l_i(\theta)| < \infty$.*

It is easily checked that these assumptions imply for all $i \in [\![1, N]\!]$ that:

1. The function $\theta \to g_i(\theta)$ is continuously differentiable on $\mathcal{T}(\Theta)$ and the Fisher identity (Fisher, 1925) holds:

$$\nabla \ell_i(\theta) = -\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \,. \tag{12}$$

2. For all $\vartheta \in \Theta$, the function $\theta \to Q_i(\theta, \vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$ and is L_i-smooth, i.e., for all $(\theta, \theta') \in \Theta$ and $L_i > 0$:

$$\|\nabla Q_i(\theta, \vartheta) - \nabla Q_i(\theta', \vartheta)\| \leq L_i \|\theta - \theta'\| \,. \tag{13}$$

3. For all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, Louis Formula (Louis, 1982) yields that:

$$\begin{aligned}
\nabla^2 l_i(\theta) = & -\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla^2 \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) & (14)\\
& -\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) & (15)\\
& + \left(\int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i)\right)^\top \int_{\mathsf{Z}_i} p_i(z_i, \theta) \nabla \log f_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \,. & (16)
\end{aligned}$$

Thus, sufficient conditions to verify M 2d. are M 2c. and the following condition: There exist a measurable function $N_i : \mathsf{Z}_i \to \mathbb{R}$ such that for all $\theta \in \Theta$, $|\nabla \log f_i(z_i, \theta)| \leq N_i(z_i)$ for μ_i almost every z_i with $\int_{\mathsf{Z}_i} p_i(z_i, \theta) N_i^2(z_i) \mu_i(\mathrm{d}z_i) < \infty$.

M 3. *For all $i \in [\![1, N]\!]$, the objective function ℓ_i is bounded from below, i.e. there exist $M_i \in \mathbb{R}$ such that for all $\theta \in \Theta$:*

$$\ell_i(\theta) \geq M_i \; . \tag{17}$$

For $\theta \in \Theta$, we say that a function $B_{i,\theta}$ is a surrogate of ℓ_i at θ if the following three properties are satisfied:

S.1 the function $\vartheta \to B_{i,\theta}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$

S.2 for all $\vartheta \in \Theta$, $B_{i,\theta}(\vartheta) \geq \ell_i(\vartheta)$

S.3 $B_{i,\theta}(\theta) = \ell_i(\theta)$ and $\nabla B_{i,\theta}(\vartheta)\Big|_{\vartheta=\theta} = \nabla \ell_i(\vartheta)\Big|_{\vartheta=\theta}$.

For all $i \in [\![1, N]\!]$ and $(\theta, \theta') \in \Theta^2$, define the Kullback-Leibler Divergence from $\mathbb{P}_{i,\theta'}$ to $\mathbb{P}_{i,\theta}$ as:

$$\mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\theta'}) \triangleq \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{p_i(z_i, \theta')} \mu_i(\mathrm{d}z_i) \tag{18}$$

and the negated entropy function $H_i(\theta)$ as:

$$H_i(\theta) \triangleq \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \; . \tag{19}$$

To analyse the MBEM algorithm, we introduce for $i \in [\![1, N]\!]$ and $\theta \in \Theta$ the function $\vartheta \to B_{i,\theta}(\vartheta)$ defined by:

$$B_{i,\theta}(\vartheta) \triangleq Q_i(\vartheta, \theta) + H_i(\theta) \; . \tag{20}$$

We will show below that for $i \in [\![1, N]\!]$ and $\theta \in \Theta$, $B_{i,\theta}$ is a surrogate of l_i at θ. Let us note that this function can be rewritten as follows:

$$\begin{aligned} B_{i,\theta}(\vartheta) &= \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{f_i(z_i, \vartheta)} \mu_i(\mathrm{d}z_i) & (21) \\ &= \int_{\mathsf{Z}_i} p_i(z_i, \theta) \log \frac{p_i(z_i, \theta)}{p_i(z_i, \vartheta)} \mu_i(\mathrm{d}z_i) + \ell_i(\vartheta) & (22) \\ &= \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta}) + \ell_i(\vartheta) \; . & (23) \end{aligned}$$

We verify **S.1** using assumption M 2. Since $\vartheta \to \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta})$ is always positive and is equal to zero if $\theta = \vartheta$, we verify **S.2** and the first part of **S.3**. The second part of **S.3** follows from the Fisher identity (12). The difference between the surrogate function and the objective function denoted, for all $\vartheta \in \Theta$, $h_i(\vartheta) \triangleq B_{i,\theta}(\vartheta) - l_i(\vartheta)$ plays a key role in the convergence analysis. Here, for all $i \in [\![1, N]\!]$ and $\vartheta \in \Theta$ the error reads $h_i(\vartheta) = \mathrm{KL}(\mathbb{P}_{i,\theta} \parallel \mathbb{P}_{i,\vartheta})$. Under M 2c. and M 2d., we obtain that for all $i \in [\![1, N]\!]$, the function $\vartheta \to h_i(\vartheta)$ is L_i-smooth. Since for all $i \in [\![1, N]\!]$ and $\theta \in \Theta$, the surrogate function $\vartheta \to B_{i,\theta}(\vartheta)$ is equal to $\vartheta \to Q_i(\vartheta, \theta)$ up to a constant, the MBEM algorithm is equivalent to the following theoretical algorithm:

Algorithm 2 Theoretical MBEM algorithm

Initialisation: given an initial parameter estimate θ^0, for all $i \in [\![1, N]\!]$ compute a surrogate function $\vartheta \to A_i^0(\vartheta) = B_{i,\theta^0}(\vartheta)$ defined by (21).
Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$.

2. For all $i \in I_k$, compute a surrogate function $\vartheta \to B_{i,\theta^{k-1}}(\vartheta)$ defined by (21).

3. Set $\theta^k \in \arg\min\limits_{\vartheta \in \Theta} \sum_{i=1}^N A_i^k(\vartheta)$ where $A_i^k(\vartheta)$ are defined recursively as follows:

$$A_i^k(\vartheta) = \begin{cases} B_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ A_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{24}$$

We remark that, for all $i \in [\![1, N]\!]$ and $\vartheta \in \Theta$:

$$A_i^k(\vartheta) = B_{i,\theta^{\tau_{i,k}}}(\vartheta) \tag{25}$$

using the notation introduced in (10). Denote by $\langle \cdot, \cdot \rangle$ the scalar product. We now state the convergence theorem of the MBEM algorithm:

Theorem 1. *Assume* **M1-M3**. *Let* $\left(\theta^k\right)_{k \geq 1}$ *be a sequence generated from* $\theta^0 \in \Theta$ *by the iterative application described by algorithm 1. Then:*

(i) $\left(\ell(\theta^k)\right)_{k \geq 1}$ *converges almost surely.*

(ii) $\left(\theta^k\right)_{k \geq 1}$ *satisfies the Asymptotic Stationary Point Condition, i.e.*

$$\liminf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq 0 \tag{26}$$

Proof. The proof is postponed to section 5.1 □

We observe that in the unconstrained case, we have:

$$\inf_{\theta \in \mathbb{R}^d} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} = -\|\nabla \ell(\theta^k)\| \ , \tag{27}$$

which yields to $\lim\limits_{k \to \infty} \|\nabla \ell(\theta^k)\| = 0$.

2.2 MBEM for a curved exponential family

In the particular case where for all $i \in \llbracket 1, N \rrbracket$ and $z_i \in \mathsf{Z}_i$, the function $\theta \to f_i(z_i, \theta)$ belongs to the curved exponential family, we assume that:

E 1. *For all $i \in \llbracket 1, N \rrbracket$ and $\theta \in \Theta$:*

$$\log f_i(z_i, \theta) = H_i(z_i) - \psi_i(\theta) + \langle \tilde{S}_i(z_i), \phi_i(\theta) \rangle. \tag{28}$$

where $\psi_i : \Theta \mapsto \mathbb{R}$ and $\phi_i : \Theta \mapsto \mathbb{R}$ are twice continuously differentiable functions of θ, $H_i : \mathsf{Z}_i \mapsto \mathbb{R}$ is a twice continuously differentiable function of z_i and $\tilde{S}_i : \mathsf{Z}_i \mapsto \mathsf{S}_i$ is a statistic taking its values in a convex subset S_i of $\mathbb{R}$ and such that $\int_{\mathsf{Z}_i} |\tilde{S}_i(z_i)| p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) < \infty$.

Define for all $\theta \in \Theta$ and $i \in \llbracket 1, N \rrbracket$ the function $\bar{s}_i : \Theta \to \mathsf{S}_i$ as:

$$\bar{s}_i(\theta) \triangleq \int_{\mathsf{Z}_i} \tilde{S}_i(z_i) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i). \tag{29}$$

Define, for all $\theta \in \Theta$ and $s = (s_i, i \in \llbracket 1, N \rrbracket) \in \mathsf{S}$ where $\mathsf{S} = \times_{n=1}^{N} \mathsf{S}_i$, the function $L(s; \theta)$ by:

$$L(s; \theta) \triangleq \sum_{i=1}^{N} \psi_i(\theta) - \sum_{i=1}^{N} \langle s_i, \phi_i(\theta) \rangle. \tag{30}$$

E 2. *There exist a function $\hat{\theta} : \mathsf{S} \mapsto \Theta$ such that for all $s \in \mathsf{S}$, :*

$$L(s; \hat{\theta}(s)) \leq L(s; \theta). \tag{31}$$

In many models of practical interest for all $s \in \mathsf{S}$, $\theta \mapsto L(s, \theta)$ has a unique minimum. In the context of the curved exponential family, the MBEM algorithm can be formulated as follows:

Algorithm 3 mini-batch EM for a curved exponential family

Initialisation: given an initial parameter estimate θ^0, for all $i \in \llbracket 1, N \rrbracket$ compute $s_i^0 = \bar{s}(\theta^0)$.

Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset \llbracket 1, N \rrbracket, \mathrm{card}(A) = p\}$.

2. For $i \in \llbracket 1, N \rrbracket$, compute s_i^k such as:

$$s_i^k = \begin{cases} \bar{s}_i(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases} \tag{32}$$

3. Set $\theta^k = \hat{\theta}(s^k)$ where $s^k = (s_i^k, i \in \llbracket 1, N \rrbracket)$.

Example 1. We observe N independent and identically distributed (i.i.d.) random variables $(y_i, i \in [\![1, N]\!])$. Each one of these observations is distributed according to a mixture model. Denote by $(c^j, j \in [\![1, J]\!])$ the distribution of the component of the mixture and $(\pi_j, j \in [\![1, J]\!])$ the associated weights. Consider the complete data likelihood for each individual $f_i(z_i, \theta)$:

$$f_i(z_i, \theta) = \prod_{j=1}^{J} \left(\pi_j c^j(y_i, \delta)\right)^{\mathbb{1}_{z_i=j}} . \tag{33}$$

We restrict this study to a mixture of Gaussian distributions. In such case $\theta = ((\pi_j, \mu_j, \sigma_j), j \in [\![1, J]\!])$ and the individual complete log likelihood is expressed as:

$$\log f_i(z_i, \theta) = \sum_{j=1}^{J} \mathbb{1}_{z_i=j} \log(\pi_j) + \sum_{j=1}^{J} \mathbb{1}_{z_i=j} \left[-\frac{(y_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2} \log \sigma_j^2 \right] . \tag{34}$$

The complete data sufficient statistics are given for all $i \in [\![1, N]\!]$ and $j \in [\![1, J]\!]$, by $\tilde{S}_i^{1,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i=j}$, $\tilde{S}_i^{2,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i=j} y_i$ and $\tilde{S}_i^{3,j}(y_i, z_i) \triangleq \mathbb{1}_{z_i=j} y_i^2$. At each iteration k, algorithm 3 consists in picking a set I_k and for $i \in I_k$, computing the following quantities:

$$(\bar{s}_i^k)^{1,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) , \tag{35}$$

$$(\bar{s}_i^k)^{2,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} y_i p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i , \tag{36}$$

$$(\bar{s}_i^k)^{3,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} y_i^2 p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i^2 , \tag{37}$$

where the quantity $p_{ij}(\theta^{k-1}) \triangleq \mathbb{P}_{i,\theta^{k-1}}(z_i = j)$ is obtained using the Bayes rule:

$$p_{ij}(\theta^{k-1}) = \frac{\mathbb{P}_i(z_i = j) p_i(y_i | z_i = j; \theta^{k-1})}{p_i(y_i; \theta^{k-1})} = \frac{\pi_j^{k-1} c^j(y_i; \mu_j^{k-1}, \sigma_j^{k-1})}{\sum_{l=1}^{J} \pi_l^{k-1} c^l(y_i; \mu_l^{k-1}, \sigma_l^{k-1})} . \tag{38}$$

For $i \notin I_k$, $j \in [\![1, J]\!]$, and $d \in [\![1, 3]\!]$ $(\bar{s}_i^k)^{d,j} = (\bar{s}_i^{k-1})^{d,j}$. Finally the maximisation step yields:

$$\pi_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}}{N} , \tag{39}$$

$$\mu_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{2,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} , \tag{40}$$

$$\sigma_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{3,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} - (\mu_j^k)^2 . \tag{41}$$

3 Convergence of the mini-batch MCEM algorithm

We now consider the stochastic version of the MBEM algorithm called the mini-batch MCEM algorithm. At iteration k, the MBMCEM approximates the quantity defined by (7) by Monte Carlo integration, i.e. for all $i \in I_k$, $\vartheta \in \Theta$ and $k \geq 1$:

$$\hat{Q}_i^k(\vartheta, \theta^{k-1}) \triangleq \frac{1}{M_k} \sum_{m=0}^{M_k-1} \log f_i(z_i^{k,m}, \vartheta) \ , \tag{42}$$

where $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ is a Monte Carlo batch. In simple scenarios, the samples $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ are conditionally independent and identically distributed with distribution $p_i(z_i, \theta^{k-1})$. Nevertheless, in most cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov Chains (MCMC) algorithm. MCMC algorithms are a class of methods allowing to sample from complex distribution over (possibly) large dimensional space.

Recall that a Markov kernel P on a measurable space $(\mathsf{E}, \mathcal{E})$ is an application on $\mathsf{E} \times \mathcal{E}$, taking values in $[0, 1]$ such that for any $z \in \mathsf{E}$, $P(z, \cdot)$ is a probability measure on $\mathcal{E}$ and for any $A \in \mathcal{E}$, $P(\cdot, A)$ is measurable. We denote by P^k the $k-$th iterate of P defined recursively as $P^0(z, A) \triangleq \mathbb{1}_A(z)$ and for $k \geq 1$, $P^k(z, A) \triangleq \int_A P^{k-1}(z, \mathrm{d}z')P(z', A)$. The probability π is said to be stationary for P if $\int_{\mathsf{E}} \pi(\mathrm{d}z)P(z, A) = \pi(A)$ for any $A \in \mathcal{E}$. We refer the reader to (Meyn and Tweedie, 2012) for the definitions of basic properties of Markov chains.

For $i \in [\![1, N]\!]$ and $\theta \in \Theta$, let $P_{i,\theta}$ be a Markov kernel with stationary distribution $\pi_{i,\theta}(A_i) = \int_{A_i} p_i(z_i, \theta)\mu_i(\mathrm{d}z_i)$ where $A_i \in \mathcal{Z}_i$. For example, $P_{i,\theta}$ might be either a Gibbs or a Metropolis-Hastings samplers with target distribution $\pi_{i,\theta}$. For $\theta \in \Theta$, let $\lambda_{i,\theta}$ be a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$. We will use $\lambda_{i,\theta}$ as an initial distribution and allow this initial distribution to depend on the parameter θ. For example, $\lambda_{i,\theta}$ might be the Dirac mass at some given point but more clever choice can be made. We denote by $\mathbb{E}_{i,\theta}$ the expectation of the canonical Markov chain $\{z_i^m\}_{m=0}^{\infty}$ with initial distribution $\lambda_{i,\theta}$ and transition kernel $P_{i,\theta}$.

In this setting, the Monte Carlo mini-batch $\{z_i^{k,m}\}_{m=0}^{M_{k-1}}$ is a realisation of a Markov Chain with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$. The MBMCEM algorithm can be summarised as follows:

```
Algorithm 4 mini-batch MCEM algorithm
Initialisation: given an initial parameter estimate $\theta^0$, for all $i \in [\![1, N]\!]$ and $m \in [\![0, M_0 - 1]\!]$, sample a Markov Chain $\{z_i^{0,m}\}_{m=0}^{M_0-1}$ with initial distribution $\lambda_{i,\theta^0}$ and transition kernel $P_{i,\theta^0}$ and compute a function $\vartheta \to \hat{R}_i^0(\vartheta) = \hat{Q}_i^0(\vartheta, \theta^0)$ defined by (42).
Iteration k: given the current estimate $\theta^{k-1}$:
    1. Pick a set $I_k$ uniformly on $\{A \subset [\![1, N]\!], \text{card}(A) = p\}$.
    2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sample a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$.
    3. For all $i \in I_k$, compute the function $\vartheta \to \hat{Q}_i^k(\vartheta, \theta^{k-1})$ defined by (42).
    4. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N \hat{R}_i^k(\vartheta)$ where $\hat{R}_i^k(\vartheta)$ are defined recursively as follows:
        $$\hat{R}_i^k(\vartheta) = \begin{cases} \hat{Q}_i^k(\vartheta, \theta^{k-1}) & \text{if } i \in I_k \\ \hat{R}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{43}$$
```

Whether we use Markov Chain Monte Carlo or direct simulation, we need to control the supremum norm of the fluctuations of the Monte Carlo approximation. Let $i \in [\![1, N]\!]$, $\{q_i(z_i, \vartheta), z_i \in \mathsf{Z}_i, \vartheta \in \Theta\}$ be a family of measurable functions, λ_i a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$. We define:

$$C_i(q_i) \triangleq \sup_{\theta \in \Theta} \sup_{M > 0} M^{-1/2} \mathbb{E}_{i,\theta} \left[\sup_{\vartheta \in \Theta} \left| \sum_{m=0}^{M-1} \left\{ q_i(z_i^m, \vartheta) - \int_{\mathsf{Z}_i} q_i(z_i, \vartheta) p_i(z_i, \theta) \lambda_i(\mathrm{d}z_i) \right\} \right| \right] . \tag{44}$$

M 4. *For all $i \in [\![1, N]\!]$:*

$$C_i(\log f_i) < \infty \quad \textit{and} \quad C_i(\nabla \log f_i) < \infty . \tag{45}$$

The assumption M 4 is based on maximal inequality for beta-mixing sequences obtained in (Doukhan et al., 1995). This condition can be translated in terms of drift and minorisation conditions (see (Meyn and Tweedie, 2012)). Finally, we consider the following assumption on the number of simulations:

M 5. *$\{M_k\}_{k \geq 0}$ is a non deacreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$.*

We now state the convergence theorem for the MBMCEM algorithm:

Theorem 2. *Assume* **M1-M5**. *Let $\left(\theta^k\right)_{k \geq 1}$ be a sequence generated from $\theta^0 \in \Theta$ by the iterative application described by algorithm 4. Then:*

(i) $\left(\ell(\theta^k)\right)_{k \geq 1}$ converges almost surely.

(ii) $\left(\theta^k\right)_{k\geq 1}$ *satisfies the Asymptotic Stationary Point Condition.*

Proof. The proof is postponed to section 5.2 □

3.1 MBMCEM for a curved exponential family

Using the notations introduced in section 2.2, we can write the mini-batch MCEM algorithm can be described as follows:

Algorithm 5 mini-batch MCEM for a curved exponential family

Initialisation: given an initial parameter estimate θ^0, for all $i \in [\![1, N]\!]$ and $m \in [\![0, M_0 - 1]\!]$, sample a Markov Chain $\{z_i^{0,m}\}_{m=0}^{M_0-1}$ with initial distribution λ_{i,θ^0} and transition kernel P_{i,θ^0} and compute $s_i^0 = \frac{1}{M_0}\sum_{m=1}^{M_0} \tilde{S}_i(z_i^{0,m})$.
Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![1, N]\!], \mathrm{card}(A) = p\}$.
2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sample a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ with initial distribution $\lambda_{i,\theta^{k-1}}$ and transition kernel $P_{i,\theta^{k-1}}$.
3. Compute s_i^k such as:

$$s_i^k = \begin{cases} \frac{1}{M_k}\sum_{m=1}^{M_k-1} \tilde{S}_i(z_i^{k,m}) & \text{if } i \in I_k \\ s_i^{k-1} & \text{otherwise} \end{cases} \tag{46}$$

4. Set $\theta^k = \hat{\theta}(s^k)$ where $s^k = (s_i^k, i \in [\![1, N]\!])$

Example 2. For illustration purposes we can apply the MBMCEM to the Gaussian Mixture Model example of Section 2.2 even though the conditional expectation (38) is tractable.

At iteration k of the MBMCEM algorithm, pick a set I_k and for $i \in I_k$, $m \in [\![0, M_k - 1]\!]$ and $j \in [\![1, J]\!]$, draw a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ from the conditional posterior distribution $p_i(z_i, \theta^{k-1})$ and update the sufficient statistics as $\tilde{S}_i^{1,j}(y_i, z_i^{k,m}) \triangleq \mathbb{1}_{z_i^{k,m}=j}$, $\tilde{S}_i^{2,j}(y_i, z_i^{k,m}) \triangleq \mathbb{1}_{z_i^{k,m}=j} y_i$ and $\tilde{S}_i^{3,j}(y_i, z_i^{k,m}) \triangleq \mathbb{1}_{z_i^{k,m}=j} y_i^2$.
For $i \notin I_k$, $j \in [\![1, J]\!]$, $m \in [\![0, M_k - 1]\!]$ and $d \in [\![1, 3]\!]$, $z_i^{k,m} = z_i^{k-1,m}$and $\tilde{S}_i^{d,j}(y_i, z_i^{k,m}) =$

$\tilde{S}_i^{d,j}(y_i, z_i^{k-1,m})$. Finally the maximisation step yields:

$$\pi_j^k = \frac{\sum_{i=1}^N \tilde{S}_i^{1,j}(y_i, z_i^{k,m})}{N} \ , \tag{47}$$

$$\mu_j^k = \frac{\sum_{i=1}^N \tilde{S}_i^{2,j}(y_i, z_i^{k,m})}{\sum_{i=1}^N \tilde{S}_i^{1,j}(y_i, z_i^{k,m})} \ , \tag{48}$$

$$\sigma_j^k = \frac{\sum_{i=1}^N \tilde{S}_i^{3,j}(y_i, z_i^{k,m})}{\sum_{i=1}^N \tilde{S}_i^{1,j}(y_i, z_i^{k,m})} - (\mu_j^k)^2 \ . \tag{49}$$

4 Numerical examples

4.1 A Linear mixed effects model

4.1.1 The model

We consider, in this section, a linear mixed effects model (Lavielle, 2014). We denote by $y = (y_i \in \mathbb{R}^{n_i}, i \in [\![1, N]\!])$ the observations where for all $i \in [\![1, N]\!]$:

$$y_i = A_i\theta + B_i z_i + \epsilon_i \quad . \tag{50}$$

$A_i \in \mathbb{R}^{n_i \times d}$ and $B_i \in \mathbb{R}^{n_i \times p}$ are design matrices, $\theta \in \mathbb{R}^d$ is a vector of parameters, $z_i \in \mathbb{R}^p$ are the latent data (i.e. the random effects in the context of mixed effects models) which are assumed to be distributed according to a multivariate Gaussian distribution $\mathcal{N}(0, \Omega)$. We also assume that the residual errors $\epsilon_i \in \mathbb{R}^{n_i}$ are distributed according to $\mathcal{N}(0, \Sigma)$ and that the sequences of variables $(z_i, i \in [\![1, N]\!])$ and $(\epsilon_i, i \in [\![1, N]\!])$ are i.i.d. and mutually independent. The covariance matrices Ω and Σ are assumed to be known. For all $i \in [\![1, N]\!]$, the conditional distribution of the observations given the latent variables $y_i|z_i$ and of the latent variables given the observations $z_i|y_i$ are respectively given by:

$$y_i|z_i \sim \mathcal{N}(A_i\theta + B_i z_i, \Sigma), \tag{51}$$

$$z_i|y_i \sim \mathcal{N}(\mu_i, \Gamma_i). \tag{52}$$

where:

$$\Gamma_i = (B_i^\top \Sigma^{-1} B_i + \Omega^{-1})^{-1}, \tag{53}$$

$$\mu_i = \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta). \tag{54}$$

This model belongs to the curved exponential family introduced in section 2.2 where for all $i \in [\![1, N]\!]$:

$$\tilde{S}_i(z_i) \triangleq z_i \quad \text{and} \quad \bar{s}_i(\theta) = \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta) \tag{55}$$

$$\psi_i(\theta) \triangleq (y_i - A_i\theta)^\top \Sigma^{-1}(y_i - A_i\theta) \tag{56}$$

$$\phi_i(\theta) \triangleq B^\top \Sigma^{-1}(y_i - A_i\theta) \tag{57}$$

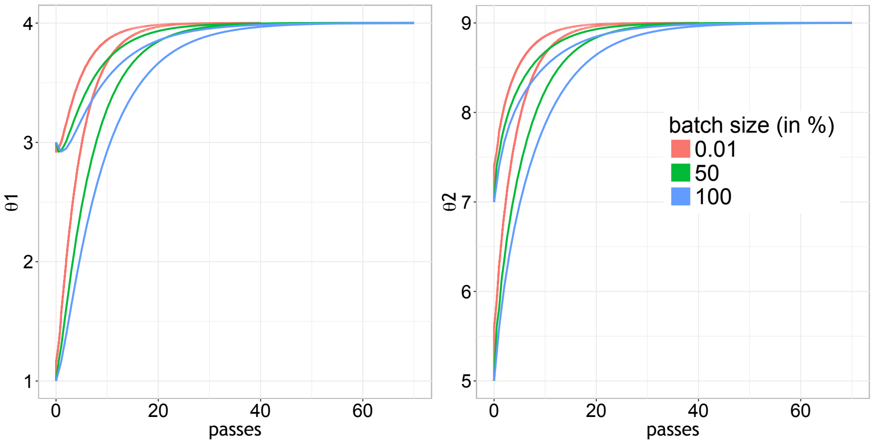

Figure 1: Convergence of the vector of parameter estimates θ^k function of passes over the data.

Maximising $L(s, \theta)$, defined in (30), with respect to θ yields the following maximisation function for all $s = (s_i \in \mathbb{R}^p, i \in [\![1, N]\!])$:

$$\hat{\theta}(s) \triangleq \left(\sum_{i=1}^{N} A_i^\top \Sigma^{-1} A_i \right)^{-1} \sum_{i=1}^{N} A_i^\top \Sigma^{-1} (y_i - B_i s_i).$$

Thus, the $k-th$ update of the MBEM algorithm consists in sampling a subset of indices I_k and computing $\theta^k = \hat{\theta}(s^k)$ where:

$$s_i^k = \begin{cases} \bar{s}_i(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases}$$

4.1.2 Simulation and runs

We generate a synthetic dataset, with $d = 2$, $\theta : (\theta_1 : 4, \theta_2 : 9)$, $N = 10000$ and for all $i \in [\![1, N]\!]$, $n_i = 10$ observations per individual and random design matrices $(A_i, i \in [\![1, N]\!])$ and $(B_i, i \in [\![1, N]\!])$. Two runs of the MBEM are executed starting from different initial values $((\theta_1^0 : 1, \theta_2^0 : 5)$ and $(\theta_1^0 : 3, \theta_2^0 : 7))$ to study the convergence behaviour of these algorithms depending on the initialisation. Figure 1 shows the convergence of the vector of parameter estimates $(\theta_1^k, \theta_2^k)_{k=0}^K$ over passes of the EM algorithm, the MBEM algorithm where half of the data is considered at each iteration and the Incremental EM algorithm (i.e. a single data point is considered at each iteration). The speed of convergence is a monotone function of the batch size in this case, the smaller the batch the faster the convergence.

4.2 Logistic regression for a binary variable

4.2.1 The model

Let $y = (y_i, i \in [\![1, N]\!])$ be the vector of binary responses where for each individual i, $y_i = (y_{ij}, j \in [\![1, n_i]\!])$ is a sequence of conditionally independent random variables taking values in $\{0, 1\}$ which corresponds to the j-th responses for the i-th subject. We consider a logistic regression problem in which the parameters depend upon each individual i. Denote by $z_i = (z_{i,d} \in \mathbb{R}, d \in [\![1, m]\!])$ the vector of regression coefficients (the latent data) for individual i and $(d_{ij} \in \mathbb{R}, j \in [\![1, n_i]\!])$ the associated explanatory variables. The conditional distribution of the observations y_i given the latent variables z_i is given by:

$$\operatorname{logit}(\mathbb{P}(y_{ij} = 0 | z_i)) = d_{ij}^\top z_i .$$

For all $i \in [\![1, N]\!]$, we assume that z_i are independent and marginally distributed according to $\mathcal{N}(\beta, \Omega)$ where $\beta = (\beta_d, d \in [\![1, D]\!])$ and $\Omega = \operatorname{diag}(\omega_d, d \in [\![1, D]\!])$. The complete log-likelihood is expressed as:

$$\log f(z, \theta) \propto \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ y_{ij} d_{ij}^\top z_i - \log(1 + e^{d_{ij}^\top z_i}) \right\} \tag{58}$$

$$- \sum_{i=1}^{N} \left\{ \frac{1}{2} \log(|\Omega|) + \frac{1}{2} \operatorname{Tr} \left(\Omega^{-1} (z_i - \beta)(z_i - \beta)^\top \right) \right\} . \tag{59}$$

Since the expectation of the complete log likelihood with respect to the conditional distribution of the latent variables given the observations is intractable, we use the MCEM and the MBMCEM algorithms, which require to simulate random draws from this conditional distribution. We use the saemix R package (Comets et al., 2017) to run a Metropolis-Hastings within Gibbs sampler (Brooks et al., 2011) where for all $i \in [\![1, N]\!]$ and dimension $d \in [\![1, D]\!]$ of the parameter, the Markov Chain is constructed as follows:

```
Algorithm 6 Random Walk Metropolis
Initialisation: given the current parameter estimates $(\beta_d^{k-1}, \omega_d^{k-1})$ set $\mathcal{T}_d^{(0)} = \omega_d^{k-1}$ and sample the initial state $z_{i,d}^{(0)} \sim \mathcal{N}(\beta_d^{k-1}, \mathcal{T}_d^{(0)})$
Iteration t: given the current chain state $z_d^{(t-1)}$:
    1. Sample a candidate state:
           $z_{i,d}^{(c)} \sim \mathcal{N}(z_{i,d}^{(t-1)}, \mathcal{T}_d^{(t-1)})$    (60)
    2. Accept with probability $\min\left(1, \alpha^{(t)}(z_{i,d}^{(c)}, z_{i,d}^{(t-1)})\right)$
    3. Update the variance of the proposal as follows:
           $\mathcal{T}_d^{(t)} = \mathcal{T}_d^{(t-1)} + (1 + \delta(\alpha^{(t)} - \alpha^*))$    (61)
```

where $\delta = 0.4$, α^t is the MH acceptance ratio and $\alpha^* = 0.4$ is the optimal acceptance ratio (Robert and Casella, 2005). This model belongs to the curved exponential family introduced in section 2.2 where for all $i \in [\![1, N]\!]$, $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$. At iteration k, the MBMCEM algorithm consists in:

1. Picking a set I_k uniformly on $\{A \subset [\![1, N]\!], \text{card}(A) = p\}$.
2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sampling a Markov Chain $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ using algorithm 6.
3. Computing $s_i^k = (s_i^{1,k}, s_i^{2,k})$ such as:

$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k}\sum_{m=0}^{M_k-1} z_i^{k,m}, \frac{1}{M_k}\sum_{m=0}^{M_k-1} (z_i^{k,m})^\top z_i^{k,m}\right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases} \tag{62}$$

4. Updating the parameters as follows:

$$\beta^k = \frac{1}{N}\sum_{i=1}^{N} s_i^{1,k} \,, \tag{63}$$

$$\Omega^k = \frac{1}{N}\sum_{i=1}^{N} s_i^{2,k} - (\beta^k)^\top \beta^k \,. \tag{64}$$

4.2.2 Simulation and runs

In the sequel, $D = 3$, $N = 1200$ and for all $i \in [\![1, N]\!]$, $n_i = 15$. For all $i \in [\![1, N]\!]$ and $j \in [\![1, n_i]\!]$, we take $d_{ij,1} = 1$, $d_{ij,2} = -20 + 5(j-1)$ and $d_{ij,3} = 10\lceil 3i/N \rceil$. The data are

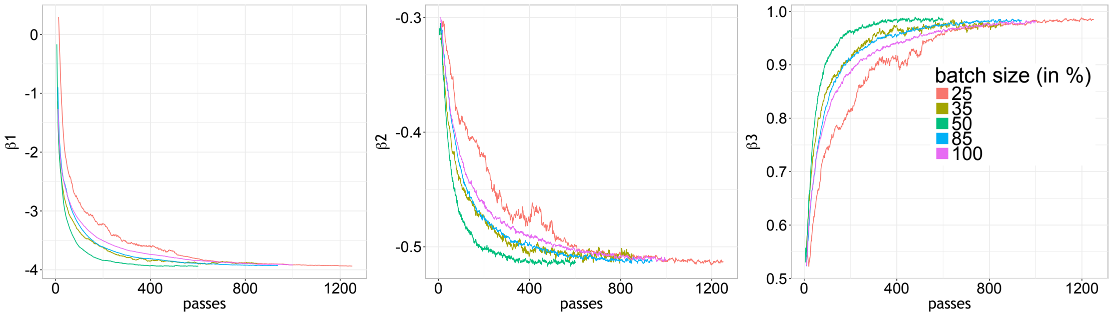

Figure 2: Convergence of the vector of fixed parameters β for different batch sizes function of passes over the data.

generated using the following values for the fixed and random effects ($\beta_1 = -4, \beta_2 = -0.5, \beta_3 = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2$). We generate a synthetic dataset using the following generating values for the fixed and random effects ($\beta_1 = -4, \beta_2 = -0.5, \beta_3 = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2$). We run the MBMCEM algorithm in this section. The size of the Monte Carlo batch increases polynomially, $M_k \triangleq M_0 + k^2$ with $M_0 = 50$. Figure 2 shows the convergence of the fixed effects ($\beta_1, \beta_2, \beta_3$) estimates obtained with both the MCEM and the MBMCEM algorithms for different batch sizes. The effect of the batch size on the convergence rate differs from the previous example. Whereas smaller batches implied faster convergence for the mini batch EM algorithm, here, an optimal batch size of 50% accelerates the algorithm. Figure 2 highlights a non monotonic evolution of the convergence rate with respect to the size of the batch.

5 Proofs

5.1 Proof of Theorem 1

5.1.1 Proof of (i)

First, let us define for $\theta \in \Theta$

$$\bar{A}^k(\theta) \triangleq \sum_{i=1}^{N} A_i^k(\theta) \ , \tag{65}$$

where for all $i \in [\![1, N]\!]$, A_i^k is defined in (24). For any $k \geq 1$ and for all $\theta \in \Theta$ the following decomposition plays a key role:

$$\bar{A}^k(\theta) = \bar{A}^{k-1}(\theta) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta) - \sum_{i \in I_k} A_i^{k-1}(\theta). \tag{66}$$

Since by construction $\bar{A}^k(\theta^k) \leq \bar{A}^k(\theta^{k-1})$, we get:

$$\bar{A}^k(\theta^k) \leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}). \tag{67}$$

Since for $i \in I_k$, $B_{i,\theta^{k-1}}$ is a surrogate of ℓ_i at θ^{k-1} we get that $B_{i,\theta^{k-1}}(\theta^{k-1}) = \ell_i(\theta^{k-1})$. On the other hand, for $i \in [\![1, N]\!]$, $A_i^{k-1} \equiv B_{i,\theta^{\tau_{i,k-1}}}$ and $B_{i,\theta^{\tau_{i,k-1}}}$ is a surrogate of ℓ_i at $\theta^{\tau_{i,k-1}}$, thus we obtain that $\ell_i(\theta^{k-1}) - A_i^{k-1}(\theta^{k-1}) \leq 0$. Plugging these two relations in (67) we obtain:

$$\begin{aligned} \bar{A}^k(\theta^k) &\leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} \ell_i(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}) && (68) \\ &\leq \bar{A}^{k-1}(\theta^{k-1}) \ . && (69) \end{aligned}$$

As a result, the sequence $\left(\bar{A}^k(\theta^k)\right)_{k \geq 0}$ is monotonically decreasing. Since, under assumption M 3, this quantity is bounded from below with probability one, we obtain its almost sure convergence. Taking the expectations with respect to the sampling distributions of the previous inequalities implies the convergence of the (deterministic) sequence $\left(\mathbb{E}[\bar{A}^k(\theta^k)]\right)_{k \geq 0}$. Let us denote for all $\theta \in \Theta$ and a subset $J \subset [\![1, N]\!]$:

$$\begin{aligned} \ell_J(\theta) &\triangleq \sum_{i \in J} \ell_i(\theta) \ , && (70) \\ A_J^{k-1}(\theta) &\triangleq \sum_{i \in J} A_i^{k-1}(\theta) \ . && (71) \end{aligned}$$

Inequality (67) gives :

$$0 \leq \sum_{k=1}^{n} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \leq \sum_{k=1}^{n} \bar{A}^{k-1}(\theta^{k-1}) - \bar{A}^k(\theta^k) = \bar{A}^0(\theta^0) - \bar{A}^n(\theta^n) \ . \tag{72}$$

Consequently, the sum of positive terms $\left(\sum_{k=1}^{n} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right)_{n\geq 1}$ converges almost surely and $\left(A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right)_{k\geq 1}$ converges almost surely to zero. The Beppo-Levi theorem and the Tower property of the conditional expectation imply:

$$\mathsf{M} \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] \tag{73}$$

$$= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \middle| \mathcal{F}_{k-1}\right]\right] , \tag{74}$$

with

$$\mathbb{E}\left[\ell_{I_k}(\theta^{k-1}) \middle| \mathcal{F}_{k-1}\right] = \frac{p}{N}\ell(\theta^{k-1}) ,$$

$$\mathbb{E}\left[[A_{I_k}^{k-1}(\theta^{k-1}) \middle| \mathcal{F}_{k-1}\right] = \frac{p}{N}\sum_{i=1}^{N} A_i^{k-1}(\theta^{k-1}) = \frac{p}{N}\bar{A}^{k-1}(\theta^{k-1}) ,$$

where $\mathcal{F}_{k-1} = \sigma(I_j, j \leq k-1)$ is the filtration generated by the sampling of the indices. We thus obtain:

$$\mathsf{M} = \frac{p}{N}\sum_{k=0}^{\infty} \mathbb{E}\left[\bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] = \frac{p}{N}\mathbb{E}\left[\sum_{k=0}^{\infty} \bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] < \infty . \tag{75}$$

This last equation shows that:

$$\lim_{k\to\infty} \bar{A}^{k}(\theta^{k}) - \ell(\theta^{k}) = 0 \quad \text{a.s.} \tag{76}$$

which implies the almost sure convergence of $\left(\ell(\theta^k)\right)_{k\geq 0}$.

5.1.2 Proof of (ii)

Let us define, for all $k \geq 0$, $\bar{h}_k$ as:

$$\bar{h}^k : \vartheta \to \sum_{i=1}^{N} A_i^k(\vartheta) - \ell_i(\vartheta) . \tag{77}$$

$\bar{h}^k$ is L-smooth with $L = \sum_{i=1}^{N} L_i$ since each of its component is L_i-smooth by definition of the surrogate functions. Using the particular parameter $\vartheta^k = \theta^k - \frac{1}{L}\nabla\bar{h}_k(\theta^k)$ we have the following classical inequality for smooth functions (cf. Lemma 1.2.3 in (Nesterov, 2007)):

$$0 \leq \bar{h}^k(\vartheta^k) \leq \bar{h}^k(\theta^k) - \frac{1}{2L}\|\nabla\bar{h}^k(\theta^k)\|_2^2 \tag{78}$$

$$\implies \|\nabla\bar{h}^k(\theta^k)\|_2^2 \leq 2L\bar{h}^k(\theta^k) . \tag{79}$$

Using (76), we conclude that $\lim_{k\to\infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle$ for any $\theta \in \Theta$ yields:

$$\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle = \langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \ . \tag{80}$$

Note that θ^k is the result of the minimisation of the sum of surrogates $\bar{A}^k(\theta)$ on the constrained set Θ, therefore $\langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k \rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\begin{align} \langle \nabla \ell(\theta^k), \theta - \theta^k \rangle &\geq -\langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \tag{81} \\ &\geq -\|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \ . \tag{82} \end{align}$$

By minimising over Θ and taking the infimum limit on k, we get:

$$\liminf_{k\to\infty} \inf_{\theta\in\Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq - \lim_{k\to\infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0 \ , \tag{83}$$

which is the Asymptotic Stationary Point Condition (ASPC).

5.2 Proof of Theorem 2

We preface the proof by the following lemma which is of independent interest:

Lemma 1. *Let $(V_k)_{k\geq 0}$ be a non negative sequence of random variables such that $\mathbb{E}[V_0] < \infty$. Let $(X_k)_{k\geq 0}$ a non negative sequence of random variables and $(E_k)_{k\geq 0}$ be a sequence of random variables such that $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|] < \infty$. If for any $k \geq 1$:*

$$V_k \leq V_{k-1} - X_k + E_k \tag{84}$$

then:

(i) for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.

(ii) the sequence $(\mathbb{E}[V_k])_{k\geq 0}$ converges and $\lim\limits_{k\to\infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.

(iii) the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.

Remark 2. *Note that the result still holds if $(V_k)_{k\geq 0}$ is a sequence of random variables which is bounded from below by a deterministic quantity $M \in \mathbb{R}$.*

Proof. We first show that for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$. Note indeed that:

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \ , \tag{85}$$

showing that $\mathbb{E}[V_k] \leq \mathbb{E}[V_0] + \mathbb{E}\left[\sum_{j=1}^{k} E_j\right] < \infty$.

Since $0 \leq X_k \leq V_{k-1} - V_k + E_k$ we also obtain for all $k \geq 0$, $\mathbb{E}[X_k] < \infty$. Moreover, since $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$, the series $\sum_{j=1}^{\infty} E_j$ converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \ . \tag{86}$$

Note that $\mathbb{E}[|W_k|] \leq \mathbb{E}[V_k] + \mathbb{E}\left[\sum_{j=k+1}^{\infty} |E_j|\right] < \infty$. For all $k \geq 1$, we get:

$$W_k \leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \ , \tag{87}$$

$$\mathbb{E}[W_k] \leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \ . \tag{88}$$

Hence the sequences $(W_k)_{k\geq 0}$ and $(\mathbb{E}[W_k])_{k\geq 0}$ are non increasing. Since for all $k \geq 0$, $W_k \geq -\sum_{j=1}^{\infty} |E_j| > -\infty$ and $\mathbb{E}[W_k] \geq -\sum_{j=1}^{\infty} \mathbb{E}[|E_j|] > -\infty$, the (random) sequence

$(W_k)_{k\geq 0}$ converges a.s. to a limit W_∞ and the (deterministic) sequence $(\mathbb{E}[W_k])_{k\geq 0}$ converges to a limit w_∞. Since $|W_k| \leq V_0 + \sum_{j=1}^{\infty} |E_j|$, the Fatou lemma implies that:

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \ , \tag{89}$$

showing that the random variable W_∞ is integrable.

In the sequel, set $U_k \triangleq W_0 - W_k$. By construction we have for all $k \geq 0$, $U_k \geq 0$, $U_k \leq U_{k+1}$ and $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$ and by the monotone convergence theorem, we get:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \ . \tag{90}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \ , \tag{91}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (87) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\sum_{j=1}^{\infty} X_j \leq W_0 - W_\infty < \infty \ , \tag{92}$$

$$\sum_{j=1}^{\infty} \mathbb{E}[X_j] \leq \mathbb{E}[W_0] - w_\infty < \infty \ , \tag{93}$$

which concludes the proof of the lemma.

□

5.2.1 Proof of (i)

To study the convergence of the MBMCEM algorithm, we consider for all $k \geq 1$, the function $\vartheta \to \hat{B}_{i,\theta^{k-1}}(\vartheta)$ defined for all $i \in I_k$ and $\vartheta \in \Theta$ by:

$$\hat{B}_{i,\theta^{k-1}}(\vartheta) \triangleq \hat{Q}_i^k(\vartheta, \theta^{k-1}) + H_i(\theta^{k-1}) \tag{94}$$

$$= -\frac{1}{M_k} \sum_{m=0}^{M_k-1} \log p_i(z_i^{k,m}, \vartheta) + l_i(\vartheta) + H_i(\theta^{k-1}) \ , \tag{95}$$

where $H_i(\theta^{k-1})$ is defined by (19). This function is a Monte Carlo approximation of the surrogate function $B_{i,\theta^{k-1}}$ defined for all $\vartheta \in \Theta$ and $i \in I_k$ as:

$$B_{i,\theta^{k-1}}(\vartheta) \triangleq -\int_{\mathsf{Z}_i} \log p_i(z_i, \vartheta) p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) + l_i(\vartheta) + H_i(\theta^{k-1}) \tag{96}$$

$$= \mathrm{KL}\big(\mathbb{P}_{i,\theta^{k-1}} \,\big\|\, \mathbb{P}_{i,\vartheta}\big) + l_i(\vartheta) \ . \tag{97}$$

Under assumption M 2, $\vartheta \to \hat{B}_{i,\theta^{k-1}}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$. Let us define, for $\theta \in \Theta$, $\hat{A}^k(\theta) \triangleq \sum_{i=1}^N \hat{A}_i^k(\theta)$ where $\hat{A}_i^k(\vartheta)$ are defined recursively as follows:

$$\hat{A}_i^k(\vartheta) = \begin{cases} \hat{B}_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ \hat{A}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{98}$$

(94) implies that for $k \geq 1$, $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N \hat{A}_i^k(\vartheta)$. Set for all $\theta \in \Theta$, $i \in [\![1, N]\!]$ and $k \geq 1$:

$$A_i^k(\theta) \triangleq B_{i,\theta^{\tau_{i,k}}}(\theta) \quad \text{and} \quad \bar{A}^k(\theta) = \sum_{i=1}^N A_i^k(\theta) \ , \tag{99}$$

where $\tau_{i,k}$ is defined by (10). For any $k \geq 1$ and $\theta \in \Theta$ the following decomposition plays a key role:

$$\hat{A}^k(\theta) = \hat{A}^{k-1}(\theta) + \sum_{i \in I_k} \{\hat{B}_{i,\theta^{k-1}}(\theta) - \hat{A}_i^{k-1}(\theta)\} \ . \tag{100}$$

Set the following notations:

$$\begin{aligned} V_k &\triangleq \bar{A}^k(\theta^k) \ , \\ X_k &\triangleq -\sum_{i \in I_k} \{B_{i,\theta^{k-1}}(\theta^{k-1}) - A_i^{k-1}(\theta^{k-1})\} \ , \\ E_k &\triangleq \sum_{i \in I_k} \{\hat{B}_{i,\theta^{k-1}}(\theta^{k-1}) - B_{i,\theta^{k-1}}(\theta^{k-1})\} \\ &\quad + \sum_{i \in I_k} \{A_i^{k-1}(\theta^{k-1}) - \hat{A}_i^{k-1}(\theta^{k-1})\} \\ &\quad + \bar{A}^k(\theta^k) - \hat{A}^k(\theta^k) + \hat{A}^{k-1}(\theta^{k-1}) - \bar{A}^{k-1}(\theta^{k-1}) \ . \end{aligned}$$

Combining (100) with $\bar{A}^k(\theta^k) = \bar{A}^k(\theta^k) - \hat{A}^k(\theta^k) + \hat{A}^k(\theta^k)$ and $\hat{A}^k(\theta^k) \leq \hat{A}^k(\theta^{k-1})$, we obtain:

$$V_k \leq V_{k-1} - X_k + E_k \ , \tag{101}$$

where A_i^{k-1} and $\bar{A}^k$ are defined in (99). We now check the assumptions of Lemma 1. Note first that the sequence $(V_k)_{k \geq 0}$ is bounded from below under assumption M 3. We now check that $X_k \geq 0$ thanks to the following relation:

$$\begin{aligned} X_k &= -0 - \sum_{i \in I_k} \ell_i(\theta^{k-1}) + \sum_{i \in I_k} \mathrm{KL}\big(\mathbb{P}_{i,\theta^{\tau_{i,k-1}}} \, \big\| \, \mathbb{P}_{i,\theta^{k-1}}\big) + \sum_{i \in I_k} \ell_i(\theta^{k-1}) && (102) \\ &= \sum_{i \in I_k} \mathrm{KL}\big(\mathbb{P}_{i,\theta^{\tau_{i,k-1}}} \, \big\| \, \mathbb{P}_{i,\theta^{k-1}}\big) \geq 0 \ . && (103) \end{aligned}$$

We finally have to prove the convergence of the series $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|]$. For this purpose, we will show that for all $i \in [\![1, N]\!]$:

$$\sum_{k=0}^{\infty} \mathbb{E}\Big[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\Big] < \infty \,. \tag{104}$$

We have, using the Tower property of the conditional expectation and the Jensen inequality:

$$\mathbb{E}\Big[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\Big] \leq \mathbb{E}\Big[\mathbb{E}_{i,\theta^{\tau_{i,k}}}\Big[\sup_{\vartheta\in\Theta} |\hat{A}_i^k(\vartheta) - A_i^k(\vartheta)|\Big]\Big] \,. \tag{105}$$

Under assumption M 4 applied with the function $\vartheta \to \hat{A}_i^k(\vartheta)$, for all $i \in [\![1, N]\!]$ we have:

$$\mathbb{E}_{i,\theta^{\tau_{i,k}}}\Big[\sup_{\vartheta\in\Theta} |\hat{A}_i^k(\vartheta) - A_i^k(\vartheta)|\Big] \leq C_i M_{\tau_{i,k}}^{-1/2} \,, \tag{106}$$

where C_i is a finite constant defined by (44) and $\tau_{i,k}$ is defined by (10). Thus, we have that:

$$\mathbb{E}\Big[|\hat{A}_i^k(\theta^k) - A_i^k(\theta^k)|\Big] \leq C_i \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] \,. \tag{107}$$

Since, any index i is included in a mini-batch with a probability equal to $\frac{p}{N}$ conditionally independently from the past, we obtain that:

$$\mathbb{E}[M_{\tau_{i,k}}^{-1/2}] = \sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \,. \tag{108}$$

Taking the infinite sum of this term yields:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] &= \sum_{k=1}^{\infty}\sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} && (109)\\
&= \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{k-(l+1)} \frac{p}{N} \mathbb{1}_{\{l\leq k-1\}} M_l^{-1/2} && (110)\\
&= \frac{p}{N} \sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{-(l+1)} M_l^{-1/2} \sum_{k=l+1}^{\infty} \left(1 - \frac{p}{N}\right)^{k} && (111)\\
&= \sum_{l=0}^{\infty} M_l^{-1/2} \,, && (112)
\end{aligned}$$

which proves identity (104), using assumption M 5. By summing over the indices $i \in [\![1, N]\!]$, (104) implies:

$$\sum_{k=0}^{\infty} \mathbb{E}\Big[|\hat{A}^k(\theta^k) - \bar{A}^k(\theta^k)|\Big] < \infty \,. \tag{113}$$

Hence, we obtain that $\sum_{k=0}^{\infty} |\hat{A}^k(\theta^k) - \bar{A}^k(\theta^k)| < \infty$ almost surely which implies that:

$$\lim_{k\to\infty} \hat{A}^k(\theta^k) - \bar{A}^k(\theta^k) = 0 \quad \text{a.s.} \tag{114}$$

Similarly, using assumption M 4 applied for all $i \in [\![1, N]\!]$, with the function $\vartheta \to \nabla \hat{A}_i^k(\vartheta)$ we obtain:

$$\lim_{k\to\infty} \nabla\hat{A}^k(\theta^k) - \nabla\bar{A}^k(\theta^k) = 0 \quad \text{a.s.} \tag{115}$$

It follows from (104) and (113) that $\sum_{k=0}^{\infty} \mathbb{E}\left[|E_k|\right] < \infty$ and that the series $\sum_{k=0}^{\infty} \epsilon_k$ converges to an almost surely finite limit. Hence by Lemma 1 and (114) we get:

- the sequence $\left(\bar{A}^k(\theta^k)\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \chi_k$ converge a.s.
- the sequence $\left(\mathbb{E}\left[\bar{A}^k(\theta^k)\right]\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \mathbb{E}\left[X_k\right]$ converge with

$$\lim_{k\to\infty} \mathbb{E}\left[\bar{A}^k(\theta^k)\right] = \mathbb{E}[\lim_{k\to\infty} \bar{A}^k(\theta^k)] \ .$$

- the sequence $\left(\hat{A}^k(\theta^k)\right)_{k\geq 0}$ converges a.s. and the sequence $\left(\mathbb{E}\left[\hat{A}^k(\theta^k)\right]\right)_{k\geq 0}$ converges.

Now, we have to prove the almost-sure convergence of the sequence $\left(\ell(\theta^k)\right)_{k\geq 0}$ and the convergence of $\left(\mathbb{E}\left[\ell(\theta^k)\right]\right)_{k\geq 0}$. Using the same argument as in (73) and (75), we have:

$$\mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] = \frac{p}{N}\,\mathbb{E}\left[\sum_{k=1}^{\infty} \{\bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\}\right] < \infty \ , \tag{116}$$

which yields to:

$$\lim_{k\to\infty} \mathbb{E}\left[\bar{A}^k(\theta^k) - \ell(\theta^k)\right] = 0 \ , \tag{117}$$

$$\lim_{k\to\infty} \bar{A}^k(\theta^k) - \ell(\theta^k) = 0 \quad \text{a.s.} \ , \tag{118}$$

showing that the sequence $\left(\mathbb{E}\left[\ell(\theta^k)\right]\right)_{k\geq 0}$ converges and that $\left(\ell(\theta^k)\right)_{k\geq 0}$ converges a.s.

5.2.2 Proof of (ii)

Consider for any $k \geq 0$, the L smooth function $\bar{h}^k$ defined by (77). Using (78) and (117) we get $\lim\limits_{k\to\infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle$ for any $\theta \in \Theta$ yields:

$$\begin{aligned} \langle \nabla \ell(\theta^k), \theta - \theta^k \rangle &= \langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle & (119) \\ &= \langle \nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k), \theta - \theta^k \rangle & (120) \\ &\quad + \langle \nabla \hat{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \ . & (121) \end{aligned}$$

Note that θ^k is the result of the minimisation of $\hat{A}^k(\theta)$ on the constrained set Θ, therefore for all $\theta \in \Theta$, $\langle \nabla \hat{A}^k(\theta^k), \theta - \theta^k \rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle \geq \langle \nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \tag{122}$$

$$\geq -\|\nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 - \|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \ . \tag{123}$$

By minimising over Θ and taking the infimum limit, we get, using (115):

$$\liminf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq - \lim_{k \to \infty} \left(\|\nabla \bar{A}^k(\theta^k) - \nabla \hat{A}^k(\theta^k)\|_2 + \|\nabla \bar{h}^k(\theta^k)\|_2 \right) = 0 \ , \tag{124}$$

which is the Asymptotic Stationary Point Condition (ASPC).

6 Conclusion

We have presented in this article, almost-sure convergence guarantees for the mini-batch variant of the EM algorithm, that were rather assumed than proved prior this work. To do so, we applied the incremental framework developed in (Mairal, 2015) under some mild assumptions on the model. We also, extended that framework to stochastic surrogates in order to prove the convergence of the mini-batch MCEM algorithm.

We believe that this stochastic incremental framework is of independent interest to analyse a wide class of optimisation algorithms based on the computation of stochastic surrogate functions. Moreover, non asymptotic rates in the nonconvex and incremental settings are particularly interesting to highlight the influence of the mini-batch size.

References

Ablin, P., Gramfort, A., Cardoso, J.-F., and Bach, F. (2018). Em algorithms for ica. *arXiv preprint arXiv:1805.10054*.

Baey, C., Trevezas, S., and Cournède, P.-H. (2016). A non linear mixed effects model of plant growth and estimation via stochastic variants of the em algorithm. *Communications in Statistics-Theory and Methods*, 45(6):1643–1669.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Booth, J. G., Hobert, J. P., and Jank, W. (2001). A survey of monte carlo algorithms for maximizing the likelihood of a two-stage hierarchical model. *Statistical Modelling*, 1(4):333–349.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L., editors (2011). *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL.

Cappé, O. (2011). Online em algorithm for hidden markov models. *Journal of Computational and Graphical Statistics*, 20(3):728–749.

Cappé, O. and Moulines, E. (2009). On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613.

Chakraborty, A. and Das, K. (2010). Inferences for joint modelling of repeated ordinal scores and time to event data. *Computational and mathematical methods in medicine*, 11(3):281–295.

Chan, K. and Ledolter, J. (1995). Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252.

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software*, 80(3):1–42.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions*, (suppl. 1):205–237. Recent results in estimation theory and related topics.

Defazio, A., Bach, F. R., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in*

Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 1646–1654.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 393–427. Gauthier-Villars.

Fisher, R. A. (1925). *Theory of statistical estimation*, volume 22.

Fort, G., Moulines, E., et al. (2003). Convergence of the monte carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220–1259.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research*, 6:2049–2073.

Hsiao, T., Khurd, P., Rangarajan, A., and Gindi, G. (2006). An overview of fast convergent ordered-subsets reconstruction methods for emission tomography based on the incremental em algorithm. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 569(2):429–433.

Hughes, J. P. (1999). Mixed effects models with censored data with application to hiv rna levels. *Biometrics*, 55(2):625–629.

Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press.

Levine, R. A. and Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439.

Likas, A. and Galatsanos, N. (2004). A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on signal processing*, 52(8):2222–2233.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 226–233.

Mairal, J. (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92(437):162–170.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Neal, R. and Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, MI, editor, *Learning in Graphical Models*, volume 89 of *NATO advanced science institutes series, series D, Behavioral and Social Sciences*, pages 355–368, PO BOX 17, 3300 AA Dordrecht, Netherlands. NATO, Springer. NATO Advanced Study Institute on Learning in Graphical Models, Erice, Italy, Sep 27-Oct 07, 1996.

Neath, R. C. et al. (2013). On convergence properties of the monte carlo em algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pages 43–62. Institute of Mathematical Statistics.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Ng, S. and McLachlan, G. (2003). On the choice of the number of blocks with the incremental EM algorithm for the fitting of normal mixtures. *Statistics and Computing*, 13(1):45–55.

Ng, S. and McLachlan, G. (2004). Speeding up the EM algorithm for mixture model-based segmentation of magnetic resonance images. *Pattern Recognition*, 37(8):1573–1589.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Roux, N. L., Schmidt, M., and Bach, F. R. (2012). A stochastic gradient method with an exponential convergence _rate for finite training sets. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc.

Sherman, R. P., Ho, Y.-Y. K., and Dalal, S. R. (1999). Conditions for convergence of monte carlo em sequences with an application to product diffusion modeling. *The Econometrics Journal*, 2(2):248–267.

Thiesson, B., Meek, C., and Heckerman, D. (2001). Accelerating EM for large databases. *Machine Learning*, 45(3):279–299.

Vlassis, N. and Likas, A. (2002). A greedy EM algorithm for Gaussian mixture learning. *Neural Processing Letters*, 15(1):77–87.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2014). High dimensional expectation-maximization algorithm: Statistical optimization and asymptotic normality. *arXiv preprint arXiv:1412.8729*.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Zhu, R., Wang, L., Zhai, C., and Gu, Q. (2017). High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 4180–4188, International Convention Centre, Sydney, Australia. PMLR.

Glossary

directional derivative Consider a function $f : \Theta \to \mathbb{R}$. For all $(\theta, \theta') \in \Theta^2$, the following limit is called the directional derivative of f at θ in the direction $\theta' - \theta$: $\nabla f(\theta, \theta' - \theta) \triangleq \lim_{t \to 0} (f(\theta + t(\theta' - \theta)) - f(\theta))/t$. 1

iterative application Let $\mathcal{X}$ be a set and $x_0 \in \mathcal{X}$ a given point. Then an iterative algorithm A with initial point x_0 is a point-to-set mapping $A \colon \mathcal{X} \to \mathcal{X}$ which generates a sequence $\{x_n\}_{n=1}^{\infty}$ according to

$$x_{n+1} \in A(x_n) \tag{125}$$

. 1, 8, 13

smooth A function $f : \Theta \to \mathbb{R}$ is called L-smooth when it is differentiable and when its gradient ∇f is L-Lipschitz continuous.. 1, 6, 8