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

January 11, 2019

Abstract

Abstract

1 Convergence of the mini-batch EM algorithm

1.1 MBEM for a curved exponential family

In the particular case where for all $i \in [\![1, N]\!]$ 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 $i \in [\![1, N]\!]$ and $\theta \in \Theta$:*

$$\log f_i(z_i, \theta) = H_i(z_i) - \psi_i(\theta) + \langle \tilde{S}_i(z_i), \phi_i(\theta) \rangle. \tag{1}$$

where $\psi_i : \Theta \mapsto \mathbb{R}$ and $\phi_i : \Theta \mapsto \mathbb{R}^{d_i}$ 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 $\mathbb{R}^{d_i}$ 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 $i \in [\![1, N]\!]$ the function $\bar{s}_i : \Theta \to \mathsf{S}_i$ as:

$$\bar{s}_i(\theta) \triangleq \int_{\mathsf{Z}_i} \tilde{S}_i(z_i) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i). \tag{2}$$

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

$$L(s; \theta) \triangleq \sum_{i=1}^{N} \psi_i(\theta) - \sum_{i=1}^{N} \langle s_i, \phi_i(\theta) \rangle. \tag{3}$$

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

$$L(s; \hat{\theta}(s)) \leq L(s; \theta). \tag{4}$$

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 1 mini-batch EM for a curved exponential family

Initialisation: given an initial parameter estimate θ^0, for all $i \in [\![1, N]\!]$ 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 $\{A \subset [\![1, N]\!], \text{card}(A) = p\}$.
2. For $i \in [\![1, N]\!]$, compute s_i^k such as:

$$s_i^k = \begin{cases} \bar{s}_i(\theta^{k-1}) & \text{if } i \in I_k. \\ s_i^{k-1} & \text{otherwise.} \end{cases} \tag{5}$$

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

1.2 MBEM, oEM and oEM-vr

To study those three algorithms we assume a slightly different exponential family as in (6):

$$L(s; \theta) \triangleq \psi(\theta) - \langle \sum_{i=1}^{N} s_i, \phi(\theta) \rangle. \tag{6}$$

Denote $s = \sum_{i=1}^N s_i$ and $\bar{s}_i(\theta) \triangleq \mathbb{E}_{p(z_i,\theta)}[\tilde{S}(z_i)]$. At iteration k of epoch e, the E-step of those three algorithms consists in picking a single data i_k and:

- **oEM algorithm:** $s^k = s^{k-1} + \rho_k(\bar{s}_{i_k}(\theta^{k-1}) - s^{k-1})$
- **oEM-vr algorithm:** $s^{e,k} = s^{e,k-1} + \rho(\bar{s}_{i_k}(\theta^{k-1}) - s_{i_k}^{e,0} + s^{e,0} - s^{k-1})$
- **MBEM algorithm:** $s^{e,k} = s^{e,k-1} + (\bar{s}_{i_k}(\theta^{k-1}) - s_{i_k}^{k-1})$

where $s^{e,0} = s^{e-1,M}$.

1.3 Local convergence of MBEM

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 $f_i(z_i, \theta)$:

$$f_i(z_i, \theta) = \prod_{j=1}^{J} (\pi_j c^j(y_i, \delta))^{\mathbb{1}_{z_i=j}} . \tag{7}$$

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:

$$\log f_i(z_i, \theta) = \sum_{j=1}^{J} \mathbb{1}_{z_i=j} \log(\pi_j) + \sum_{j=1}^{J} \mathbb{1}_{z_i=j} \left[-\frac{(y_i - \mu_j)^2}{2\sigma_j^2} - \frac{1}{2} \log \sigma_j^2 \right] . \tag{8}$$

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 1 consists in picking a set I_k and for $i \in I_k$, computing the following quantities:

$$(\bar{s}_i^k)^{1,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) , \tag{9}$$

$$(\bar{s}_i^k)^{2,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} y_i p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i , \tag{10}$$

$$(\bar{s}_i^k)^{3,j} = \int_{\mathsf{Z}_i} \mathbb{1}_{z_i=j} y_i^2 p_i(z_i, \theta^{k-1}) \mu_i(\mathrm{d}z_i) = p_{ij}(\theta^{k-1}) y_i^2 , \tag{11}$$

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

$$p_{ij}(\theta^{k-1}) = \frac{\mathbb{P}_i(z_i = j) p_i(y_i | z_i = j; \theta^{k-1})}{p_i(y_i; \theta^{k-1})} = \frac{\pi_j^{k-1} c^j(y_i; \mu_j^{k-1}, \sigma_j^{k-1})}{\sum_{l=1}^{J} \pi_l^{k-1} c^l(y_i; \mu_l^{k-1}, \sigma_l^{k-1})} . \tag{12}$$

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:

$$\pi_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}}{N} , \tag{13}$$

$$\mu_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{2,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} , \tag{14}$$

$$\sigma_j^k = \frac{\sum_{i=1}^{N} (\bar{s}_i^k)^{3,j}}{\sum_{i=1}^{N} (\bar{s}_i^k)^{1,j}} - (\mu_j^k)^2 . \tag{15}$$

2 Numerical examples

2.1 A Linear mixed effects model

2.1.1 The model

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

$$y_i = A_i\theta + B_i z_i + \epsilon_i \quad . \tag{16}$$

$A_i \in \mathbb{R}^{n_i \times p}$ and $B_i \in \mathbb{R}^{n_i \times m}$ are design matrices, $\theta \in \mathbb{R}^p$ is a vector of parameters, $z_i \in \mathbb{R}^m$ 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 $\mathcal{N}(0, \Omega)$ where $\Omega \in \mathbb{R}^{m \times m}$. We also assume that the residual errors $\epsilon_i \in \mathbb{R}^{n_i}$ are distributed according to $\mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{n_i \times n_i}$, 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:

$$y_i|z_i \sim \mathcal{N}(A_i\theta + B_i z_i, \Sigma), \tag{17}$$
$$z_i|y_i \sim \mathcal{N}(\mu_i, \Gamma_i). \tag{18}$$

where:

$$\Gamma_i = (B_i^\top \Sigma^{-1} B_i + \Omega^{-1})^{-1}, \tag{19}$$
$$\mu_i = \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta). \tag{20}$$

This model belongs to the curved exponential family introduced in section 1.1 where for all $i \in [\![1, N]\!]$:

$$\tilde{S}_i(z_i) \triangleq z_i \quad \text{and} \quad \bar{s}_i(\theta) = \Gamma_i B_i^\top \Sigma^{-1}(y_i - A_i\theta) \tag{21}$$
$$\psi_i(\theta) \triangleq (y_i - A_i\theta)^\top \Sigma^{-1}(y_i - A_i\theta) \tag{22}$$
$$\phi_i(\theta) \triangleq B^\top \Sigma^{-1}(y_i - A_i\theta) \tag{23}$$

Maximising $L(s, \theta)$, defined in (6), with respect to θ yields the following maximisation function for all $s = (s_i \in \mathbb{R}^m, i \in [\![1, N]\!])$:

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

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:

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

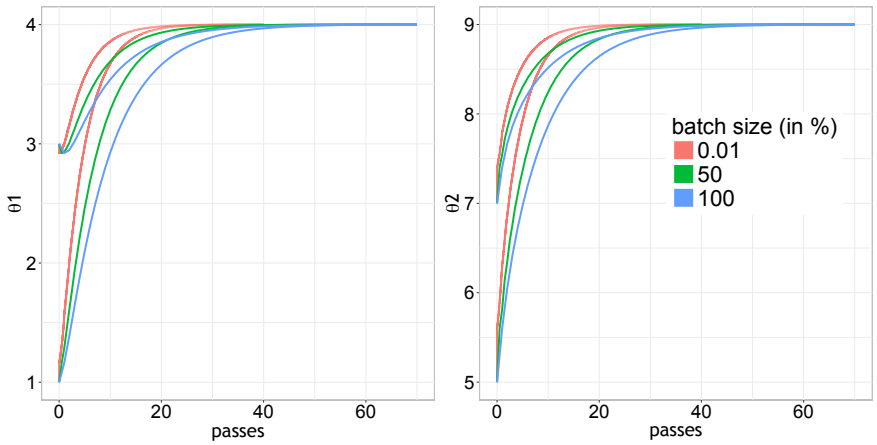

Figure 1: Convergence of the vector of parameter estimates θ^k function of passes over the data.

2.1.2 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.

3 Proofs

3.1 Proof of Theorem 1

3.1.1 Proof of (i)

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

$$\bar{A}^k(\theta) \triangleq \sum_{i=1}^{N} A_i^k(\theta) \ , \tag{24}$$

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

$$\bar{A}^k(\theta) = \bar{A}^{k-1}(\theta) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta) - \sum_{i \in I_k} A_i^{k-1}(\theta). \tag{25}$$

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

$$\bar{A}^k(\theta^k) \leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} B_{i,\theta^{k-1}}(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}). \tag{26}$$

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 (26) we obtain:

$$\begin{aligned} \bar{A}^k(\theta^k) &\leq \bar{A}^{k-1}(\theta^{k-1}) + \sum_{i \in I_k} \ell_i(\theta^{k-1}) - \sum_{i \in I_k} A_i^{k-1}(\theta^{k-1}) & (27) \\ &\leq \bar{A}^{k-1}(\theta^{k-1}) \ . & (28) \end{aligned}$$

As a result, the sequence $\left(\bar{A}^k(\theta^k)\right)_{k \geq 0}$ is monotonically decreasing. Since, under assumption **??**, 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 $J \subset [\![1, N]\!]$:

$$\ell_J(\theta) \triangleq \sum_{i \in J} \ell_i(\theta) \ , \tag{29}$$

$$A_J^{k-1}(\theta) \triangleq \sum_{i \in J} A_i^{k-1}(\theta) \ . \tag{30}$$

Inequality (26) gives :

$$0 \leq \sum_{k=1}^{n} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \leq \sum_{k=1}^{n} \bar{A}^{k-1}(\theta^{k-1}) - \bar{A}^k(\theta^k) = \bar{A}^0(\theta^0) - \bar{A}^n(\theta^n) \ . \tag{31}$$

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:

$$\mathsf{M} \triangleq \mathbb{E}\left[\sum_{k=0}^{\infty} A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1})\right] \tag{32}$$

$$= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}\left[A_{I_k}^{k-1}(\theta^{k-1}) - \ell_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right]\right], \tag{33}$$

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:

$$\mathsf{M} = \frac{p}{N}\sum_{k=0}^{\infty} \mathbb{E}\left[\bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] = \frac{p}{N}\,\mathbb{E}\left[\sum_{k=0}^{\infty} \bar{A}^{k-1}(\theta^{k-1}) - \ell(\theta^{k-1})\right] < \infty . \tag{34}$$

This last equation shows that:

$$\lim_{k\to\infty} \bar{A}^k(\theta^k) - \ell(\theta^k) = 0 \quad \text{a.s.} \tag{35}$$

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

3.1.2 Proof of (ii)

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

$$\bar{h}^k : \vartheta \to \sum_{i=1}^{N} A_i^k(\vartheta) - \ell_i(\vartheta) . \tag{36}$$

$\bar{h}^k$ 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 (**?**)):

$$0 \leq \bar{h}^k(\vartheta^k) \leq \bar{h}^k(\theta^k) - \frac{1}{2L}\|\nabla\bar{h}^k(\theta^k)\|_2^2 \tag{37}$$

$$\implies \|\nabla\bar{h}^k(\theta^k)\|_2^2 \leq 2L\bar{h}^k(\theta^k) . \tag{38}$$

Using (35), 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:

$$\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle = \langle \nabla \bar{A}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \; . \tag{39}$$

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:

$$\begin{aligned} \langle \nabla \ell(\theta^k), \theta - \theta^k \rangle &\geq -\langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle & (40) \\ &\geq -\|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \; . & (41) \end{aligned}$$

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

$$\liminf_{k\to\infty} \inf_{\theta\in\Theta} \frac{\langle \nabla \ell(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq - \lim_{k\to\infty} \|\nabla \bar{h}^k(\theta^k)\|_2 = 0 \; , \tag{42}$$

which is the Asymptotic Stationary Point Condition (ASPC).