Convergence of the Monte-Carlo Expectation Maximization for curved exponential family

Gersende Fort *LMC-IMAG, Grenoble, France*

Eric Moulines *TSI-ENST, Paris, France*

Workshop: MCMC methodology

December 2001, Lancaster

http://www-lmc.imag.fr/lmc-sms/Gersende.Fort/biblio.html

(A-1) Problem

How to maximize on a domain Θ,

$$g(\theta) := \int_{\mathcal{X}} h(z;\theta)\ \mu(dz) \qquad\qquad h(\cdot;\theta) > 0 \quad \mu \text{ a.s}$$

when (continuous or discrete) integration cannot be explicitly solved ?

- A tool for inference in missing data models (dependance upon the observations is omitted).

$g(\theta)$**:** incomplete data likelihood, with parameter θ

$h(z;\theta)$**:** complete data likelihood, z is the vector of the *missing data*.

- Curved exponential family:

$$\log h(z;\theta) =: \phi(\theta) + \langle S(z); \psi(\theta) \rangle$$

(A-2) Expectation Maximization (EM) and Monte-Carlo EM (MCEM)

[Dempster, Laird, Rubin 1977] *[Wei and Tanner 1991, Tanner 1996]*

For curved exponential family: $\log h(z;\theta) =: \phi(\theta) + \langle S(z); \psi(\theta)\rangle$

Define the posterior distribution of the missing data: $\pi_\theta(dz) := h(z;\theta)/g(\theta)\ \ \mu(dz)$

EM: iterative and deterministic algorithm

E-step:

$$\bar{S}(\theta_n) := \int_{\mathcal{X}} S(z)\, \pi_{\theta_n}(dz)$$

M-step:

$$\theta_{n+1} := \mathrm{argmax}_{\theta\in\Theta} \left\{\phi(\theta) + \left\langle \bar{S}(\theta_n); \psi(\theta)\right\rangle\right\}$$

▶ $\theta_{n+1} = \hat{\theta}\left(\bar{S}(\theta_n)\right)$

MCEM: iterative and random algorithm

S-step: m_n samples $\{Z_j^n\}$ of a MC
with invariant distribution $\pi_{\theta_n}(dz)$
with initial distribution λ constant over iteration

E-step:

$$\tilde{S}_n := m_n^{-1} \sum_{j=1}^{m_n} S(Z_j^n)$$

M-step:

$$\theta_{n+1} := \mathrm{argmax}_{\theta\in\Theta} \left\{ \phi(\theta) + \left\langle \tilde{S}_n; \psi(\theta)\right\rangle\right\}$$

▶ $\theta_{n+1} = \hat{\theta}\left(\tilde{S}_n\right)$

(A-3-a) Convergence Analysis: EM algorithm

$$\theta_{n+1} = T(\theta_n)$$

[Wu 1983, Celeux and Diebolt 1992, Lange 1995]

$$g(\theta_{n+1}) \geq g(\theta_n)$$

Under suitable "*regularity conditions*",

- Convergence of $\{g(\theta_n)\}$: There exists θ^*, $\quad \{g(\theta_n)\} \longrightarrow \{\theta \in \Theta, g(\theta) = g(\theta^*)\}$.
- Convergence of $\{\theta_n\}$: If $\{\theta_n\} \longrightarrow \theta^*$, $\quad \theta^* \in \{\theta \in \Theta, T(\theta) = \theta\}$
- Convergence to a stationnary point of g: $\quad \{\theta \in \Theta, T(\theta) = \theta\} = \{\theta \in \Theta, \nabla g(\theta) = 0\}$.
- Convergence to a local maximum of g:

  $|\nabla T(\theta^*)| < 1$ iff θ^* is a local maximum of g.

  $|\nabla T(\theta^*)| > 1$ iff θ^* is a saddle point of g.

  $|\nabla T(\theta^*)^{-1}| < 1$ iff θ^* is a local minimum of g.

$\hookrightarrow$ convergence to a local maximum of g.

(A-3-b) Convergence Analysis: MCEM algorithm

▶ *[Chan and Ledolter 1995]*

Constant simulation schedule $m_n := m$,

Initial value θ_0 in a neighborhood of a local max θ^* of g,

For all $\mathcal{K}$ compact, $\sup_{\theta \in \mathcal{K}} \left| m^{-1} \sum_{j=1}^{m} S(\Phi_j) - \pi_\theta(S) \right| \longrightarrow_{\mathbb{P}} 0$ as $m \to \infty$.

Then, $\forall \epsilon > 0,$ $\exists\, \kappa_\epsilon$ $\lim_m P\left(|\theta_k - \theta^*| < \epsilon, \text{for some } k < \kappa_\epsilon \right) = 1.$

▶ Litterature on iterated random maps *[Shapiro, Wardi 1996, Pierre-Loti-Viaud 1995]*

EM: $\theta_{n+1} = T(\theta_n)$ MCEM: $\theta_{n+1} = F_n(\theta_n)$

For all $\mathcal{K}$ compact, $\sup_{\theta \in \mathcal{K}} \left| g\left(T(\theta)\right) - g\left(F_n(\theta)\right) \right| \longrightarrow_{a.s.} 0$, as $n \to \infty$.

$\hookrightarrow$ Uniform law of large numbers.

(A-4) Contribution

▶ A stable MCEM algorithm: a.s. boundedness of the random sequence $\{\theta'_n\}$.

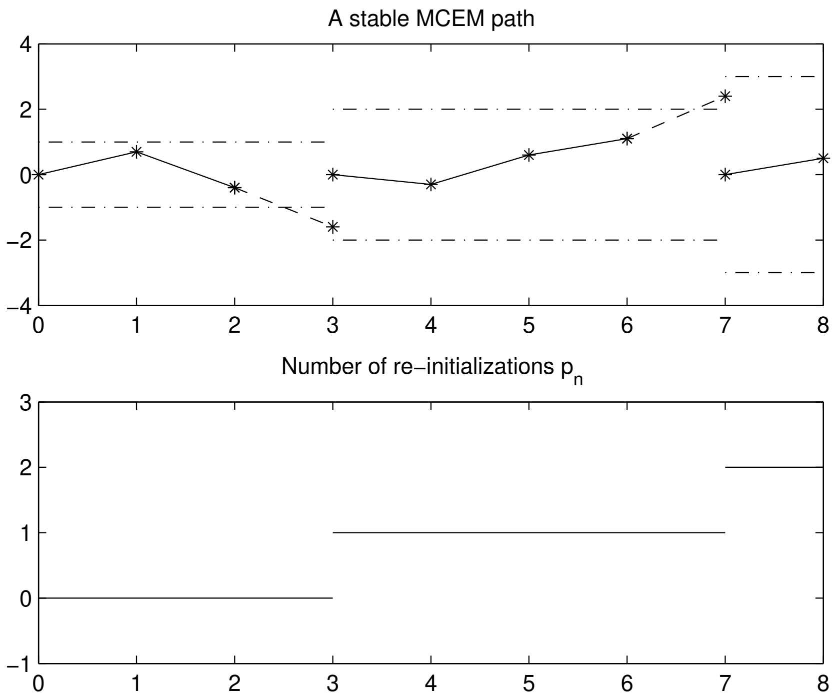
▶ Sufficient conditions for convergence on

the model:	Similar to what is assumed for EM.
the Monte-Carlo appproximation:	Control the L^p norm of the fluctuation $\tilde{S}_n - \bar{S}(\theta'_n)$.
the simulation schedule $\{m_n\}$.	

▶ Rate of convergence

of stable MCEM:	sub-optimal compared to other Stoch. Approx. $\cdots$
of a new method based on averaging:	optimal !!

(B-1) Stable MCEM $\{\theta'_n\}$ Adapted from *[Chen, Guo, Gao 1998]*



For given compact sets $\{\mathcal{K}_n\}$ such that

$$\mathcal{K}_n \subset \mathcal{K}_{n+1} \qquad \Theta = \bigcup_n \mathcal{K}_n,$$

Choose $\theta'_0 \in \mathcal{K}_0$,

Set $p_0 := 0$.

If $\hat{\theta}(\tilde{S}_n) \in \mathcal{K}_{p_n}$,

$$\theta'_{n+1} := \hat{\theta}(\tilde{S}_n),$$
$$p_{n+1} := p_n.$$

Otherwise,

$$\theta'_{n+1} := \theta'_0,$$
$$p_{n+1} := p_n + 1.$$

(B-2-a) Assumptions: Model

Remind: $\mathrm{argmax}_\Theta\, g(\theta) = \int_{\mathcal{X}} h(z;\theta)\, \mu(dz)$ $\log h(z;\theta) = \phi(\theta) + \langle S(z); \psi(\theta)\rangle$

(A1) $\Theta \subset \mathbb{R}^l$ convex, $\mathcal{X} \subset \mathbb{R}^d$ open, μ σ-finite measure.

- ϕ, ψ, S continuous.
- g is positive, continuous and $\{\theta, g(\theta) \geq M\}$ is compact, ($\forall\, M > 0$)
- $\pi_\theta(dz) := h(z;\theta)/g(\theta)\, \mu(dz)$ and $\bar{S}(\theta) := \pi_\theta(S)$ continuous.
- $s \mapsto \hat{\theta}(s) := \mathrm{argmax}_\theta \phi(\theta) + \langle s; \psi(\theta)\rangle$ exists and is continuous.

(B-2-b) Assumptions: Monte-Carlo approximations

Remind: At iteration $n+1$

Draw m_n samples of a MC with stationary distribution $\pi_{\theta'_n}(dz)$, and initial distribution λ.

Approximate $\pi_{\theta'_n}(S)$ by the sample mean.

$\hookrightarrow$ A family of kernel $\{P_\theta\}$, $\theta \in \Theta$, with invariant distribution π_θ and a canonical space $\mathbb{E}_{\lambda,\theta}$

(A2) Let $p \geq 2$ and λ probability on $\mathcal{X}$. For any compact set $\mathcal{K} \subset \Theta$, there exists C

$$\sup_{\theta \in \mathcal{K}} \mathbb{E}_{\lambda,\theta} \left| \sum_{k=1}^{n} \{S(\Phi_k) - \pi_\theta(S)\} \right|^p \leq C\, n^{p/2}$$

$\hookrightarrow$ Uniform in θ Rosenthal's inequality for MC $\{\Phi_n\}$

(A3) $\sum_n m_n^{-p/2} < \infty$.

(B-3-a) The uniform Rosenthal's inequality

Assume

(i) P is ψ-irreducible, aperiodic,

(ii) The state space $\mathcal{X}$ is petite: $P^m(x,\cdot) \geq \epsilon\, \nu_m(\cdot)$.

Then for any $p \geq 2$, any bounded Borel function g,

$$\mathbb{E}_x \left| \sum_{k=1}^{n} \{g(\Phi_k) - \pi(g)\} \right|^p \leq 6^p C_p \left(\sup_{\mathcal{X}} |g|^p \right) \left\{ 1 + 2\{1 - (1-\epsilon)^{1/m}\}^{-1} \right\} n^{p/2},$$

where C_p is the Rosenthal's constant.

Ex: Random Walk HM on a compact state space $\mathcal{X}$:

continuous proposal

$\pi_\theta(dz) =: p(z;\theta)\mu(dz)$ and $(z,\theta) \mapsto p(z;\theta)$ continous.

(B-3-b) The uniform Rosenthal's inequality (continued)

Assume

(i) P is ψ-irreducible, aperiodic,

(ii) C is petite: $\quad P^m(x,\cdot) \geq \epsilon\, \nu_m(\cdot)\, \mathbb{1}_C(x)$

(iii) For some unbounded off petite set function V, $\quad PV \leq \lambda V + b\mathbb{1}_C \qquad \lambda < 1, b < \infty$

Then For any $p \geq 2$ and function $|g| \leq V^{1/p}$,

$$\mathbb{E}_x \left| \sum_{k=1}^{n} \{g(\Phi_k) - \pi_\theta(g)\} \right|^p \leq C(x)\, n^{p/2}$$

where

$$C(x) := \frac{12^p C_p M}{1-\lambda} \{M^{1/p}(1-\lambda^{1/p}) - b^{1/p}\}^{-p} \cdots$$
$$\times \left\{ \frac{b}{1-\lambda} + 5\epsilon^{-1}(2M + bm) + \frac{2M}{M(1-\lambda) - b}\{V(x) + \frac{b}{1-\lambda}\} \right\}$$

and C_p is the Rosenthal's constant and $M := 1 + \sup_C V + b(1-\lambda^{1/p})^{-p}$

(B-4-a) Stability and Convergence

Given compact sets $\{\mathcal{K}_n\}$, $\quad \mathcal{K}_n \subset \mathcal{K}_{n+1} \quad \Theta = \bigcup_n \mathcal{K}_n$,

Given $\theta_0' \in \mathcal{K}_0$,

Given a distribution λ on $\mathcal{X}$,

- $\limsup_n p_n < \infty$ a.s. and $\quad \limsup_n |\theta_n'| < \infty$ a.s. [a.s. boundedness of the paths].
- Convergence of $\{g(\theta_n')\}$: a.s. to a connected component of $g\left(\{\theta, T(\theta) = \theta\}\right)$.
- Convergence of $\{\theta_n'\}$: If $g\left(\{\theta, T(\theta) = \theta\}\right)$ has an empty interior [g is l-times differentiable],

  (a) $\{g(\theta_n')\} \longrightarrow_{a.s.} g^*$

  (b) $\{\theta_n'\} \longrightarrow_{a.s.} \{\theta, T(\theta) = \theta\} \cap \{\theta, g(\theta) = g^*\}$.
- Convergence to (a) stationary point(s) of g: If ϕ and ψ cont. diff and Θ open,

$$\{\theta, T(\theta) = \theta\} = \{\theta, \nabla g(\theta) = 0\} \text{ and } \{\theta_n'\} \longrightarrow_{a.s.} \{\theta, \nabla g(\theta) = 0\}.$$

$\hookrightarrow$ Convergence to a local maximum of g?

(B-4-b) Convergence to a local maximum

Assume

- Θ open
- Regularity conditions (cont. diff or twice cont.diff) on ϕ, ψ, $\bar{S}$, $\hat{\theta}$, g.
- The stationary points of g are isolated [g is a power series]
- For every stationary point θ^*, the matrices

$$-\nabla_\theta^2 L(\bar{S}(\theta^*);\theta^*) \qquad \int_{\mathcal{X}} \nabla_\theta L(S(z);\theta^*) \ {}^t\nabla_\theta L(S(z);\theta^*) \, \pi_{\theta^*}(dz)$$

are positive definite, $\quad L(s;\theta) := \phi(\theta) + < s ; \psi(\theta) >$.

▶ From now on, $\quad \{\theta'_n\} \longrightarrow_n \theta^* \quad$ and θ^* isolated $\quad$ and $\theta^* = T(\theta^*)$.

(B-4-c) Convergence to a local maximum of g (continued)

$\{\theta'_n\} \longrightarrow_n \theta^*$ and $\theta^* = T(\theta^*)$

Remind, for large n: $\theta'_n \hookrightarrow \tilde{S}_n \sim \bar{S}(\theta'_n) \hookrightarrow \theta'_{n+1} = \hat{\theta}(\tilde{S}_n) \hookrightarrow \tilde{S}_{n+1} \cdots$

$\theta'_{n+1} = \hat{\theta}(\tilde{S}_n)$ so $\tilde{S}_n \longrightarrow_n \bar{S}(\theta^*) \Longrightarrow \theta'_n \longrightarrow_n \theta^*$

$$\tilde{S}_n - \bar{S}(\theta^*) = \nabla T(\theta^*) \left(\tilde{S}_{n-1} - \bar{S}(\theta^*)\right) + \eta_n$$

► **No traps** for "strongly perturbed" stochastic algorithms *[Brandière, Duflo 1998]*

If η_n "sufficiently strong", and

$|\nabla T(\theta^*)| < 1$ then $\tilde{S}_n \longrightarrow_n \bar{S}(\theta^*)$ a.s.

$|\nabla T(\theta^*)| > 1$ or $|\nabla T(\theta^*)^{-1}| < 1$, then $\{\tilde{S}_n \longrightarrow_n \bar{S}(\theta^*)\}$ is negligible.

Remind: $|\nabla T(\theta^*)| < 1$ iff θ^* local maximum of g.

(C-1) Rate of convergence of the stable MCEM algorithm

Assume $|\nabla T(\theta^*)| =: \gamma < 1$ (θ^* is a stable fixed point)

$$\tilde{S}_n - \bar{S}(\theta^*) =: \mu_n + \rho_n,$$

$$\mu_{n+1} = \nabla T(\theta^*)\, \mu_n + \epsilon_n$$ ϵ_n, martingale with L^p-norm bounded by $1/\sqrt{m_n}$

▶ If $1 \leq \lim_n m_{n+1}/m_n < \gamma^{-2}$,

$$\mu_n = O_{L^p}\left(1/\sqrt{m_n}\right) \qquad \rho_n \mathbb{1}_{\lim_n \tilde{S}_n = \bar{S}(\theta^*)} = o_{w.p.1}\left(1/\sqrt{m_n}\right).$$

▶ That is, on the simulation time-scale *i.e.* after n simulations

Geometrical schedule: $m_n \sim m^n$ $\qquad \mu_n = O_{L^p}\left(n^{-1/2}\right)$

Polynomial schedule: $m_n \sim n^\alpha$ $\qquad \mu_n = O_{L^p}\left(n^{-\alpha/2(1+\alpha)}\right)$

↪ Depends on the schedule

↪ Rate of convergence of the SAEM algorithm is $n^{-1/2} \cdots$ *[Delyon, Lavielle, Moulines 1998]*

(C-2) An averaging procedure Adapted from *[Polyak 1990, Polyak and Juditsky 1992]*

$$\tilde{S}_n = \bar{S}(\theta^*) + \Xi_n \qquad Var(\Xi_n) \sim m_n^{-1}$$

▶ Compute (in parallel) $\Sigma_n := M_n^{-1}\left(\sum_{k=1}^n m_k \tilde{S}_k\right)$ $\quad M_n := \sum_{k=1}^n m_k$

and choose as estimate: $\bar{\theta}_n := \hat{\theta}(\Sigma_n)$ instead of $\theta'_n := \hat{\theta}(\tilde{S}_n)$

$$\Sigma_n - \bar{S}(\theta^*) =: \bar{\mu}_n + \bar{\rho}_n$$

▶ If $1 \leq \lim_n m_{n+1}/m_n < \gamma^{-2}$,

$$\bar{\mu}_n = O_{L^p}\left(M_n^{-1/2}\right) \qquad \bar{\rho}_n \mathbb{1}_{\lim_n \tilde{S}_n = \bar{S}(\theta^*)} = o_{w.p.1}\left(nM_n^{-1/2}\right).$$

▶ That is, on the simulation time-scale

Geometrical schedule: $m_n \sim m^n$ $\qquad \bar{\mu}_n = O_{L^p}\left(n^{-1/2}\right)$

Polynomial schedule: $m_n \sim n^\alpha, \alpha > 1$ $\qquad \bar{\mu}_n = O_{L^p}\left(n^{-1/2}\right)$

-END-

(X) Some plots

Observations $Y_k \sim \mathcal{P}\left(\exp(\theta_t + Z_k)\right) \qquad Z_k = 0.4 Z_{k-1} + \mathcal{N}(0,1) \qquad \theta_t = 2$

Simulation Step RW HM Hybrid sampler: Gaussian proposal s.t. $\alpha \sim 39\%$.