The Annals of Statistics
2003, Vol. 31, No. 4, 1220–1259

CONVERGENCE OF THE MONTE CARLO EXPECTATION MAXIMIZATION FOR CURVED EXPONENTIAL FAMILIES

BY GERSENDE FORT AND ERIC MOULINES

LMC-IMAG and ENST

The Monte Carlo expectation maximization (MCEM) algorithm is a versatile tool for inference in incomplete data models, especially when used in combination with Markov chain Monte Carlo simulation methods. In this contribution, the almost-sure convergence of the MCEM algorithm is established. It is shown, using uniform versions of ergodic theorems for Markov chains, that MCEM converges under weak conditions on the simulation kernel. Practical illustrations are presented, using a hybrid random walk Metropolis Hastings sampler and an independence sampler. The rate of convergence is studied, showing the impact of the simulation schedule on the fluctuation of the parameter estimate at the convergence. A novel averaging procedure is then proposed to reduce the simulation variance and increase the rate of convergence.

Introduction. Many problems in computational statistics reduce to the maximization of a criterion

$$g(\theta) := \int_{\mathcal{X}} h(z;\theta)\mu(dz), \qquad h(\cdot;\theta) > 0,\ \mu\text{-a.s.,} \tag{1}$$

on a feasible set Θ, when g cannot be computed in closed form. In the terminology of the missing data problem, g is the incomplete data likelihood, that is, the likelihood of the observations for the value of the parameter θ, $z \in \mathcal{X}$ is the missing data vector and h is the complete data likelihood with respect to (w.r.t.) the reference measure μ, that is, h is the likelihood of the observations and of the missing data.

The expectation maximization (EM) algorithm [Dempster, Laird and Rubin (1977)] is a popular iterative procedure for maximizing g. The *E step* of the algorithm requires the computation of the expectation of the complete log-likelihood w.r.t. the posterior distribution of the missing data. In many situations, this step is intractable. To solve this problem, many approximations of the EM algorithm, which use simulations as an intermediate step, have been proposed [see, e.g., Tanner (1996); Celeux and Diebolt (1992); Delyon, Lavielle and Moulines (1999)]. Perhaps the most popular algorithm for this purpose is the Monte Carlo EM (MCEM) initially proposed by Wei and Tanner (1990) and later used and

Received September 1999; revised April 2002.
AMS 2000 subject classifications. Primary 65C05, 62-04; secondary 60J10.
Key words and phrases. EM algorithm, Monte Carlo EM algorithm, Metropolis Hastings algorithms, averaging procedure.

studied by many authors [see Sherman, Ho and Dalal (1999) and references therein].

The basic principle behind this algorithm is to replace the expectation step by a blending of Monte Carlo integration procedure with Markov Chain Monte Carlo (MCMC) sampling techniques such as the Gibbs or the Metropolis Hastings algorithm. The MCEM algorithm has been successfully applied in many different settings, including nonlinear time-series models [Chan and Ledolter (1995)], generalized linear mixed models with missing data [Chan and Kuk (1997)], full-information item factor models [Meng and Schilling (1996)], genetic models [Guo and Thompson (1991)] and blind deconvolution [Cappé, Doucet, Lavielle and Moulines (1999)].

Analysis of the convergence of the MCEM algorithm was first formally addressed by Biscarat (1994) as a specific example of a random iterative algorithm. The conditions in Biscarat (1994) were later weakened by Chan and Ledolter (1995). The assumptions in these works are, however, rather restrictive, because they involve a uniform law of large numbers, that is, uniform convergence in probability of the Monte Carlo expectation to the corresponding sample average over θ in a compact subset of the feasible set Θ. This assumption fails to be verified when Monte Carlo integration is carried out along a single run MCMC algorithm in the simulation step. It can, however, be verified under reasonable assumptions when Monte Carlo integration is done using independent chains, as shown by Sherman, Ho and Dalal [(1999), Theorem 2]. [The difficulty when moving from single run to multiple runs was overlooked by Chan and Ledolter (1995).] Convergence of random iterative algorithms was also considered by Shapiro and Wardi (1996), Pierre-Loti-Viaud (1995) and Brandière (1998), also under restrictive assumptions.

Sherman, Ho and Dalal (1999) addressed a different class of results. They focused on the missing data problem, for which $g(\theta)$ is the incomplete data likelihood, depending on the sample size, say N (the dependence on this parameter is implicit in our work; all the results we obtain are conditional to N). They assumed that the Monte Carlo integration was carried out by means of independent chains and that the number of independent chains, the number of iterations for each chain at each step and the number of iterations of the algorithm are functions of N. Under these assumptions, they derived the rate of convergence of the Monte Carlo estimator obtained as $N \to \infty$.

The purpose of this paper is to complement the results above by providing a convergence analysis of the MCEM algorithm which remains valid under assumptions that are verified for a wide class of MCMC simulation techniques, including both single run and multiple run chains. The proof of convergence is rather different from the schemes used before, avoiding any form of uniform law of large numbers. An averaging technique to improve the rate of convergence is also presented that is based on a modification of the averaging techniques [Polyak (1990)].

This paper is organized as follows. In Section 1, we present the MCEM algorithm and define the *stable MCEM* algorithm which guarantees the almost-sure (a.s.) boundedness of the random recursion. In Section 2, we study the convergence of stable MCEM for curved exponential families when the simulation step is based on MCMC techniques by assuming a *uniform* ergodic behavior of the MCMC kernels. In Section 3, the rate of convergence is derived and how this rate can be improved, with a very small computational overhead, by using an *averaging* approach, is shown. Section 4 is devoted to an application. The proofs are postponed to Sections 5–7.

1. The Monte Carlo expectation maximization algorithm. In this contribution, we use the terminology of the missing data problem. Let $\Theta \subseteq \mathbb{R}^l$, let $\mathcal{X} \subseteq \mathbb{R}^d$ be endowed with the Borel σ-field, let μ be a σ-finite Borel measure on $\mathcal{X}$ and let $\{h(z;\theta), \theta \in \Theta\}$ be a family of positive μ-integrable functions. Any iteration of EM may be formally decomposed into two steps. At iteration $n+1$, the E step consists of evaluating

$$\mathcal{Q}(\theta, \theta_n) := \int_{\mathcal{X}} \log h(z;\theta) p(z;\theta) \mu(dz),$$

where

$$p(z;\theta) := h(z;\theta)/g(\theta),$$

so that

$$\pi_\theta(dz) := p(z;\theta)\mu(dz)$$

is a probability distribution which may be interpreted as the posterior distribution of the missing data. In the M step, the new value of the parameter θ_{n+1} is set as the maximum over Θ of $\theta \mapsto \mathcal{Q}(\theta, \theta_n)$, $\theta_{n+1} := \arg\max_{\phi \in \Theta} \mathcal{Q}(\phi, \theta_n)$. It is assumed for simplicity that this maximum exists and is unique [see Wu (1983) for details]. The key property of EM is that increasing $\mathcal{Q}(\theta, \theta_n)$ forces an increase of g, the function to maximize. It is known that under regularity assumptions, EM instances $\{\theta_n\}$ converge to the set of the stationary points of g [Wu (1983)]. In some situations, the E step is intractable and to deal with these cases, Wei and Tanner (1990) proposed to replace the expectation by a Monte Carlo integration, leading to the so-called Monte Carlo EM. The MCMC approach consists of sampling an $\mathcal{X}$-valued Markov chain $\{Z_j^n\}_j$ from a Markov kernel P_{θ_n}, with stationary distribution π_{θ_n} and initial distribution λ (assumed to be constant over iterations). In the E step we compute

$$\mathcal{Q}_n(\theta, \theta_n) := m_n^{-1} \sum_{j=1}^{m_n} \log h(Z_j^n;\theta), \qquad m_n \in \mathbb{Z}_+,$$

whereas the M step remains unchanged. A difficulty when dealing with the random sequence $\{\theta_n\}$ is to guarantee the stability (a.s. boundedness). To avoid unnecessary technical conditions, we present a simple modification of the iterative scheme by adapting the algorithm presented by Chen, Guo and Gao (1988).

The stable MCEM algorithm. A new sequence $\{\theta'_n\}$ is obtained by truncating the original recursion: whenever $\arg\max_{\phi\in\Theta} \mathcal{Q}_n(\phi, \theta'_n)$ is outside a specific set, it is reinitialized at a point θ'_0. In the technique proposed by Chen, Guo and Gao (1988), the truncation bounds are random functions of the recursion index n. The advantage of this approach (compared to projection) is that the truncation *does not* modify the set of stationary points of the original recursion. More formally, let $\{\mathcal{K}_n\}$ be a sequence of compact subsets such that, for any $n \geq 0$,

$$\mathcal{K}_n \subsetneq \mathcal{K}_{n+1}, \qquad \Theta = \bigcup_{n\geq 0} \mathcal{K}_n. \tag{2}$$

Set $p_0 := 0$ and choose $\theta'_0 \in \mathcal{K}_0$. The stable MCEM algorithm is defined as follows:

$$\begin{aligned} &\text{if } \arg\max_{\phi\in\Theta} \mathcal{Q}_n(\phi, \theta'_n) \in \mathcal{K}_{p_n}, && \theta'_{n+1} := \arg\max_{\phi\in\Theta} \mathcal{Q}_n(\phi, \theta'_n) \text{ and } p_{n+1} := p_n, \\ &\text{if } \arg\max_{\phi\in\Theta} \mathcal{Q}_n(\phi, \theta'_n) \notin \mathcal{K}_{p_n}, && \theta'_{n+1} := \theta'_0 \text{ and } p_{n+1} := p_n + 1. \end{aligned} \tag{3}$$

Note that p_n counts the number of reinitializations. It is shown in the sequel that, under appropriate assumptions, $\{p_n\}$ is a.s. finite, meaning that along any trajectory of the algorithm, the number of reinitialization is finite.

2. Convergence of the MCEM algorithm for a curved exponential family.

2.1. *Model assumptions.* We further restrict our attention to the case where the complete data likelihood h is from the class of *curved exponential densities*. We consider the following assumptions which are satisfied in many scenarios.

M1. $\Theta \subseteq \mathbb{R}^l$, $\mathcal{X} \subseteq \mathbb{R}^d$ and μ is a σ-finite positive Borel measure on $\mathcal{X}$.

Denote by $\langle \cdot ; \cdot \rangle$ the scalar product, denote by $|\cdot|$ the Euclidean norm and denote by ∇ the differentiation operator. Let $\phi : \Theta \to \mathbb{R}$, $\psi : \Theta \to \mathbb{R}^q$ and $S : \mathcal{X} \to \mathcal{S} \subseteq \mathbb{R}^q$. Define $L : \mathcal{S} \times \Theta \to \mathbb{R}$ and $h : \mathcal{X} \times \Theta \to \mathbb{R}^+ \setminus \{0\}$:

$$L(s; \theta) := \phi(\theta) + \langle s; \psi(\theta)\rangle, \qquad h(z; \theta) := \exp\big(L(S(z); \theta)\big).$$

M2. Assume that:

(a) ϕ, ψ are continuous on Θ and S is continuous on $\mathcal{X}$;

(b) for all $\theta \in \Theta$, $\bar{S}(\theta) := \pi_\theta(S)$ is finite and continuous on Θ;

(c) there exists a continuous function $\hat{\theta} : \mathcal{S} \to \Theta$, such that for all $s \in \mathcal{S}$, $L(s; \hat{\theta}(s)) = \sup_{\theta\in\Theta} L(s; \theta)$;

(d) g is positive, finite and continuous on Θ and, for any $M > 0$, the level set $\{\theta \in \Theta,\ g(\theta) \geq M\}$ is compact.

Let $\mathcal{L}$ be the set of stationary points of the EM algorithm. With the notation above, $\mathcal{L}$ is given by

$$\mathcal{L} := \{\theta \in \Theta,\ \hat{\theta} \circ \bar{S}(\theta) = \theta\}. \tag{4}$$

As shown by Wu [(1983), Theorem 2], under M1 and M2, if Θ is open, and ϕ and ψ are differentiable on Θ, then g is differentiable on Θ and $\mathcal{L} = \{\theta \in \Theta, \nabla g(\theta) = 0\}$. Hence, the set of fixed points of EM coincides with the set of stationary points of g.

M3. Assume either that:

(a) the set $g(\mathcal{L})$ is compact or
(a′) for all compact sets $\mathcal{K} \subseteq \Theta$, $g(\mathcal{L} \cap \mathcal{K})$ is finite.

Note that under M2(d), $g(\mathcal{L})$ is compact iff $\mathcal{L}$ is compact.

EXAMPLE (Poisson count with random effect). For the purpose of illustration, we consider the estimation of a location parameter in a model of Poisson counts. This model is adapted from Zeger (1988) [see also Chan and Ledolter (1995)]. Conditional on the latent variables $Z_0, Z_1, \dots, Z_d$, the counts $Y_1, \dots, Y_d$ are independent and Poisson variables with intensity $\exp(\theta + Z_k)$, where θ is the unknown translation parameter to estimate in the maximum likelihood sense. $\{Z_k\}$ is a stationary autoregressive process of order 1, $Z_k = aZ_{k-1} + \sigma\varepsilon_k$, where $\{\varepsilon_k\}$ is i.i.d. standard Gaussian noise and the coefficients $|a| < 1, \sigma > 0$ are known. Set $\mathbf{z} := (z_0, \dots, z_d)$, a $\mathbb{R}^{d+1}$-valued vector. The complete likelihood may be written as

$$h(\mathbf{z}; \theta) = \exp\left(\theta \sum_{k=1}^{d} Y_k - \exp(\theta) \sum_{k=1}^{d} \exp(z_k)\right), \tag{5}$$

the dominating measure μ is absolutely continuous w.r.t. the Lebesgue measure on $\mathcal{X} := \mathbb{R}^{d+1}$ and the density is given up to an irrelevant normalization factor by

$$\exp\left(\sum_{k=1}^{d} Y_k z_k - (2\sigma^2)^{-1}\left(\sum_{k=1}^{d} (z_k - a z_{k-1})^2 + (1-a)^2 z_0^2\right)\right). \tag{6}$$

Here $\Theta := \mathbb{R}$, $\phi(\theta) := \theta \sum_{k=1}^{d} Y_k$, $\psi(\theta) := -e^{\theta}$ and $S(\mathbf{z}) := \sum_{k=1}^{d} e^{z_k} \in \mathcal{S} := \mathbb{R}^+ \setminus \{0\}$. Assumption M2(a) is trivially verified. Observe that for $y > 0$, $z \in \mathbb{R}$, $\theta \in \mathbb{R}$, we have $y\theta - e^{\theta + z} \le -yz + y(\ln(y) - 1)$, so that

$$h(\mathbf{z}; \theta) \le \exp\left(\sum_{k=1}^{d} Y_k\bigl(\log(Y_k) - 1\bigr) - \sum_{k=1}^{d} Y_k z_k\right) \qquad \forall \mathbf{z} \in \mathbb{R}^{d+1},\ \theta \in \mathbb{R}. \tag{7}$$

We easily deduce from (7) that $\sup_{\theta \in \mathbb{R}} g(\theta) < \infty$. Equation (7) also implies that g is uniformly bounded on Θ and is continuous. Since $\lim_{\theta\to-\infty} g(\theta) = \lim_{\theta\to+\infty} g(\theta) = 0$, then the level sets are compact and M2(d) is verified. As g is continuous, M2(b) is trivially checked using similar arguments. M2(c) is verified with

$$\hat{\theta}(s) := \log\left(\sum_{k=1}^{d} Y_k\right) - \log(s).$$

Finally, $\theta \mapsto g(\theta)$ and its derivatives are analytic on Θ and analytic functions have only a finite number of zeros in any compact set. Whereas $\mathcal{L} = \{\theta \in \Theta, \nabla g(\theta) = 0\}$, then for all compact $\mathcal{K} \subset \Theta$, $\mathcal{L} \cap \mathcal{K}$ is finite and M3(a′) is verified.

2.2. *Monte Carlo approximation.* Let $\{\mathcal{K}_n\}$ be a sequence of compact sets satisfying (2). Given $\theta'_0 \in \mathcal{K}_0$ and a probability measure λ on $\mathcal{X}$, the stable MCEM sequence $\{\theta'_n\}$ is then defined as [see (3)]

$$\begin{aligned} &\text{if } \hat{\theta}(\tilde{S}_n) \in \mathcal{K}_{p_n}, \qquad \theta'_{n+1} := \hat{\theta}(\tilde{S}_n) \text{ and } p_{n+1} := p_n, \\ &\text{if } \hat{\theta}(\tilde{S}_n) \notin \mathcal{K}_{p_n}, \qquad \theta'_{n+1} := \theta'_0 \text{ and } p_{n+1} := p_n + 1, \end{aligned} \tag{8}$$

where

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

$\{Z_j^n\}$ is sampled from a Markov kernel $P_{\theta'_n}$ with invariant distribution $\pi_{\theta'_n}$ and $Z_0^n \sim \lambda$. To go further, we need to control the L^p-norm of the fluctuations of the Monte Carlo approximation of $\bar{S}(\theta'_n)$ by $\tilde{S}_n$.

M4. There exist $p \geq 2$ and λ, a probability measure on $\mathcal{X}$, such that for any compact set $\mathcal{K} \subseteq \Theta$,

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

$$\sup_{\theta\in\mathcal{K}} \sup_{n\geq 1} \sum_{k\geq 1} |\lambda P_\theta^k(S) - \pi_\theta(S)| < \infty,$$

where $\mathbb{E}_{\lambda,\theta}$ is the expectation of the canonical Markov chain $\{\Phi_n\}$ with transition kernel P_θ and initial distribution λ.

We now state practical conditions upon which M4 is verified. The simplest case is when the kernel P_θ is uniformly ergodic. [See Meyn and Tweedie (1993) for relevant definitions on Markov chains.] Let P be a Markov kernel on $\mathcal{X}$.

PROPOSITION 1. *Let P be a ψ-irreducible aperiodic Markov transition kernel on $\mathcal{X}$. Assume that the whole state space is ν_m-small with minorizing* *constant* $\varepsilon > 0$. *Then* P *possesses a unique invariant probability measure* π. *In addition, for any* $p \geq 2$ *and any bounded Borel function* $g : \mathcal{X} \to \mathbb{R}^q$,

$$\sum_{k=1}^{\infty} |P^k g(x) - \pi(g)| \leq 2\Big(\sup_{\mathcal{X}} |g|\Big)(1 - (1-\varepsilon)^{1/m})^{-1} \qquad \forall x \in \mathcal{X},$$

and for all $n \geq 1$, $x \in \mathcal{X}$,

$$\mathbb{E}_x \left| \sum_{k=1}^{n} \{g(\Phi_k) - \pi(g)\} \right|^p \leq 6^p C_p \Big(\sup_{\mathcal{X}} |g|^p\Big)(1 + 2\{1 - (1-\varepsilon)^{1/m}\}^{-1})^{p+1} n^{p/2}, \tag{9}$$

where C_p *is Rosenthal's constant* [*see Hall and Heyde* (1980), *Theorem* 2.12].

The proof is given in Section 6.

Using this result, assumption M4 is verified provided that $\sup_{\mathcal{X}} |S| < \infty$, P_θ is for all $\theta \in \Theta$ uniformly ergodic, that is, $\mathcal{X}$ is ν_{m_θ}-small with minorizing constant ε_θ, and for all θ in a compact subset of Θ, (a) ε_θ is bounded away from zero and (b) m_θ is bounded. This condition is often verified when $\mathcal{X}$ is compact and the kernel depends continuously on θ (see Section 4 for an illustration). To deal with noncompact state space, the following proposition (proved in Section 6) provides convenient sufficient conditions based on the Foster–Lyapunov drift criterion (10).

PROPOSITION 2. *Let* P *be a* ψ*-irreducible aperiodic transition kernel on* $\mathcal{X}$ *and let* C *be an accessible petite set. Assume that there exist some constants* $0 < \rho < 1$, $b < \infty$ *and a Borel norm-like function* $V : \mathcal{X} \to [1, \infty)$, *bounded on* C *such that*

$$PV \leq \rho V + b\mathbb{1}_C. \tag{10}$$

Let $p \geq 2$. *Choose* $M > \sup_C V \vee b/(1 - \rho^{1/p})^p$. *Then the set* $\{V \leq M\}$ *is* ν_m*-small with minorizing constant* $\varepsilon > 0$ *and for any Borel function* $g : \mathcal{X} \to \mathbb{R}^q$, $|g| \leq V^{1/p}$, *it holds that for all* $x \in \mathcal{X}$, $n \geq 1$,

$$\sum_{k=1}^{\infty} |P^k g(x) - \pi(g)| \leq C\varepsilon^{-1}(m+1)M^{1/p}A^{-1}V^{1/p}(x)$$

and

$$\mathbb{E}_x \left| \sum_{k=1}^{n} \{g(\Phi_k) - \pi(g)\} \right|^p \leq C\varepsilon^{-(p+1)}(m+1)^{p+1}M^2 A^{-2p} V(x) n^{p/2},$$

where $A := ((1-\rho)^{1/p} - (b/M)^{1/p})$ *and* C *is a constant which depends only on* p.

Hence, if the kernel P depends on a parameter θ, all the quantities appearing in Proposition 2 may depend on θ and the condition M4 is verified if, for any compact subset $\mathcal{K} \subset \Theta$, (a) $\sup_{\theta \in \mathcal{K}} \rho_\theta < 1$, $\sup_{\theta \in \mathcal{K}} b_\theta < \infty$, $\sup_{\theta \in \mathcal{K}} M_\theta < \infty$ and $\sup_{\theta \in \mathcal{K}} m_\theta < \infty$, (b) $\inf_{\theta \in \mathcal{K}} \varepsilon_\theta > 0$ and (c) there exists a measure of probability λ on $\mathcal{X}$ such that $\sup_{\theta \in \mathcal{K}} \lambda(V_\theta) < \infty$.

Finally, we need to assume that the number of simulations at each iteration increases at a given rate $\{m_n\}$. The rate of increase depends on the control of the fluctuation of the Monte Carlo sum. More precisely:

M5. $\{m_n\}$ is a sequence of integers such that $\sum_n m_n^{-p/2} < \infty$, where p is given by M4.

EXAMPLE [Poisson count with random effect (continued)]. To impute the missing values, we use the hybrid sampler *random scan symmetric random walk Metropolis Hastings* (henceforth denoted RSM). At each iteration, a single component of the missing data vector $\mathbf{z}$ drawn at random is updated using a one-dimensional random walk Metropolis Hastings algorithm, with a proposal distribution that has a positive, continuous and symmetric density q w.r.t. the Lebesgue measure on $\mathbb{R}$. This sampler was studied by Fort, Moulines, Roberts and Rosenthal (2003). The key findings are summarized here:

- For any $\theta \in \Theta$, the RSM kernel P_θ is Lebesgue-irreducible and aperiodic. In addition, for any compact sets $C \subset \mathbb{R}^{d+1}$ and $\mathcal{K} \subset \Theta$, there exist a constant $\varepsilon > 0$ and a probability measure ν on $\mathbb{R}^{d+1}$ such that $P_\theta^{d+1}(\mathbf{z}, \cdot) \geq \varepsilon \nu(\cdot)$ for all $\theta \in \mathcal{K}, \mathbf{z} \in C$.
- Choose $0 < s < 1$ such that $s(1-s)^{1/s-1} < (2d-2)^{-1}$ and set $V_\theta(\mathbf{z}) := \pi_\theta(\mathbf{z})^{-s}$. Then, for any compact $\mathcal{K} \subset \Theta$,

$$\limsup_{|z| \to +\infty} \sup_{\theta \in \mathcal{K}} \frac{P_\theta V_\theta(\mathbf{z})}{V_\theta(\mathbf{z})} < 1.$$

Consequently, by applying Proposition 2, it is proved that assumption M4 holds with any real $p \geq 2$ and any probability measure λ such that for any compact set $\mathcal{K} \subset \Theta$, $\sup_{\theta \in \mathcal{K}} \lambda(V_\theta) < \infty$.

2.3. *Almost-sure convergence.* We now state the main results of our contribution. Under assumptions M1 and M2, any iteration of the EM algorithm can be written as $\theta_{n+1} = \hat{\theta} \circ \bar{S}(\theta_n) =: T(\theta_n)$, where $T : \Theta \to \Theta$ is continuous. Wu [(1983), Theorem 1] proved that (a) $\{g(\theta_n)\}$ converges to $g(\theta^*)$ for some θ^* in the set $\mathcal{L}$ of the fixed points of T and (b) the limit points of $\{\theta_n\}$ are in $\mathcal{L}$. Under assumptions M1–M4, we obtain a similar result for the stable MCEM algorithm. The convergence results hold almost surely w.r.t. $\bar{\mathbb{P}}$, the probability on the canonical space associated to the trajectories of stable MCEM, started at θ'_0, given λ, the initial distribution of the Markov chains, and $\{\mathcal{K}_n\}$, the sequence of compact sets (see Section 5.2 for a precise definition of $\bar{\mathbb{P}}$). Denote by $\mathrm{Cl}(A)$ the closure of the set A.

THEOREM 3. *Assume* M1–M5. *Let* $\{\mathcal{K}_n\}$ *be a sequence of compact sets satisfying* (2), *let* $\theta'_0 \in \mathcal{K}_0$ *and let* λ *be given as in* M4. *Consider the stable MCEM random sequence* $\{\theta'_n\}$ *defined by* (8). *Then*:

(i) (a) $\lim_n p_n < \infty$ *w.p.*1 *and* $\limsup_n |\theta'_n| < \infty$ *w.p.*1;

(b) $\{g(\theta'_n)\}$ *converges w.p.*1 *to a connected component of* $g(\mathcal{L})$, *where* $\mathcal{L}$ *is given by* (4).

(ii) *If, in addition*, $g(\mathcal{L} \cap \mathrm{Cl}(\{\theta'_n\}))$ *has an empty interior, then* $\{g(\theta'_n)\}$ *converges w.p.*1 *to* g^* *and* $\{\theta'_n\}$ *converges to the set* $\mathcal{L}_{g^*} := \{\theta \in \mathcal{L}, g(\theta) = g^*\}$.

The proof is given in Section 5.

REMARK 4. Using Sard's theorem [Bröcker (1975)], it is known that $g(\{\nabla g = 0\})$ has an empty interior as soon as the function g is l-times differentiable (where l is the dimension of the parameter space). Hence, Theorem 3(ii) applies under very weak regularity assumptions.

In many instances, the set $\mathcal{L}$ is made up of isolated points and, under suitable conditions, the previous convergence results imply pointwise convergence to some stationary point of g. Depending on the values of the Hessian of g, these limiting points are either local maxima, local minima or saddle points. A question of interest is to state conditions upon which the stationary points coincide only with local maxima. To that goal, we formulate some additional regularity assumptions:

M6. (a) Θ is open;
(b) for any $s \in \mathcal{S}$, $\theta \mapsto L(s;\theta)$ is twice continuously differentiable on Θ;
(c) $\theta \mapsto \bar{S}(\theta)$ is twice continuously differentiable on Θ;
(d) $\theta \mapsto g(\theta)$ is continuously differentiable on Θ;
(e) $\mathcal{S}$ is open and the convex hull of $S(\mathbb{R}^d)$ is included in $\mathcal{S}$;
(f) $s \mapsto \hat{\theta}(s)$ is twice continuously differentiable on $\mathcal{S}$.

M7. The stationary points of g are isolated. For every stationary point θ^* of g, the matrices $-\nabla^2_\theta L(\bar{S}(\theta^*);\theta^*)$ and

$$\int_{\mathcal{X}} \nabla_\theta L\big(S(z);\theta^*\big)^t \,\nabla_\theta L\big(S(z);\theta^*\big) p(z;\theta^*)\mu(dz)$$

are positive definite.

It was shown by Delyon, Lavielle and Moulines (1999) that under M6 and M7, the matrix

$$\nabla T(\theta^*) = [\nabla^2_\theta L(s^*;\theta^*)]^{-1}\big(\nabla^2_\theta L(s^*;\theta^*) - \nabla^2 \log g(\theta^*)\big), \qquad s^* := \bar{S}(\theta^*),$$

is diagonalizable with positive real eigenvalues. If θ^* is a stable fixed point of T, then the modulus of all the eigenvalues of $\nabla T(\theta^*)$ is strictly less than 1 and θ^* is a proper maximizer of g. If θ^* is hyperbolic (resp. unstable), then it is a saddle point of g (resp. a local minimum of g). Recall finally that if the stationary points of g are isolated (i.e., under M7), convergence to hyperbolic and unstable points (i.e., convergence to saddle points and local minima of g) never occurs w.p.1 for the MCEM sequence, as shown in Brandière (1998).

EXAMPLE [Poisson count with random effects (continued)]. Assumption M6 is readily verified. Note that

$$\nabla \log g(\theta) = \sum_{k=1}^{d} Y_k - e^{\theta} \int_{\mathbb{R}^{d+1}} S(\mathbf{z}) p(\mathbf{z}; \theta) \mu(d\mathbf{z})$$

and a stationary point θ^* solves the equation

$$\sum_{k=1}^{d} Y_k = e^{\theta^*} \int S(\mathbf{z}) p(\mathbf{z}; \theta^*) \mu(d\mathbf{z}),$$

that is,

$$\sum_{k=1}^{d} Y_k = e^{\theta^*} \bar{S}(\theta^*).$$

Since g is analytic (see Section 2.1), any compact subset of Θ contains only a finite number of stationary points of g. For a stationary point θ^*, note that $-\nabla_\theta^2 L(\bar{S}(\theta^*); \theta^*) = e^{\theta^*} \bar{S}(\theta^*)$ and

$$\int \nabla_\theta L\big(S(\mathbf{z}); \theta^*\big)^t \, \nabla_\theta L\big(S(\mathbf{z}); \theta^*\big) p(\mathbf{z}; \theta^*) \mu(d\mathbf{z})$$
$$= e^{\theta^*} \int \big(S(\mathbf{z}) - \bar{S}(\theta^*)\big)^2 p(\mathbf{z}; \theta^*) \mu(d\mathbf{z}),$$

so that M7 holds.

3. Rate of convergence and averaging. We now study the rate of convergence of $\{\theta'_n\}$ (given $\{\mathcal{K}_n\}$, $\theta'_0 \in \mathcal{K}_0$ and λ) to a local maximum θ^* of g. Rate of convergence is useful to understand how we should ideally tune the number of simulations m_n as a function of the iteration index. It also allows us to derive an accelerated version of the algorithm based on averaging.

Define $G(s) := \bar{S} \circ \hat{\theta}(s)$. The mapping G gives another way to consider an iteration of the EM algorithm, not directly in the parameter space Θ, but in the space of the complete data sufficient statistics $\mathcal{S}$. If θ^* is a fixed point of T, that is, $\theta^* = T(\theta^*) = \hat{\theta} \circ \bar{S}(\theta^*)$, then $s^* := \bar{S}(\theta^*)$ is a fixed point of G, that is, $s^* = G(s^*) = \bar{S} \circ \hat{\theta}(s^*)$. In addition, $\nabla T(\theta^*) = \nabla\hat{\theta}(s^*)\nabla\bar{S}(\theta^*)$ and $\nabla G(s^*) = \nabla\bar{S}(\theta^*)\nabla\hat{\theta}(s^*)$. Hence $\nabla G(s^*)$ has the same eigenvalues as $\nabla T(\theta^*)$, counting multiplicities together with $(q - l)$ additional eigenvalues equal to zero. The stability properties can thus be directly translated in terms of the stability of s^*: when θ^* is stable, then s^* is stable and vice versa.

3.1. *Rate of convergence.* We begin by informally discussing the results. Let θ^* be a fixed point of T and let $s^* := \bar{S}(\theta^*)$. There are a priori multiple possible limiting points, so we need to restrict our attention to the set of trajectories that converges to a given limiting point s^*. For large enough n, we may decompose the recursion as

$$\tilde{S}_n - s^* = \big(G(\tilde{S}_{n-1}) - G(s^*)\big) + \tilde{S}_n - G(\tilde{S}_{n-1}) = \Gamma(\tilde{S}_{n-1} - s^*) + \varepsilon_n + \eta_n,$$

where $\Gamma := \nabla G(s^*)$ and $\{\varepsilon_n\}$ is a martingale difference sequence w.r.t. the filtration $\mathcal{F}_n := \sigma(\tilde{S}_0, \dots, \tilde{S}_n)$,

$$\varepsilon_n := \big(\tilde{S}_n - \overline{\mathbb{E}}[\tilde{S}_n|\mathcal{F}_{n-1}]\big)\mathbb{1}_{\{|\tilde{S}_{n-1}-s^*|\le\delta\}}, \qquad n \ge 1,\ \delta > 0,\ \varepsilon_0 := 0.$$

The remainder term η_n can be expressed as $\eta_n := \eta_n^{(1)} + \eta_n^{(2)}$, where for $n \ge 1$,

$$\eta_n^{(1)} := \big(\tilde{S}_n - G(\tilde{S}_{n-1})\big)\mathbb{1}_{\{|\tilde{S}_{n-1}-s^*|\ge\delta\}} + \big(\overline{\mathbb{E}}[\tilde{S}_n|\mathcal{F}_{n-1}] - G(\tilde{S}_{n-1})\big)\mathbb{1}_{\{|\tilde{S}_{n-1}-s^*|\le\delta\}}, \tag{11}$$

$$\eta_n^{(2)} := \big(G(\tilde{S}_{n-1}) - G(s^*) - \Gamma(\tilde{S}_{n-1} - s^*)\big) = \sum_{i,j} R_{n-1}(i,j)(\tilde{S}_{n-1,i} - s_i^*)(\tilde{S}_{n-1,j} - s_j^*), \tag{12}$$

and R_n is defined componentwise as

$$R_n(i,j) := \int_0^1 (1-t)\frac{\partial^2 G(s^* + t(\tilde{S}_n - s^*))}{\partial s_i\, \partial s_j}\, dt.$$

It is convenient to decompose the error $\tilde{S}_n - s^*$ as a sum of a linear term μ_n that obeys a linear difference equation driven by the martingale difference ε_n,

$$\mu_n = \Gamma\mu_{n-1} + \varepsilon_n, \qquad n \ge 1 \text{ and } \mu_0 := 0, \tag{13}$$

and a remainder term

$$\rho_n := \tilde{S}_n - s^* - \mu_n, \qquad n \ge 0, \tag{14}$$

which will be shown to be negligible along the trajectories converging to s^*. We stress that, because there are possibly several convergence points, the remainder term ρ_n as defined above will be small *only* along trajectories that converge to s^*.

As shown in the previous section, under the stated assumptions, $\tilde{S}_n$ may converge only to stable points of G (hyperbolic points and unstable points are avoided w.p.1), which are associated with a local maximum of the incomplete likelihood g. Hence, we may assume that s^* is stable, which implies that all the eigenvalues of Γ have modulus less than 1 and, thus, that there exist $\gamma < 1$ and a constant $C < \infty$ such that for all k, $|\Gamma^k| \leq C\gamma^k$, where $|\cdot|$ is any matrix norm. This implies that the linear control model (13) is stable and that

$$\mu_n = \sum_{k=0}^{n} \Gamma^k \varepsilon_{n-k}.$$

In many situations, γ is very close to 1, which explains why the EM algorithm is sometimes slow to converge [see Jamshidian and Jennrich (1997)]. Most often, γ is unknown. It can, however, be estimated using, for example, the Louis information principle [see Delyon, Lavielle and Moulines (1999)], but this generally involves a significant computational overhead. By construction, the driving error $\{\varepsilon_n\}$ is a martingale increment. Observe that if we assume that, for all n, $|\tilde{S}_{n-1} - s^*| \leq \delta$ for some deterministic s^* and δ, then there exists a deterministic compact $\mathcal{K} \subseteq \Theta$ such that for all n, $\theta_n' \in \mathcal{K}$. From this remark and M4, it may be asserted that the L^p-norm of the martingale ε_n is inversely proportional to $\sqrt{m_n}$, the square root of the number of simulations at step n. Hence,

$$\mu_n = O_{L^p}\left(\sum_{k=0}^{n} \gamma^{n-k} m_k^{-1/2}\right).$$

We say that $X_n = O_{L^p}(\alpha_n)$, where $\alpha_n \neq 0$ if $\alpha_n^{-1} X_n$ is bounded in L^p. A more explicit expression for the rate of μ_n can be obtained by using the following lemma, from Pólya and Szegö [(1976), Result 178, page 39]:

LEMMA 5. *Let $\{a_n\}$ and $\{b_n\}$, $b_n \neq 0$, be two sequences such that* (i) *the power series $f(x) := \sum_{n=1}^{\infty} a_n x^n$ has a radius of convergence r and* (ii) $\lim_{n\to\infty} b_n/b_{n+1} =: q$ *with $|q| < r$. Define $c_n := \sum_{k=0}^{n} a_k b_{n-k}$. Then* $\lim_{n\to\infty} c_n b_n^{-1} = f(q)$.

Hence, provided that $\lim_n m_{n+1}/m_n < \gamma^{-2}$, the linear term $\mu_n = O_{L^p}(m_n^{-1/2})$. The constraint $\lim_n m_{n+1}/m_n < \gamma^{-2}$ is always satisfied when $\{m_n\}$ is subexponential. When $\limsup \gamma^{2n} m_n = \infty$, the constraint is no longer satisfied and the rate is strictly lower than $m_n^{-1/2}$. Of course, this analysis makes sense only if we can prove that μ_n is the leading term of the error $\tilde{S}_n - s^*$, that is, ρ_n is negligible w.r.t. μ_n along the trajectories of $\tilde{S}_n$ that converge to s^*. More specifically, we have to show that (see Lemma 14, Section 7)

$$\rho_n \mathbb{1}_{\{\lim_n \tilde{S}_n = s^*\}} = o_{\text{w.p.1}}(m_n^{-1/2}). \tag{15}$$

We say that $X_n = o_{\text{w.p.1}}(\alpha_n)$ [resp. $X_n = O_{\text{w.p.1}}(\alpha_n)$], where $\alpha_n \neq 0$ if $\lim_n \alpha_n^{-1}|X_n| = 0$ w.p.1 (resp. $\alpha_n^{-1}|X_n|$ is bounded w.p.1).

The discussion above is summarized in the following theorem.

THEOREM 6. *Assume* M1–M7. *Let s^* be a stable fixed point of the map G. Let $\gamma < 1$ be the modulus of the largest eigenvalue of $\nabla G(s^*)$. Assume that $1 \le \lim_n m_{n+1}/m_n < \gamma^{-2}$. Then $\mu_n = O_{L^p}(m_n^{-1/2})$ and $\rho_n \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = o_{\text{w.p.1}}(m_n^{-1/2})$, where μ_n and ρ_n are given by* (13) *and* (14).

Theorem 6 shows that, under weak conditions on the sequence $\{m_n\}$, along any trajectory converging to a stable fixed point s^*, the error $\theta'_n - \theta^*$ (or equivalently $\tilde{S}_n - s^*$) is asymptotically given by μ_n. In addition, the L^p-norm of μ_n decreases as the square root of the number of simulations at step n.

To compare the rate of convergence of the MCEM algorithm with other stochastic versions of the EM algorithm, such as the stochastic approximation EM (SAEM), it is worthwhile to compute the rate as a function of the number of simulations rather than as a function of the number of iterations. For a generic sequence $\{X_n\}$, define the interpolated sequence $X_n^{(i)} = X_{\phi(n)}$, where ϕ is defined as the largest integer such that

$$\sum_{k=0}^{\phi(n)} m_k < n \le \sum_{k=0}^{\phi(n)+1} m_k.$$

The subscript n for the interpolated sequence $\theta_n'^{(i)}$ refers to the total number of simulations, while for the original sequence $\{\theta'_n\}$, it coincides with the number of iterations. Assume first that the number of simulations is increasing at a polynomial rate, that is, $m_n := n^\alpha$ so that $\phi(n) \sim [(1+\alpha)n]^{1/(1+\alpha)}$. On the simulation time scale, $\mu_n^{(i)} = O_{L^p}(n^{-\alpha/(2(1+\alpha))})$ and $\rho_n^{(i)} = o_{\text{w.p.1}}(n^{-\alpha/(2(1+\alpha))})$. Hence the rate of convergence is always smaller than $n^{-1/2}$, which is the rate of the SAEM algorithm [Delyon, Lavielle and Moulines (1999), Theorem 7]. It is interesting to note that the rate is improved by choosing large values of α, whereas small values of α can lead to rather inefficient estimates. In practice, this means that it is better to increase the number of simulations rapidly when the algorithm is approaching convergence, giving thus a theoretical background to well established practice. Assume now that $m_n := m^n$, $m > 1$. This choice is advocated in Chan and Ledolter (1995) and in several earlier works on the subject. We get similarly that $\mu_n^{(i)} = O_{L^p}(n^{-1/2})$ and $\rho_n^{(i)} = o_{\text{w.p.1}}(n^{-1/2})$ whenever $1 < m < \gamma^{-2}$: in this case, the rate of convergence is $n^{-1/2}$, provided that m is small enough.

3.2. *The averaging procedure.* The preceding discussion evidences that the performance depends critically on the choice of the schedule, which is of course a serious practical drawback. Recently, a data-driven procedure was proposed by Booth and Hobert (1999). This procedure requires evaluation of the variance of $\tilde{S}_n - G(\tilde{S}_{n-1})$ which is a challenging problem when MCMC is used to sample the missing data.

We consider here an alternative procedure adapted from a technique developed by Polyak (1990) to improve the rate of convergence for stochastic approximation procedures. To motivate the construction, recall that

$$\tilde{S}_n = s^* + \Xi_n, \qquad \Xi_n := \sum_{k=0}^{n} \Gamma^{n-k} \varepsilon_k + \rho_n.$$

Each value of $\tilde{S}_n$ may be seen as an estimator of s^* affected by a noise term. The stable MCEM algorithm estimates s^* by $\tilde{S}_n$ which is an inefficient estimation strategy. By analogy with the regression problem, an estimator of s^* with reduced variance can be obtained by averaging and weighting the successive estimates $\tilde{S}_n$ of s^*. Because the regression noise Ξ_n is both correlated and heteroscedastic, the best unbiased linear estimator of s^* would require us to know (or estimate) both the correlation and the variance of Ξ_n, which is a difficult task. For simplicity, we consider the weighted average

$$\Sigma_n := M_n^{-1} \sum_{j=0}^{n} m_j \tilde{S}_j \quad \text{and} \quad M_n := \sum_{j=0}^{n} m_j, \tag{16}$$

where $\tilde{S}_n$ is weighted by m_n, which is a rough estimate of the inverse of the variance of Ξ_n. Σ_n may thus be seen as a weighted least-squares estimate of s^*, where the weights are (roughly) proportional to the inverse of the noise variance.

Using the decomposition above, $\Sigma_n - s^*$ may be written as $\Sigma_n - s^* = \bar{\mu}_n + \bar{\rho}_n$, where

$$\begin{aligned} \bar{\mu}_n &:= M_n^{-1} \sum_{k=0}^{n} \left(\sum_{j=0}^{n-k} m_{j+k} \Gamma^j \right) \varepsilon_k, \\ \bar{\rho}_n &:= M_n^{-1} \sum_{k=0}^{n} m_k \rho_k. \end{aligned} \tag{17}$$

Under M4, $\overline{\mathbb{E}}[|\varepsilon_n|^p | \mathcal{F}_n] \le 2^p C m_n^{-p/2}$, where C, given by M4, does not depend on the simulation schedule. Then the martingale form of Rosenthal's inequality implies that

$$\begin{aligned} \|\bar{\mu}_n\|_{L^p} \le C(p) \Biggl(\Biggl(\sum_{k=0}^{n} m_k^{-1} \Biggl(\sum_{j=0}^{n-k} m_{j+k} \gamma^j \Biggr)^2 \Biggr)^{1/2} \\ + \Biggl(\sum_{k=0}^{n} m_k^{-p/2} \Biggl(\sum_{j=0}^{n-k} m_{j+k} \gamma^j \Biggr)^p \Biggr)^{1/p} \Biggr) M_n^{-1}, \end{aligned}$$

where $C(p)$ is a constant depending only on p. A more explicit expression for the rate of $\bar{\mu}_n$ can be obtained from the following lemma (the proof of which is postponed to Section 7).

LEMMA 7. *Let $0 < \gamma < 1$ and $\{m_n\}$ be a positive sequence such that $1 \le \lim_n m_{n+1}/m_n =: m < \gamma^{-2}$. Define for some positive integer r,*

$$\xi_n^{(r)} := \left(\sum_{k=0}^n m_k^{r/2}\right)^{-1/r} \left(\sum_{k=0}^n m_k^{-r/2}\left(\sum_{j=0}^{n-k} m_{j+k}\gamma^j\right)^r\right)^{1/r}.$$

Then $\lim_n \xi_n^{(r)} =: \mathcal{B}_r(m;\gamma)$, where

$$\mathcal{B}_r(m;\gamma) := \Bigg((1-m\gamma)^{-r} \times \left[1 + (m^{r/2}-1)\sum_{l=0}^{r-1}\binom{r}{l}(-1)^{r-l}(m^{l-r/2}\gamma^{l-r}-1)^{-1}\right]\Bigg)^{1/r} \quad \text{if } m\gamma \neq 1,$$

$$\mathcal{B}_r(\gamma^{-1};\gamma) := \left((1-\gamma^{r/2})\sum_n (n+1)^r\gamma^{nr/2}\right)^{1/r}.$$

Hence, provided that $\lim_n m_{n+1}/m_n =: m < \gamma^{-2}$, this shows that

$$\lim_n M_n^{1/2}\|\bar{\mu}_n\|_{L^p} \le C(p)\mathcal{B}_2(m,\gamma) + C(p)\mathcal{B}_p(m,\gamma)\lim_n\left(\sum_{k=0}^n m_k^{p/2}\right)^{1/p} M_n^{-1/2}. \tag{18}$$

If $m = 1$ [this happens, e.g., for polynomial schedules $m_n \propto n^\alpha$ or subgeometrical schedules $m_n \propto \exp(n^\alpha)$, $\alpha < 1$], then $\sum_{k=0}^n m_k^{p/2} \sim n m_n^{p/2}$ and $\lim_n (\sum_{k=0}^n m_k^{p/2})^{1/p} M_n^{-1/2} = 0$. Hence,

$$\lim_n M_n^{1/2}\|\bar{\mu}_n\|_{L^p} \le C(p)\mathcal{B}_2(1,\gamma).$$

If $1 < m$, then Lemma 5 implies that $\lim_n (\sum_{k=0}^n m_k^{p/2})^{1/p} M_n^{-1/2} = (m-1)^{1/2} \times (m^{p/2}-1)^{-1/p}$. Hence,

$$\lim_n M_n^{1/2}\|\bar{\mu}_n\|_{L^p} \le C(p)\mathcal{B}_2(m,\gamma) + C(p)\mathcal{B}_p(m,\gamma)(m-1)^{1/2}(m^{p/2}-1)^{-1/p}.$$

This discussion shows that the L^p-norm of the term $\bar{\mu}_n$ decreases as $M_n^{-1/2}$, the inverse of the square root of the total number of simulations up to iteration n. In addition, $m \mapsto \mathcal{B}_2(m,\gamma)$ increases on $[1,\gamma^{-2}) \setminus \{\gamma^{-1}\}$ and the minimum is $\mathcal{B}_2(1,\gamma) = (1-\gamma)^{-1}$; when $m = \gamma^{-1}$, $\mathcal{B}_2(\gamma^{-1},\gamma) = (1+\gamma)^{1/2}(1-\gamma)^{-3/2} > \mathcal{B}_2(1,\gamma)$. This implies that the upper bound in (18) is minimal for $m = 1$ and that the upper bound for the error term is minimum when $\lim_n m_{n+1}/m_n = 1$.

The term $\bar{\mu}_n$ is the leading term in $\Sigma_n - s^*$ provided that, along any trajectories that converge to s^*, $\bar{\rho}_n$ is negligible w.r.t. $\bar{\mu}_n$, that is, $\bar{\rho}_n \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = o_{\text{w.p.1}}(M_n^{-1/2})$. By (37) and (38), $\bar{\rho}_n \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = O_{\text{w.p.1}}(1) O_{L^p}(n M_n^{-1})$. Hence, $\bar{\rho}_n$ is negligible compared to $\bar{\mu}_n$ whenever the simulation schedule checks the condition $n M_n^{-1/2} = o(1)$. For example, for geometrical schedules, this condition is always checked, whereas for polynomial schedules $m_n \propto n^\alpha$, one has to choose $\alpha > 1$.

The discussion above is summarized in the following theorem.

THEOREM 8. *Assume* M1–M7. *Let s^* be a stable fixed point of the map G and denote $\Gamma := \nabla G(s^*)$. Let $\gamma < 1$ be the modulus of the largest eigenvalue of $\nabla G(s^*)$. Let M_n, $\bar{\mu}_n$ and $\bar{\rho}_n$ be given by* (16) *and* (17). *Assume that* (i) $1 \le \lim_n m_{n+1}/m_n < \gamma^{-2}$ *and* (ii) $n M_n^{-1/2} = o(1)$. *Then $\bar{\mu}_n = O_{L^p}(M_n^{-1/2})$ and $\bar{\rho}_n \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = o_{\text{w.p.1}}(M_n^{-1/2})$.*

Theorem 8 shows that under weak conditions on the sequence $\{m_n\}$, along any trajectory converging to a stable fixed point s^*, the error $\Sigma_n - s^*$ behaves asymptotically as $\bar{\mu}_n$; thus, the estimator $\bar{\theta}_n := \hat{\theta}(\Sigma_n)$ (or equivalently Σ_n) has a rate proportional to $M_n^{-1/2}$, that is, a rate inversely proportional to the square root of the total number of simulations up to iteration n. Expressed on the simulation time scale, the previous result shows that the L^p-norm of the leading term $\bar{\mu}_n^{(i)}$ is proportional to $n^{-1/2}$.

Hence, the averaging procedure improves the rate of convergence. In addition, the discussion above evidences that when averaging is used, use of geometrical schedules is not recommended. It is better to choose m_n in such a way that $\lim_n m_{n+1}/m_n = 1$ and $n M_n^{-1/2} = o(1)$, which is verified, for example, if m_n grows polynomially.

EXAMPLE [Poisson count with random effects (continued)]. A plot of $N = 100$ observations $Y_1, \dots, Y_{100}$, obtained with $\theta_{\text{true}} = 2$, $a = 0.4$ and $\sigma^2 = 1$ is given in Figure 1. To implement stable MCEM, the compact sets $\{\mathcal{K}_n\}$ are chosen as a ball of radius $(n+1)$ centered at θ_0'. The Monte Carlo approximations are computed by use of the hybrid sampler described in Section 2.2. The proposal distribution for each component is a standard Gaussian variable on $\mathbb{R}$ (the mean acceptance rate is $\approx 40\%$). The chains are initialized in a compact ball of radius $r = 11$ according to a *concatenation* rule: if the last sample $Z_{m_n}^n$ at iteration n is in this ball, then it is the starting point of the following chain, that is, $Z_0^{n+1} := Z_{m_n}^n$; otherwise, we set $Z_0^{n+1} := r Z_{m_n}^n / |Z_{m_n}^n|$. The simulation schedule increases polynomially, $m_n := 1000 + n^2$. In Figure 2, we plot three paths of stable MCEM started, respectively, at $\theta_0' = \log(N^{-1} \sum Y_k) \approx 2.41$, $\theta_0' = -2$ and $\theta_0' = 4$. After 0, 3 and 2 reinitializations, respectively, convergence to the point

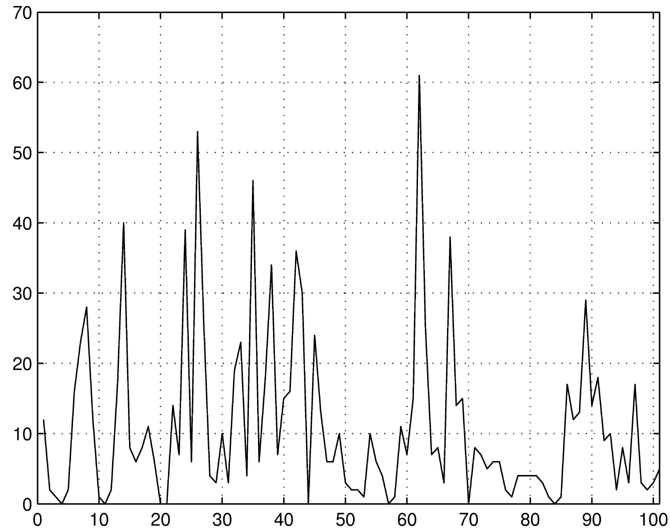

FIG. 1. 100 *observations from the Poisson count data model.*

$\theta^* \approx 1.88$ may be observed. In Figure 3, we plot a stable MCEM path started from $\theta_0' = \log(N^{-1} \sum Y_k)$ and its averaged counterpart [i.e., the sequence $\bar{\theta}_n$ given by $\bar{\theta}_n := \hat{\theta}(\Sigma_n)$]. Observe that the variation of the averaged path decreases more rapidly than the variation of the stable MCEM path, which illustrates the discussion in Section 3.2.

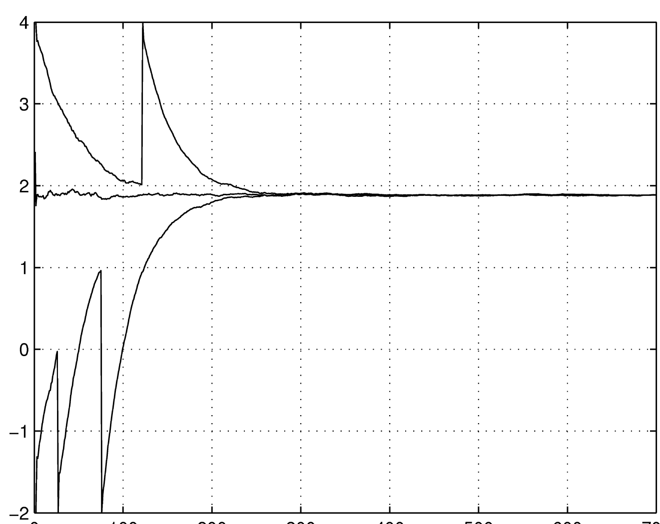

FIG. 2. *Stable MCEM sequences for different initial values and* $m_n = [n^2]$. *The paths all converge to* $\theta^* = 1.88$ *after a finite number of reinitializations.*

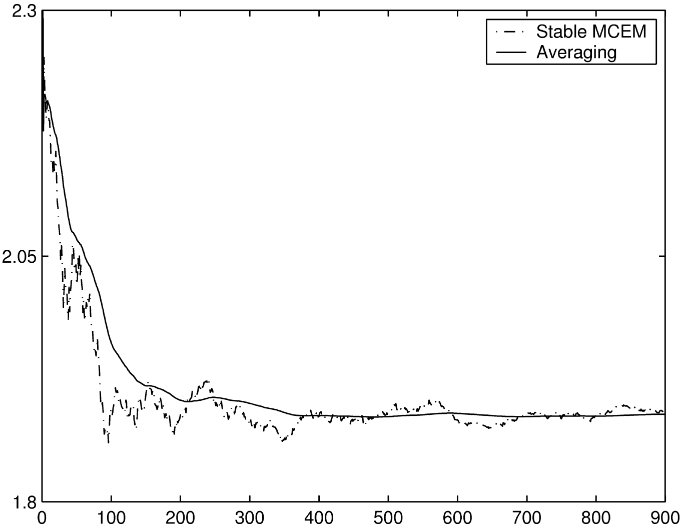

FIG. 3. *Stable MCEM sequence with and without averaging both started from* $\theta_0' = 2.41$; *polynomial schedule* $m_n = [n^2]$.

4. An application to product diffusion modeling. We illustrate the previous results by considering the Bass product diffusion model which consists in predicting market penetration of new products and services. Sherman, Ho and Dalal (1999) proved convergence in the case where the missing data were obtained (at each step) from m independent runs of a Gibbs sampler. These authors assumed uniform geometric ergodicity in the total variation distance and uniform convergence in L^2 [Assumptions (C5) and (C6)] which seemed difficult to directly verify in practice.

The observations $y := \{(t_1, n_1), \dots, (t_d, n_d)\}$ are the cumulative numbers n_j of adopters at a set of increasing instants t_j. We set $t_0 = n_0 := 0$. It is assumed that the n_j's are realizations of a process $N(t)$ at time t_j and the t_j's are selected independently of the adoption process. $N(t)$ is a pure birth Markov process with stationary transition probabilities and population adoption rate

$$\Lambda(t) := \big(M\pi - N(t)\big)\big(\varrho + \varsigma N(t)\big),$$

where M is the population size (M is known and constant over time), π is the proportion of potential adopters, $\varrho \geq 0$ is the innovator coefficient and $\varsigma \geq 0$ is the imitator coefficient. For all $0 \leq i \leq n_d - 1$, $\Lambda(t_i)$ has to be positive. In addition, for the expected number of adopters not to exceed the number of eventual adopters, we require $\varrho + \varsigma n_d \leq 1$. Hence $(\varrho, \varsigma, \pi) \in \Upsilon$, where

$$\Upsilon := \{(\varrho, \varsigma, \pi) \in (0, 1] \times [0, 1] \times [n_d/M, 1],\ 0 < \varrho + \varsigma n_d \leq 1\}.$$

Our purpose is to compute the maximum likelihood estimator for $\vartheta := (\varrho, \varsigma, \pi)$ or, equivalently, the maximum likelihood estimator for $\theta = (\alpha, \beta, \gamma) := \zeta(\vartheta)$ defined as

$$\zeta(\varrho, \varsigma, \pi) := \begin{bmatrix} -\varsigma \\ \varsigma M\pi - \varrho \\ \varrho M\pi \end{bmatrix},$$

$$\zeta^{-1}(\alpha, \beta, \gamma) := \begin{bmatrix} 1/2(-\beta + \sqrt{\beta^2 - 4\alpha\gamma}) \\ -\alpha \\ 2\gamma M^{-1}(-\beta + \sqrt{\beta^2 - 4\alpha\gamma})^{-1} \end{bmatrix}$$

so that $\zeta : \Upsilon \to \Theta := \zeta(\Upsilon)$ is continuous. Hence, we want to maximize on Θ the incomplete data likelihood g given by

$$g(\theta) := \prod_{j=1}^{d} \left(\prod_{k=n_{j-1}}^{n_j - 1} \lambda_k(\theta) \right) \times \sum_{i=n_{j-1}}^{n_j} \left(\exp\bigl(-\lambda_i(\theta)(t_j - t_{j-1})\bigr) \prod_{\substack{k=n_{j-1} \\ k \neq i}}^{n_j} \{\lambda_k(\theta) - \lambda_i(\theta)\}^{-1} \right),$$

where $\lambda_i(\theta) := \alpha i^2 + \beta i + \gamma$. Computation and maximization of g are not tractable [see Dalal and Weerahandi (1995)]. We thus implement the stable MCEM algorithm and solve a missing data problem, where missing data are individual adoption times. We write $g(\theta) := \int_{\mathcal{X}} h(\mathbf{z}; \theta)\mu(d\mathbf{z})$, where [see Sherman, Ho and Dalal (1999), Equation (11)]

$$\mathbf{z} := (z_1, \dots, z_{n_d}), \qquad z_0 := 0, \qquad \mathcal{X} := [0, t_d]^{n_d},$$

$$h(\mathbf{z}; \theta) := \prod_{i=0}^{n_d - 1} \lambda_i(\theta) \exp\bigl(-\lambda_i(\theta)(z_{i+1} - z_i)\bigr) \exp\bigl(-\lambda_{n_d}(\theta)(t_d - z_{n_d})\bigr)$$

and μ is absolutely continuous w.r.t. the Lebesgue measure on $\mathbb{R}^{n_d}$:

$$\mu(d\mathbf{z}) := \mathbb{1}_{0 < z_1 < \cdots < z_{n_d}} \prod_{j=1}^{d-1} \mathbb{1}_{z_{n_j} \le t_j < z_{n_j + 1}} \mathbb{1}_{z_{n_d} \le t_d} \, d\mathbf{z}.$$

Define $\psi(\theta) := \theta$ and

$$\phi(\theta) := -\lambda_{n_d}(\theta) t_d + \sum_{k=0}^{n_d - 1} \ln \lambda_k(\theta), \qquad S(\mathbf{z}) := \left[\sum_{k=1}^{n_d} (2k-1) z_k; \sum_{k=1}^{n_d} z_k; 0 \right],$$

so that $\log h(\mathbf{z}; \theta) = \phi(\theta) + \langle S(z); \theta \rangle$. M2(a) is readily verified and, as g is continuous on Θ, M2(b) follows from an application of the Lebesgue theorem. It is trivial to verify that for all $\theta \in \Theta$, $s \in \mathcal{S}$, $-\nabla_\theta^2 L(s;\theta)$ is positive definite. Then, for all $s \in \mathcal{S}$, the function $\theta \mapsto L(s;\theta)$ is strictly concave on Θ and $s \mapsto \hat{\theta}(s)$ is well defined on $\mathcal{S}$. By applying the implicit function theorem, $\hat{\theta}$ is also continuous. M2(c) is thus verified. $\vartheta \mapsto g \circ \zeta(\vartheta)$ is a positive and continuous function on Υ and $\lim_{\varrho \to 0} g \circ \zeta(\varrho, \varsigma, \pi) = 0$ for any (ς, π), showing that the level sets $\{g \circ \zeta \geq M\}$, $M > 0$, are compact subsets of Υ. As ζ is continuous, the level sets $\{g \geq M\}$ are compact subsets of Θ and M2(d) holds. Finally, $\mathcal{L}$ is a closed subset of the bounded set Θ which proves M3(a).

To impute the missing values $\mathbf{z}$, we use a Metropolis Hastings independent sampler (IS) with proposal distribution $q\,d\mu$ which is chosen as the product of d distributions of the order statistics of $(n_k - n_{k-1})$ independent random variables uniformly distributed on $[t_{k-1}, t_k]$, $1 \leq k \leq d$, that is,

$$q(\mathbf{z})\mu(d\mathbf{z}) := \left[\prod_{k=1}^{d} \frac{(t_k - t_{k-1})^{n_k - n_{k-1}}}{(n_k - n_{k-1})!}\right]^{-1} \mathbb{1}_{0 < z_1 < \cdots < z_{n_d}} \times \prod_{j=1}^{d-1} \mathbb{1}_{z_{n_j} \leq t_j < z_{n_j+1}} \mathbb{1}_{z_{n_d} \leq t_d}\, d\mathbf{z}.$$

Recall that for a homogeneous Poisson process of rate λ, the conditional distributions of the arrivals in a given interval given the number of arrivals is i.i.d. uniform over that interval so that the choice of the proposal is well matched to the target density. With these definitions, the IS kernel, P_θ, is Lebesgue-irreducible and aperiodic. It is easily seen that the target density $p(\mathbf{z};\theta)$ is uniformly bounded for θ in a compact set $\mathcal{K} \subseteq \Theta$. Thus, there exists some minorizing constant $0 < \varepsilon < 1$ such that $\varepsilon p(\mathbf{z};\theta) \leq q(\mathbf{z})$ for all $\theta \in \mathcal{K}$, $\mathbf{z} \in \mathcal{X}$. Hence, for $\mathbf{z} \in \mathcal{X}$, any measurable set A,

$$P_\theta(\mathbf{z}, A) \geq \int_A \alpha_\theta(\mathbf{z}, \mathbf{z}') q(\mathbf{z}')\mu(d\mathbf{z}') \geq \varepsilon \int_A p(\mathbf{z}';\theta)\mu(d\mathbf{z}') = \varepsilon \pi_\theta(A),$$

where $\alpha_\theta(\mathbf{z}, \mathbf{z}')$ is the acceptance ratio. The condition M4 follows from Proposition 1, with any $p \geq 2$ and any probability measure λ on $\mathcal{X}$.

Simulations 1. We generate $d := 30$ observations at time $t_j := 0.25j$ by choosing $M := 2000$, $(\varrho_t, \varsigma_t, \pi_t) := (0.03, 0.0004, 0.5)$, that is, $(\alpha_t, \beta_t, \gamma_t) = (-0.0004, 0.37, 30)$. The corresponding cumulative numbers n_j appear as stars in Figure 4 (we have $n_d = 651$). The parameter space Θ is covered by the increasing sequence of compact sets

$$\mathcal{K}_n := \zeta\big(\{(\varrho, \varsigma),\ 0.0003/2^n \leq \varrho \leq 1,\ 0 \leq \varsigma \leq 1,\ 0 \leq \varrho + \varsigma n_d \leq 1\} \times [n_d/M, 1]\big), \qquad n \geq 0.$$

The initial distribution λ of the Markov chains coincides with the proposal distribution of the independent sampler $q\,d\mu$ described above.

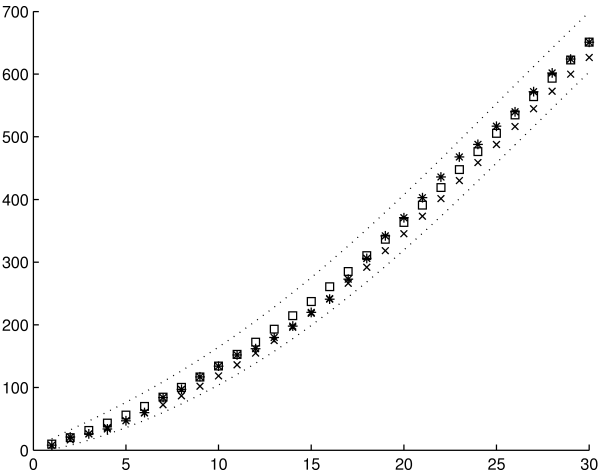

FIG. 4. *Cumulative numbers observed at time* $t_j = 0.25j$, $j = 1, \dots, 30$, *and estimated means of the count process.*

Two paths of stable MCEM started, respectively, at $\theta'_0 = (-5 \times 10^{-5}, 0.0321, 0.3260)$ (path 1) and $\theta'_0 = (-4 \times 10^{-5}, -0.24, 450)$ (path 2) are run for 300 iterations. The number of simulations at each iteration increases polynomially, $m_n = 20 + n^{1.2}$. After four and zero reinitializations, respectively, and a small number of iterations, the convergence of both paths to $\theta^* \sim (-0.00027, 0.2965, 37.41)$ can be observed. In Figure 5, we plot the stable MCEM sequences $\{\gamma_n\}$ that both

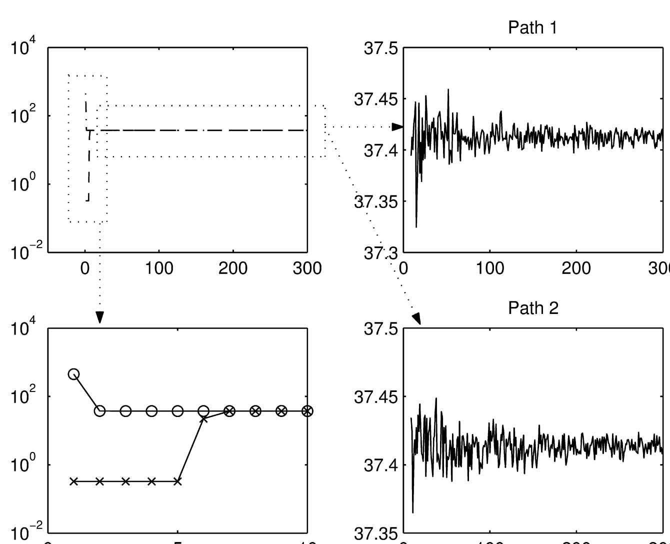

FIG. 5. *Stable MCEM sequences for different initial values and* $m_n = [n^{1.2}]$. *The paths converge to* $\gamma^* = 37.41$ *after a finite number of reinitializations.*

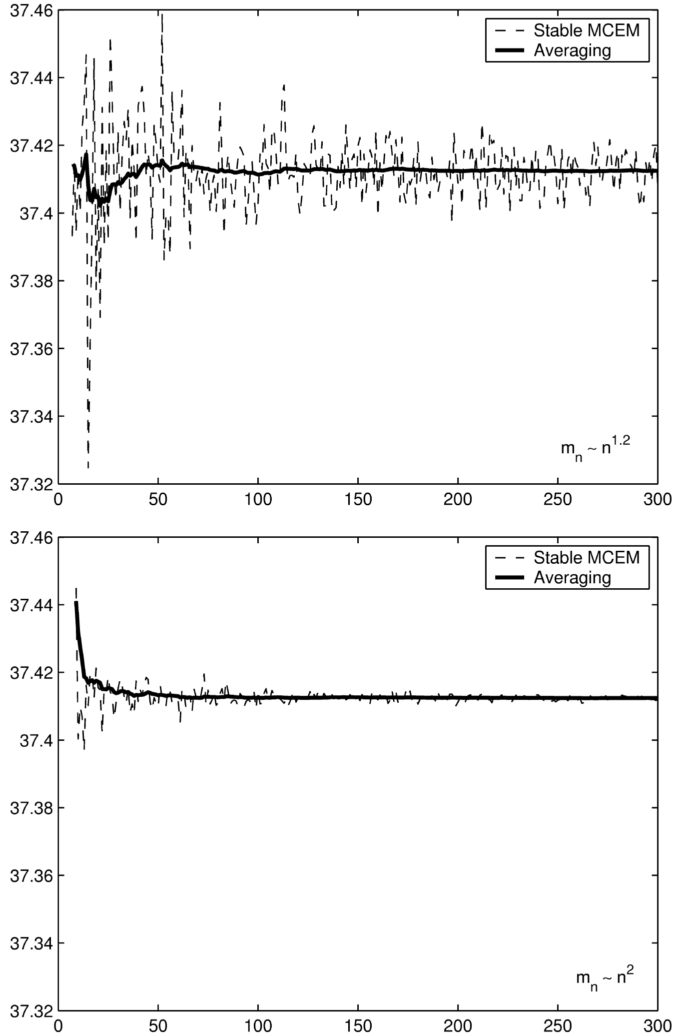

FIG. 6. *MCEM sequence and averaged MCEM sequence for different polynomial schedules. The* 10 *initial values are omitted.*

converge to $\gamma^* = 37.41$. In the lower left-hand corner, the first 10 values are drawn, showing (a) the four reinitializations on path 1 and (b) for both paths, the rapid move toward a neighborhood of the limiting value γ^*. The two paths are drawn in the right subplots (from iteration 9 to 300), showing the convergence to the same limiting point γ^* and a similar variation of the paths.

We then observed the performance of stable MCEM and the averaged counterpart for two polynomial schedules, $m_n \sim n^{1.2}$ and $m_n \sim n^2$. The procedures, run for 300 iterations, start from $\theta_0' = (-5 \times 10^{-5}, 0.0321, 0.3260)$. In Figure 6, we plot the sequences $\{\gamma_n\}$ and $\{\bar{\gamma}_n\}$ obtained by the stable MCEM algorithm and the averaging procedure, respectively (the first 10 values are discarded). In all cases, convergence to $\gamma^* = 37.41$ can be observed. Contrary to the variation of the averaged stable MCEM path, the variation of stable MCEM paths depends on the simulation schedule. Observe that averaging smooths out the trajectory and improves the rate of convergence. The same conclusions can be drawn from the sequences $\{\alpha_n\}$, $\{\beta_n\}$, $\{\bar{\alpha}_n\}$ and $\{\bar{\beta}_n\}$, the plots of which are omitted.

Dalal and Weerahandi (1992) derived approximations of mean and variance of the Poisson process $N(t)$. The estimates of the mean functions $\mathbb{E}[N(t_j)]$ computed from the true value of the parameter θ_t (resp. the stable MCEM estimate θ^*) appear as $\times$ marks (resp. squares) on Figure 4. The dotted curves interpolate points that correspond to ± 2 estimated standard errors from the estimates of the mean $\mathbb{E}[N(t_j)]$.

Simulations 2. Consider now prediction of the number of wireless telecommunication services in the United States. The Cellular Telecommunications Industry Association performed semiannual surveys, collected in June and December, from January 1985 to June 2001 (the data are available on the web site www.wow-com.com/industry/stats/surveys/). In Figure 7, the 34 observations collected at times $1, 2, \ldots, 34$ appear as stars. We assume that this count follows a pure birth Markov model (our results suggest it is a good approximation). Since the same person may subscribe to different wireless services, the (true) population size M is unknown. As discussed in Sherman, Ho and Dalal (1999) M and π enter the model through the product $M\pi$, so any value $M > n_d$ is convenient. As $n_d \sim 9 \times 10^7$, we set $M = 10^9$.

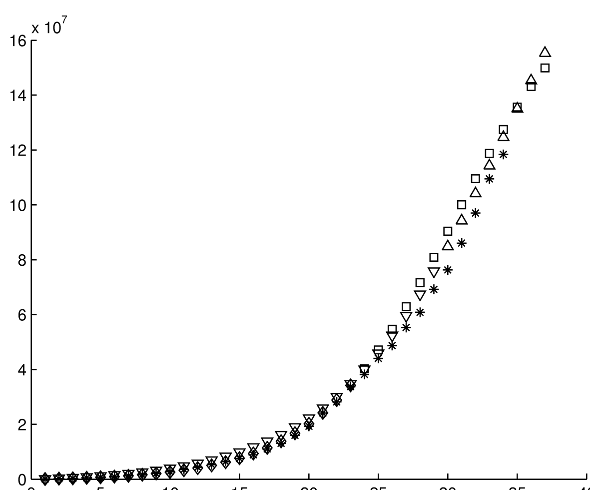

FIG. 7. *Cumulative numbers and estimated means of the count process.*

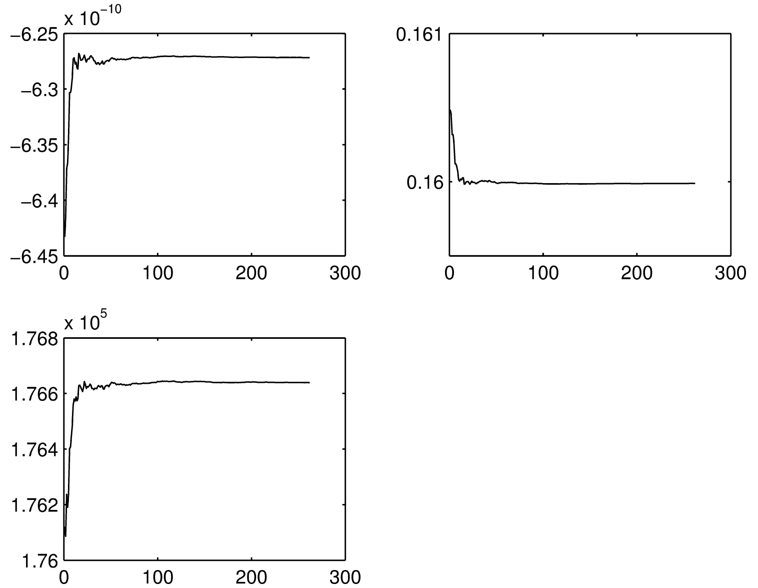

FIG. 8. *Stable MCEM sequence with averaging*: $\{\bar{\alpha}_n\}$ (*upper left*), $\{\bar{\beta}_n\}$ (*upper right*), $\{\bar{\gamma}_n\}$ (*lower left*).

Our estimate is computed from the 29 values collected from January 1985 to December 1998 and the last values are used to cross-validate the result. The estimate is computed as the limiting value of a path $\{\bar{\theta}_n\}$ of the averaged procedure run for 260 iterations with $m_n \sim n^2$ and started at $\theta'_0 = (-5 \times 10^{-5}$, 0.0341, 0.3465). The paths of $\{\bar{\alpha}_n\}$, $\{\bar{\beta}_n\}$ and $\{\bar{\gamma}_n\}$ are plotted in Figure 8. The limiting value is $\theta^* = (-6.27 \times 10^{-10}, 1.60 \times 10^{-1}, 1.77 \times 10^5)$, that is, $(\varrho^*, \varsigma^*, \pi^*) = (6.9 \times 10^{-4}, 6.27 \times 10^{-10}, 0.26)$. The fitted values (resp. the predicted values) of the mean function $\mathbb{E}[N(t_j)]$ for $j \in [1, 29]$ (resp. $j \in [30, 37]$) appear as down triangles in Figure 7 (resp. up triangles). Sherman, Ho and Dalal (1999) provided an estimate θ^* of θ_t based on the first 23 values collected from January 1985 to December 1995. They obtained $(\varrho^*, \varsigma^*, \pi^*) = (4.3 \times 10^{-4}, 1.06 \times 10^{-10}, 0.19)$, that is, $\theta^* = (-1.06 \times 10^{-9}, 0.20, 8.17 \times 10^5)$. Their fitted values (resp. their predicted values) of the mean function $\mathbb{E}[N(t_j)]$ for $j \in [1, 23]$ (resp. $j \in [24, 37]$) are represented as diamonds in Figure 7 (resp. squares). In both cases, the extrapolated values track the observed data well.

5. Proof of Theorem 3. Let $T : \Theta \to \Theta$ be a point-to-point map. Let $\mathcal{L}$ be a nonempty subset of Θ. A positive function W defined on Θ is said to be a Lyapunov function relative to $(T, \mathcal{L})$ when (i) for all $u \in \Theta$, $W \circ T(u) - W(u) \geq 0$ and (ii) for any compact set $\mathcal{K} \subseteq \Theta \setminus \mathcal{L}$, $\inf_{u \in \mathcal{K}}\{W \circ T(u) - W(u)\} > 0$. In the literature, convergence of random iterative maps $\{F_n\}$ that approximate a deterministic iterative map T having a Lyapunov function W is addressed under the assumption that, for all compact set $\mathcal{K}$,

$$\lim_n \sup_{u \in \mathcal{K}} |W \circ F_n(u) - W \circ T(u)| = 0.$$

Applied to the present problem, this condition is often not checked when MCMC algorithms are used to perform Monte Carlo integration. In this section, we show how this condition can be replaced by the weaker condition

$$\lim_n |W \circ F_n(u_n) - W \circ T(u_n)| \mathbb{1}_{u_n \in \mathcal{K}} = 0.$$

5.1. *Deterministic results.*

PROPOSITION 9. *Let $\Theta \subseteq \mathbb{R}^l$, let $\mathcal{K}$ be a compact subset of Θ and let $\mathcal{L} \subseteq \Theta$ be such that $\mathcal{L} \cap \mathcal{K}$ is compact. Let W be a continuous Lyapunov function relative to $(T, \mathcal{L})$. Assume that there exists a $\mathcal{K}$-valued sequence $\{u_n\}$ such that $\lim_n |W(u_{n+1}) - W \circ T(u_n)| = 0$. Then $\{W(u_n)\}$ converges to a connected component of $W(\mathcal{L} \cap \mathcal{K})$. If $W(\mathcal{L} \cap \mathcal{K})$ has an empty interior, $\{W(u_n)\}$ converges to $w^\star$ and $\{u_n\}$ converges to the set $\mathcal{L}_{w^\star} \cap \mathcal{K}$, where $\mathcal{L}_{w^\star} := \{\theta \in \mathcal{L}, W(\theta) = w^\star\}$.*

PROOF. Define the compact set $\mathcal{D} := W(\mathcal{L} \cap \mathcal{K})$. Let $\mathcal{D}_\alpha$ be the α neighborhood of the closed set $\mathcal{D}$ in $\mathbb{R}$, $\mathcal{D}_\alpha := \{x \in \mathbb{R}, d(x, \mathcal{D}) < \alpha\}$. Where, as $\mathcal{D}$ is compact, $\mathcal{D} = \bigcap_{\alpha > 0} \mathcal{D}_\alpha$. Let $\alpha > 0$. Since $\mathcal{D}_\alpha$ is a *finite* union of disjoint bounded open intervals, there exist $n_\alpha \geq 0$ and two increasing real-valued sequences $\{a_\alpha(k)\}$ and $\{b_\alpha(k)\}$, $1 \leq k \leq n_\alpha$, such that

$$\mathcal{D}_\alpha = \bigcup_{k \in \{1, \ldots, n_\alpha\}} \bigl(a_\alpha(k), b_\alpha(k)\bigr). \tag{19}$$

$W^{-1}(\mathcal{D}_{\alpha/2})$ is an open neighborhood of $\mathcal{L} \cap \mathcal{K}$, and we define

$$\varepsilon_\alpha := \inf_{\{u \in \mathcal{K} \setminus W^{-1}(\mathcal{D}_{\alpha/2})\}} \{W \circ T(u) - W(u)\} \quad \text{and} \quad \rho_\alpha := \varepsilon_\alpha \wedge \alpha. \tag{20}$$

Since $\mathcal{K} \setminus W^{-1}(\mathcal{D}_{\alpha/2})$ is a compact subset of $\mathbb{R}^d$, ε_α and ρ_α are both positive. We define $\eta_{n+1} := W(u_{n+1}) - W \circ T(u_n)$. Then

$$W(u_{n+1}) - W(u_n) = W \circ T(u_n) - W(u_n) + \eta_{n+1} \tag{21}$$

and there exists $N_\alpha \geq 0$, such that for any $n \geq N_\alpha$,

$$|\eta_{n+1}| \leq \rho_\alpha / 2. \tag{22}$$

By (20) and (21),

$$\bigl(n \geq N_\alpha \text{ and } u_n \in \mathcal{K} \setminus W^{-1}(\mathcal{D}_{\alpha/2})\bigr) \implies W(u_{n+1}) - W(u_n) \geq \rho_\alpha / 2. \tag{23}$$

Define $k_\alpha^\star := \min\{1 \leq k \leq n_\alpha, \limsup_n W(u_n) < b_\alpha(k)\}$ and $I(\alpha) := (a_\alpha(k_\alpha^\star); b_\alpha(k_\alpha^\star))$. Equation (23) shows that $\{W(u_n)\}$ is infinitely often (i.o.) in $\mathcal{D}_{\alpha/2} \subset \mathcal{D}_\alpha$, and since $\mathcal{D}_\alpha$ is a finite union of intervals, $\{W(u_n)\}$ is i.o. in an interval of (19); thus, $\limsup_n W(u_n) \in I(\alpha)$. Let $p \geq N_\alpha$ such that $W(u_p) \in I(\alpha)$. We prove by induction that for all $n \geq p$, $W(u_n) \in I(\alpha)$. By definition, $W(u_p) \in I(\alpha)$. Assume now that for $p \leq k \leq n$, $W(u_k) \in I(\alpha)$.

- If $W(u_n) \in \mathcal{D}_{\alpha/2}$, we have $W(u_n) \geq a_\alpha(k_\alpha^*) + \alpha/2$. Thus,
$$W(u_{n+1}) \geq W(u_n) + \eta_{n+1} \geq a_\alpha(k_\alpha^*) + \alpha/2 - \rho_\alpha/2 \geq a_\alpha(k_\alpha^*).$$
- If $W(u_n) \in \mathcal{D}_\alpha \setminus \mathcal{D}_{\alpha/2}$, then under (20), $W \circ T(u_n) - W(u_n) \geq \rho_\alpha$, and (21) and (22) imply that $W(u_{n+1}) \geq a_\alpha(k_\alpha^*) + \rho_\alpha/2 \geq a_\alpha(k_\alpha^*)$.

Hence, the set of the limit points $\mathcal{I}$ of $\{W(u_n)\}$ is nonempty and included in the interval $I(\alpha)$. Let $0 < \alpha_1 < \alpha_2$. By definition, $\mathcal{D}_{\alpha_1} \subset \mathcal{D}_{\alpha_2}$; thus $I(\alpha_1) \subset I(\alpha_2)$ and $\mathcal{I} \subset I(\alpha_1) \cap I(\alpha_2)$. Let $\{\alpha_n\}$ be a decreasing sequence such that $\lim_n \alpha_n = 0$. Then $\mathcal{I} \subset \bigcap_n I(\alpha_n)$. $\{I(\alpha_n)\}$ is a decreasing sequence of intervals, $\bigcap_n I(\alpha_n)$ is an interval and $\bigcap_n I(\alpha_n) \subset W(\mathcal{L} \cap \mathcal{K})$. Hence, $\{W(u_n)\}$ converges to this interval which concludes the first part of the proof. The last part is a consequence of (21). $\square$

It is proved in Proposition 10 that the compactness assumption of the sequence $\{u_n\}$ can be replaced by a recurrence condition, provided that there exists a Lyapunov function that controls excursion outside the compact sets of Θ. In Proposition 11, we propose a stabilization procedure that ensures this recurrence property for sequences $\{u_n\}$ defined by inhomogeneous maps, $u_{n+1} = F_n(u_n)$.

PROPOSITION 10. *Let $\Theta \subseteq \mathbb{R}^l$, $T : \Theta \to \Theta$ and $\mathcal{L} \subset \Theta$. Assume the following*:

A1. *There exists a continuous Lyapunov function W for $(T, \mathcal{L})$ such that* (a) *for all $M > 0$, the level set $\{\theta \in \Theta, W(\theta) \geq M\}$ is compact and* (b) $\Theta = \bigcup_{n \geq 1}\{\theta \in \Theta, W(\theta) \geq n\}$.

A2. *$W(\mathcal{L})$ is compact or* A2′ *$W(\mathcal{L} \cap \mathcal{K})$ is finite for all compact sets $\mathcal{K} \subseteq \Theta$.*

A3. *There exists a Θ-valued sequence $\{u_n\}$ such that* (a) *$\{u_n\}$ is infinitely often in a compact subset $\mathcal{G} \subseteq \Theta$ and* (b) *for any compact set $\mathcal{K} \subseteq \Theta$, $\lim_n |W(u_{n+1}) - W \circ T(u_n)| \mathbb{1}_{u_n \in \mathcal{K}} = 0$.*

Then $\{u_n\}$ is in a compact subset of Θ.

PROOF (under assumption A2). Let $\alpha > 0$. Under A1(b) and A2, there exists $M > 0$ such that
$$\mathcal{G} \cup \mathcal{L}_\alpha \subset \{\theta \in \Theta,\ W(\theta) \geq M\},$$
where $\mathcal{L}_\alpha$ is the α neighborhood of $\mathcal{L}$. Define
$$\varepsilon := \inf_{\{\theta \in \Theta, W(\theta) \geq M-1\} \setminus \mathcal{L}_\alpha} \{W \circ T(\theta) - W(\theta)\} \quad \text{and} \quad \rho := \varepsilon \wedge 1. \tag{24}$$

By assumption, $\varepsilon > 0$ and $\rho > 0$. Define $\eta_{n+1} := W(u_{n+1}) - W \circ T(u_n)$. Under A3, there exists N such that

$$(n \geq N \text{ and } u_n \in \{\theta \in \Theta, W(\theta) \geq M - 1\}) \quad \Longrightarrow \quad |\eta_{n+1}| \leq \rho/2. \tag{25}$$

Note that

$$W(u_{n+1}) - W(u_n) = W \circ T(u_n) - W(u_n) + \eta_{n+1}. \tag{26}$$

Since $\{u_n\}$ is infinitely often in the compact set $\mathcal{G}$, there exists $p \geq N$ such that $W(u_p) \geq M - 1$. We show by induction that for all $n \geq p$, $W(u_n) \geq M - 1$. The property holds for $n = p$. Assume it holds for $p \leq k \leq n$.

- If $u_n \in \{\theta \in \Theta,\ W(\theta) \geq M\}$, then (24)–(26) imply that $W(u_{n+1}) \geq W(u_n) - \rho/2 \geq M - 1/2 \geq M - 1$.
- If $u_n \in \{\theta \in \Theta,\ W(\theta) \geq M - 1\} \setminus \mathcal{L}_\alpha$, then (24)–(26) imply that $W(u_{n+1}) \geq W(u_n) + \varepsilon - \rho/2 \geq W(u_n) \geq M - 1$.

Hence for any $q \geq n$, u_q is in the compact set $\{\theta \in \Theta,\ W(\theta) \geq M - 1\}$. □

PROOF (under assumption A2′). By assumption, there exists M such that $\mathcal{G} \subset \{\theta \in \Theta,\ W(\theta) \geq M\}$. As $W(\mathcal{L} \cap \{\theta, W(\theta) \geq M - 1\})$ is finite, there exist $\alpha > 0$ and $M - 1 \leq M'' < M' < M$, such that

$$\mathcal{L}_\alpha \cap \{\theta \in \Theta,\ W(\theta) \geq M''\} \subset \{\theta \in \Theta,\ W(\theta) \geq M'\}.$$

Define

$$\varepsilon := \inf_{\{\theta \in \Theta, W(\theta) \geq M''\} \setminus \mathcal{L}_\alpha} \{W \circ T(\theta) - W(\theta)\} \quad \text{and} \quad \rho := \varepsilon \wedge (M' - M'').$$

It may be proved that for all large q, u_q is in the compact set $\{\theta \in \Theta, W(\theta) \geq M''\}$. The proof is along the same lines as the previous one and is omitted for brevity. □

Let $\{F_n\} : \Theta \to \Theta$ be a family of point-to-point maps. Choose a sequence of compact subsets $\{\mathcal{K}_n\}$ of Θ such that for any $n \geq 0$,

$$\mathcal{K}_n \subsetneq \mathcal{K}_{n+1}, \qquad \Theta = \bigcup_{n \geq 0} \mathcal{K}_n.$$

Let $u_0 \in \mathcal{K}_0$. Set $p_0 := 0$ and for $n \geq 0$,

$$\begin{array}{lll} \text{if } F_n(u_n) \in \mathcal{K}_{p_n}, & u_{n+1} := F_n(u_n) \text{ and } p_{n+1} := p_n; \\ \text{if } F_n(u_n) \notin \mathcal{K}_{p_n}, & u_{n+1} := u_0 \text{ and } p_{n+1} := p_n + 1. \end{array} \tag{27}$$

PROPOSITION 11. *Let $\Theta \subseteq \mathbb{R}^l$, and let T and $\{F_n\}$ be point-to-point maps onto Θ. Let $\{u_n\}$ be the sequence given by* (27). *Assume* (a) A1 and A2 *hold*, (b) *for all $u \in \mathcal{K}_0$, $\lim_n |W \circ F_n - W \circ T|(u) = 0$ and* (c) *for any compact subset $\mathcal{K} \subseteq \Theta$, $\lim_n |W \circ F_n(u_n) - W \circ T(u_n)| \mathbb{1}_{u_n \in \mathcal{K}} = 0$. Then $\limsup_n p_n < \infty$ and $\{u_n\}$ is a compact sequence.*

The proof is along the same lines as Proposition 10 and is omitted for brevity.

5.2. *Proof of Theorem* 3. Given λ, θ'_0 and the sequence of compact sets $\{\mathcal{K}_n\}$, the process $\{\theta'_n\}$ is defined on the canonical space of the inhomogeneous Markov chain $\{(\tilde{S}_n, p_n)\}$. We denote by $\bar{\mathbb{P}}$ (resp. $\overline{\mathbb{E}}$) the probability (resp. the expectation) of this canonical Markov chain (the dependence upon λ, θ'_0 and $\{\mathcal{K}_n\}$ is omitted). We apply Propositions 9 and 11 with the EM map $T := \hat{\theta} \circ \bar{S}$ and the random sequence of maps $\{F_n\}$, $F_n(\theta) := \arg\max_{\phi \in \Theta} \mathcal{Q}_n(\phi, \theta)$.

PROOF OF (i)(a). We check the conditions of Proposition 11. It is well known that the incomplete data likelihood g is a natural Lyapunov function relative to the EM map T and to the set $\mathcal{L}$ of the fixed points of T. Under M1–M3, the conditions A1 and A2 are verified with $W = g$. Let $\varepsilon > 0$ and let $\mathcal{K} \subseteq \Theta$ be compact. We prove that $\sum_n \mathbb{1}_{\{|g \circ F_n(\theta'_n) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon\}}$ is finite w.p.1. By the second Borel–Cantelli lemma, the convergence of the series is implied by the convergence of $\sum_n \bar{\mathbb{P}}(|g \circ F_n(\theta'_n) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon | \mathcal{F}_{n-1})$ w.p.1, where $\mathcal{F}_n := \sigma(\tilde{S}_k,\ k \leq n)$. By assumption, $\bar{S}(\mathcal{K})$ is a compact subset of $\mathcal{S}$. For $\delta > 0$, define the compact $\bar{S}(\mathcal{K}, \delta) := \{s \in \mathbb{R}^q,\ \inf_{t \in \mathcal{K}} |t - s| \leq \delta\}$. Then there exists $\eta(\varepsilon, \delta)$ such that for any $x, y \in \bar{S}(\mathcal{K}, \delta)$,

$$|x - y| \leq \eta(\varepsilon, \delta) \quad \Longrightarrow \quad |g \circ \hat{\theta}(x) - g \circ \hat{\theta}(y)| \leq \varepsilon.$$

Hence,

$$\begin{aligned}
&\bar{\mathbb{P}}\big(|g \circ F_n(\theta'_n) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon \big| \mathcal{F}_{n-1}\big) \\
&\quad = \bar{\mathbb{P}}\big(|g \circ \hat{\theta}(\tilde{S}_n) - g \circ \hat{\theta}(\bar{S}(\theta'_n))| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon \big| \mathcal{F}_{n-1}\big) \\
&\quad = \bar{\mathbb{P}}\big(|g \circ \hat{\theta}(\tilde{S}_n) - g \circ \hat{\theta}(\bar{S}(\theta'_n))| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon, |\tilde{S}_n - \bar{S}(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \leq \delta \big| \mathcal{F}_{n-1}\big) \\
&\qquad + \bar{\mathbb{P}}\big(|g \circ \hat{\theta}(\tilde{S}_n) - g \circ \hat{\theta}(\bar{S}(\theta'_n))| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon, |\tilde{S}_n - \bar{S}(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} > \delta \big| \mathcal{F}_{n-1}\big) \\
&\quad \leq 2\bar{\mathbb{P}}\big(|\tilde{S}_n - \bar{S}(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \alpha \big| \mathcal{F}_{n-1}\big)
\end{aligned}$$

with $\alpha := \delta \wedge \eta(\varepsilon, \delta)$. Thus,

$$\begin{aligned}
&\bar{\mathbb{P}}\big(|g \circ F_n(\theta'_n) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon \big| \mathcal{F}_{n-1}\big) \\
&\quad \leq 2\alpha^{-p} \overline{\mathbb{E}}\big[|\tilde{S}_n - \bar{S}(\theta'_n)|^p \big| \mathcal{F}_{n-1}\big] \mathbb{1}_{\theta'_n \in \mathcal{K}} \\
&\quad \leq 2\alpha^{-p} m_n^{-p} \mathbb{E}_{\lambda, \theta'_n}\left[\left|\sum_{j=1}^{m_n} \{S(\Phi_j) - \pi_{\theta'_n}(S)\}\right|^p\right] \mathbb{1}_{\theta'_n \in \mathcal{K}},
\end{aligned}$$

where p is given by M4. Then M4 implies that there exists a finite constant $C := C(\mathcal{K})$ such that

$$\mathbb{E}_{\lambda, \theta'_n}\left[\left|\sum_{j=1}^{m_n} \{S(\Phi_j) - \pi_{\theta'_n}(S)\}\right|^p\right] \mathbb{1}_{\theta'_n \in \mathcal{K}} \leq C m_n^{p/2}$$

and, under M5, the proof is concluded. □

PROOF OF (i)(b) AND (ii). We check the conditions of Proposition 9. It remains to prove that for any compact set $\mathcal{K} \subseteq \Theta$,

$$\lim_n |g(\theta'_{n+1}) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} = 0, \qquad \bar{\mathbb{P}}\text{-a.s.}$$

We proceed as above and consider the a.s. convergence of the random series

$$\sum_n \bar{\mathbb{P}}\big(|g(\theta'_{n+1}) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon \big| \mathcal{F}_{n-1}\big). \tag{28}$$

By definition, either $\theta'_{n+1} = F_n(\theta'_n)$ or $\theta'_{n+1} = \theta'_0$ and $p_{n+1} = p_n + 1$. We have just proved that the number of reinitializations is finite w.p.1, so that the series

$$\sum_n \bar{\mathbb{P}}\big(|g(\theta'_{n+1}) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon,\ \theta'_{n+1} = \theta'_0, p_{n+1} = p_n + 1 \big| \mathcal{F}_{n-1}\big)$$

is finite $\bar{\mathbb{P}}$-a.s. Then (28) is finite iff $\sum_n \bar{\mathbb{P}}(|g(\theta'_{n+1}) - g \circ T(\theta'_n)| \mathbb{1}_{\theta'_n \in \mathcal{K}} \geq \varepsilon, \theta'_{n+1} = F_n(\theta'_n) | \mathcal{F}_{n-1})$ is finite $\bar{\mathbb{P}}$-a.s., which was established above. $\square$

6. Uniform Rosenthal's inequality. Let $f : \mathcal{X} \to [1, \infty)$ be a measurable function. For some function $g : \mathcal{X} \to \mathbb{R}^q$ (resp. for some signed measure ν on $\mathcal{X}$), define

$$\|g\|_f := \sup_{\mathcal{X}} \frac{|g|}{f}, \qquad \mathcal{L}_f := \{g : \mathcal{X} \to \mathbb{R}^q,\ \|g\|_f < \infty\},$$
$$\|\nu\|_f := \sup_{\{g, |g| \leq f\}} |\nu(g)|.$$

PROPOSITION 12. *Let $(\Omega, \mathcal{A}, \mathcal{F}_n, \{\phi_n\}, P_x)$ be a canonical Markov chain with invariant probability measure π on $\mathcal{X}$. Assume that there exist $p \geq 2$, some measurable functions $1 \leq f_0 \leq V_0 \leq V_0^p \leq V_1 < \infty$ and some constants $C_i < \infty$, $i = 0, 1$, such that for any $x \in \mathcal{X}$,*

$$\begin{aligned} &\sum_n \|P^n(x, \cdot) - \pi(\cdot)\|_{f_0} \leq C_0 V_0(x), \\ &\sum_n \|P^n(x, \cdot) - \pi(\cdot)\|_{V_0^p} \leq C_1 V_1(x). \end{aligned} \tag{29}$$

Then for any Borel function $g : \mathcal{X} \to \mathbb{R}^q$, $g \in \mathcal{L}_{f_0}$,

$$\mathbb{E}_x \left| \sum_{k=1}^n \{g(\Phi_k) - \pi(g)\} \right|^p \leq \|g\|_{f_0}^p 6^p C_p C_0^p \big(C_1 V_1(x) + \pi(V_0^p)\big) n^{p/2}, \qquad x \in \mathcal{X},$$

where C_p is Rosenthal's constant.

PROOF. Denote by $\hat{g}(x) := \sum_{k=0}^{\infty}\{P^k g(x) - \pi(g)\}$ the unique solution (up to a constant) of the Poisson equation $\hat{g} - P\hat{g} = g - \pi(g)$. Then $\hat{g} \in \mathcal{L}_{V_0}$ and $\|\hat{g}\|_{V_0} \le C_0\|g\|_{f_0}$. Write

$$\sum_{k=1}^{n}\{g(\Phi_k) - \pi(g)\} = \sum_{k=1}^{n}\{\hat{g}(\Phi_k) - P\hat{g}(\Phi_{k-1})\} - P\hat{g}(\Phi_n) + P\hat{g}(\Phi_0),$$

where $\{\hat{g}(\Phi_k) - P\hat{g}(\Phi_{k-1})\}$ is an L^p-martingale increment (w.r.t. the initial distribution δ_x). By applying Minkowsky's inequality and Rosenthal's inequality [Hall and Heyde (1980), Theorem 2.12], we get

$$\begin{aligned}
&\mathbb{E}_x\left[\left|\sum_{k=1}^{n}\{g(\Phi_k) - \pi(g)\}\right|^p\right] \\
&\quad \le 3^{p-1}\Biggl\{C_p\mathbb{E}_x\left[\left(\sum_{k=1}^{n}\mathbb{E}_x\bigl[|\hat{g}(\Phi_k) - P\hat{g}(\Phi_{k-1})|^2\big|\mathcal{F}_{k-1}\bigr]\right)^{p/2}\right] \\
&\qquad\qquad + C_p\mathbb{E}_x\left[\sum_{k=1}^{n}|\hat{g}(\Phi_k) - P\hat{g}(\Phi_{k-1})|^p\right] \\
&\qquad\qquad\qquad + \mathbb{E}_x\bigl[|P\hat{g}(\Phi_n)|^p\bigr] + |P\hat{g}(x)|^p\Biggr\},
\end{aligned}$$

where C_p is Rosenthal's constant and $\{\mathcal{F}_n\}$ is the natural filtration of the Markov chain $\{\Phi_n\}$. In addition,

$$\begin{aligned}
&\left(\sum_{k=1}^{n}\mathbb{E}_x\bigl[|\hat{g}(\Phi_k) - P\hat{g}(\Phi_{k-1})|^2\big|\mathcal{F}_{k-1}\bigr]\right)^{p/2} \\
&\qquad \le \left(\sum_{k=1}^{n}P|\hat{g}|^2(\Phi_{k-1})\right)^{p/2} \le n^{p/2-1}\sum_{k=1}^{n}P|\hat{g}|^p(\Phi_{k-1}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}_x\left[\left|\sum_{k=1}^{n}\{g(\Phi_k) - \pi(g)\}\right|^p\right] \\
&\quad \le 3^{p-1}\left(C_p(n^{p/2-1} + 2^p)\sum_{k=1}^{n}P^k|\hat{g}|^p(x) + P|\hat{g}|^p(x) + P^{n+1}|\hat{g}|^p(x)\right) \\
&\quad \le 3^{p-1}\bigl(C_p(n^{p/2-1} + 2^p) + 1\bigr) \\
&\qquad \times\left(\sum_{k\ge 1}\bigl|P^k|\hat{g}|^p(x) - \pi(|\hat{g}|^p)\bigr| + n\pi(|\hat{g}|^p)\right) \\
&\quad \le 6^p C_p n^{p/2}\left(\sum_{k\ge 1}\bigl|P^k|\hat{g}|^p(x) - \pi(|\hat{g}|^p)\bigr| + \pi(|\hat{g}|^p)\right).
\end{aligned}$$

Since $\hat{g} \in \mathcal{L}_{V_0}$, $\pi(|\hat{g}|^p) \le \|\hat{g}\|_{V_0}^p \pi(V_0^p) < \infty$ and by assumption,

$$\sum_{k=0}^{\infty} \big| P^k |\hat{g}|^p(x) - \pi(|\hat{g}|^p) \big| \le \|\hat{g}\|_{V_0}^p C_1 V_1(x).$$

This yields the desired result. □

PROOF OF PROPOSITION 1. When the state space is ν_m-small, it is easily seen that

$$\sum_n \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \le 2\big(1 - (1-\varepsilon)^{1/m}\big)^{-1},$$

and the proof of (9) is a trivial application of Proposition 12. □

The following proposition gives sufficient conditions, based on nested drift conditions, leading to the explicit bounds (29).

PROPOSITION 13. *Let P be a ψ-irreducible and aperiodic transition kernel on a general state space $\mathcal{X}$. Let $C \subseteq D$ be some accessible ν_m-small sets. Assume there exist some Borel functions $f, V : \mathcal{X} \to [1, \infty)$, $f \le V$, some constants $b < \infty$ and $0 < a < 1$ such that $\sup_D V < \infty$ and*

$$PV(x) \le V(x) - f(x) + b\mathbb{1}_C(x),$$
$$f(x) \ge b/(1-a), \qquad x \in D^c.$$

Then P possesses an invariant probability measure π, $\pi(f) < \infty$ and for any probability measure (λ, μ) on $\mathcal{X} \times \mathcal{X}$,

$$\sum_{n=0}^{\infty} |\lambda P^n g - \mu P^n g| \le \|g\|_f \big(\varepsilon^{-1} M_V + a^{-1}\big(\lambda(V) + \mu(V)\big)\big), \tag{30}$$

where

$$\begin{aligned} M_V :=& \sup_{(x,x') \in C \times D} \sum_{k=1}^{m-1} \{P^k f(x) + P^k f(x')\} \\ &+ \sup_{(x,x') \in C \times D} \left(\sum_{k=1}^{m-1} \{P^k f(x) + P^k f(x')\} + a^{-1}\{P^m V(x) + P^m V(x')\} \right) \\ \le& \ 4a^{-1}\left(bm + \sup_D V\right), \end{aligned}$$

with the convention that $\sum_{k=1}^{0} P^k f(x) = 0$.

PROOF. By Theorem 14.0.1 of Meyn and Tweedie (1993), there exists an invariant probability measure π such that $\pi(f) < \infty$. For simplicity, the proof of (30) is restricted to the case $m = 1$. The proof of (30) is based on a coupling technique which may be summarized as follows. Let $\Delta := (C \times D) \cup (D \times C)$ and let R be the residual kernel defined as

$$R(x, \cdot) := \big(1 - \mathbb{1}_D(x)\varepsilon\big)^{-1}\big(P(x, \cdot) - \varepsilon \mathbb{1}_D(x)\nu_1(\cdot)\big).$$

We define a $\mathcal{X} \times \mathcal{X} \times \{0, 1\}$-valued process $Z := \{\Omega, \mathcal{A}, Z_n = (X_n, X'_n, d_n), P_{x,x',d}\}$ such that (a) $P_{x,x',0}(X_n \in \cdot) = P^n(x, \cdot)$ and $P_{x,x',0}(X'_n \in \cdot) = P^n(x', \cdot)$ for all $(x, x') \in \mathcal{X} \times \mathcal{X}$ and (b) there exists a random time T and $X_n \mathbb{1}_{T \le n} = X'_n \mathbb{1}_{T \le n}$. Set $Z_0 := (x, x', 0)$. Each time (X_k, X'_k, d_k) hits the set $\Delta \times \{0\}$, an ε-biased coin is tossed. If the coin comes up heads, then the coupling is successful: the next value of $X_{k+1} = X'_{k+1}$ is simulated from ν_1, $d_{k+1} = 1$, and the two components remain forever coupled. Otherwise, the next values X_{k+1} and X'_{k+1} are drawn independently from the residual kernel R and $d_{k+1} = 0$. If $(X_k, X'_k, d_k) \in \Delta^c \times \{0\}$, then the processes are updated independently from P.

Define the coupling time $T := \inf\{n \ge 1,\ d_n = 1\}$ (with the convention that $\inf \varnothing = \infty$), $T_0 := \inf\{k \ge 0,\ (X_k, X'_k) \in \Delta\}$ and, for $i \ge 1$, $T_i := \inf\{k > T_{i-1}, (X_k, X'_k) \in \Delta\}$ as the successive hitting times on Δ. By definition of T, we have $X_n \mathbb{1}_{T \le n} = X'_n \mathbb{1}_{T \le n}$ and for any Borel function $g \in \mathcal{L}_f$,

$$\begin{aligned} &\sum_{n \ge 0} \int \lambda(dx)\mu(dy)|P^n g(x) - P^n g(y)| \\ &\qquad \le \|g\|_f \ \mathbb{E}_{\lambda,\mu,0}\Bigg[\sum_{n=0}^{T-1}\{f(X_n) + f(X'_n)\}\Bigg]. \end{aligned} \tag{31}$$

Define

$$\begin{aligned} A(f) &:= (1 - \varepsilon) \\ &\quad \times \sup_{(x,x') \in \Delta} \int R(x, dy)R(x', dy')\mathbb{E}_{y,y',0}\Bigg[\sum_{n=0}^{T_0}\{f(X_n) + f(X'_n)\}\Bigg]. \end{aligned} \tag{32}$$

The first step in the proof consists of showing that

$$\begin{aligned} &\mathbb{E}_{x,x',0}\Bigg[\sum_{n=0}^{T_0}\{f(X_n) + f(X'_n)\}\Bigg] \\ &\qquad \le \mathbb{1}_\Delta(x, x')\{f(x) + f(x')\} + a^{-1}\mathbb{1}_{\Delta^c}(x, x')\{V(x) + V(x')\}. \end{aligned} \tag{33}$$

The case $(x, x') \in \Delta$ is trivial. For $(x, x') \in \Delta^c$, under the stated assumptions,

$$\begin{aligned} &\mathbb{E}_{x,x',0}[V(X_1) + V(X'_1)] \\ &\qquad \le V(x) + V(x') - \big(f(x) + f(x')\big) + b\big(\mathbb{1}_C(x) + \mathbb{1}_C(x')\big). \end{aligned}$$

Since $(x, x') \in \Delta^c$, $x \in C$ (resp. $x' \in C$) implies that $x' \in D^c$ (resp. $x \in D^c$), so that

$$f(x') - b\mathbb{1}_C(x) \geq af(x'), \qquad f(x) - b\mathbb{1}_C(x') \geq af(x).$$

Hence,

$$\mathbb{E}_{x,x',0}[V(X_1) + V(X_1')] \leq V(x) + V(x') - a(f(x) + f(x')), \qquad (x, x') \in \Delta^c,$$

and the proof of (33) follows from the so-called Dynkin formula [Meyn and Tweedie (1993), Proposition 11.3.2]. Note that by (33), $\mathbb{E}_{x,x',0}[T_0] < \infty$, which implies that $P_{x,x',0}(T < \infty) = 1$ for all $(x, x') \in \mathcal{X} \times \mathcal{X}$. We now prove that

$$\begin{aligned} &\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T-1}\{f(X_n) + f(X_n')\}\right] \\ &\qquad \leq \mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_0}\{f(X_n) + f(X_n')\}\right] + \varepsilon^{-1}A(f). \end{aligned} \tag{34}$$

By the strong Markov property and by noting that $P_{x,x',0}(d_{T_j} = 0) = (1 - \varepsilon)^j$, for $j \geq 0$,

$$\begin{aligned} &\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_{j+1}}\{f(X_n) + f(X_n')\}\mathbb{1}_{\{0\}}(d_{T_j+1})\right] \\ &\quad = (1-\varepsilon)\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_j}\{f(X_n) + f(X_n')\}\mathbb{1}_{\{0\}}(d_{T_{j-1}+1})\right] \\ &\qquad + \mathbb{E}_{x,x',0}\left[\mathbb{1}_{\{0\}}(d_{T_j+1})\mathbb{E}_{X_{T_j+1},X'_{T_j+1},0}\left[\sum_{n=0}^{T_0}\{f(X_n) + f(X_n')\}\right]\right], \\ &\quad \leq (1-\varepsilon)\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_j}\{f(X_n) + f(X_n')\}\mathbb{1}_{\{0\}}(d_{T_{j-1}+1})\right] + A(f)(1-\varepsilon)^j, \end{aligned}$$

with the convention $T_{-1} + 1 = 0$. By straightforward recursion,

$$\begin{aligned} &\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_{j+1}}\{f(X_n) + f(X_n')\}\mathbb{1}_{\{0\}}(d_{T_j+1})\right] \\ &\qquad \leq (1-\varepsilon)^j\left((1-\varepsilon)\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_0}\{f(X_n) + f(X_n')\}\right] + (j+1)A(f)\right). \end{aligned} \tag{35}$$

Hence,

$$\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T-1}\{f(X_n)+f(X'_n)\}\right]$$
$$=\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_0}\{f(X_n)+f(X'_n)\}\mathbb{1}_{d_{T_0+1}=1}\right]$$
$$+\sum_{j=0}^{\infty}\mathbb{E}_{x,x',0}\left[\sum_{n=0}^{T_{j+1}}\big(f(X_n)+f(X'_n)\big)\mathbb{1}_{\{0\}}(d_{T_j+1})\mathbb{1}_{\{1\}}(d_{T_{j+1}+1})\right]$$

and (34) follows by noting that $P_{X_{T_j},X'_{T_j},0}(d_{T_j+1}=1)=\varepsilon$. The proposition follows from (31)–(35).

The drift condition implies

$$\sup_{(x,x')\in C\times D}\left(\sum_{k=1}^{m-1}\{P^k f(x)+P^k f(x')\}+\{P^m V(x)+P^m V(x')\}\right)$$
$$\le 2bm+\sup_{(x,x')\in C\times D}\{V(x)+V(x')\}$$

from which it is easily seen that $M_V\le 4a^{-1}(bm+\sup_D V)$. $\square$

PROOF OF PROPOSITION 2. The first step is to prove that the level set $D:=\{V\le M\}$ is small. By assumption, $\sup_{x\in D}\mathbb{E}_x[\tau_C]<\infty$. Then for any $\eta>0$, there exists n_0 such that $P_x(\sigma_C\ge n)\le\eta$, $x\in D$ and $n\ge n_0$. Then we can define a distribution $\alpha=\{\alpha(n)\}$ on $\mathbb{Z}_+$ such that for $x\in D$ and $0<l<1$, $\sum_n\alpha(n)P^n(x,C)\ge\sum_{n\le n_0}\alpha(n)P^n(x,C)\ge l(1-\eta)$. Whereas C is petite, there exist some measure ν on $\mathcal{X}$ and some distribution $\beta=\{\beta(n)\}$ on $\mathbb{Z}_+$ such that $\sum_n\alpha*\beta(n)P^n(x,A)\ge l(1-\eta)\nu(A)$, which proves that D is petite. The smallness property of D deduces from Theorem 5.5.7. of Meyn and Tweedie (1993). Note in addition that, by definition, $D\supseteq C$. Define

$$\begin{aligned}
f_0&:=V^{1/p}, & V_0&:=V^{1/p}/(1-\rho^{1/p}),\\
f_1&:=V_0^p, & V_1&:=V/\{(1-\rho)(1-\rho^{1/p})^p\},\\
b_0&:=b^{1/p}/(1-\rho^{1/p}), & a_0&:=1-b^{1/p}/\{(1-\rho^{1/p})M^{1/p}\},\\
b_1&:=b/\{(1-\rho)(1-\rho^{1/p})^p\}, & a_1&:=1-b/\{(1-\rho)M\}.
\end{aligned}$$

It is easily seen that $PV_i\le V_i-f_i+b_i\mathbb{1}_C$, $i=0,1$, $1\le f_0\le V_0\le V_0^p=f_1\le V_1$, $0<a_i<1$ and $f_i\ge b_i/(1-a_i)$ on D^c, $i=0,1$. By applying Proposition 13, the inequalities (29) are verified and the constants C_i, $i = 0, 1$, are upper bounded by (this upper bound is not optimal)

$$C_0 \le 5\varepsilon^{-1}(m+1)M^{1/p}(1-\rho^{1/p})\big((1-\rho^{1/p})-(b/M)^{1/p}\big)^{-1},$$

$$C_1 V_1(x) + \pi(V_0^p) \le 5\varepsilon^{-1}(m+1)M(1-\rho^{1/p})^{-p}(1-\rho-b/M)^{-1}V(x).$$

This yields the desired result. □

7. Proofs of Lemmas 7 and 14.

7.1. *Proof of Lemma* 14.

LEMMA 14. *Under the assumptions of Theorem* 6, *we have*

$$\rho_n \mathbb{1}_{\{\lim_n \tilde{S}_n = s^*\}} = o_{\text{w.p.1}}(m_n^{-1/2}). \tag{36}$$

PROOF. The remainder term ρ_n also follows a difference equation of the form

$$\rho_n = \Gamma\rho_{n-1} + \eta_n = (H_{n-1}+\Gamma)\rho_{n-1} + r_{n-1} + \eta_n^{(1)}$$

since $\eta_n^{(2)}$ may be decomposed as $\eta_n^{(2)} = H_{n-1}\rho_{n-1} + r_{n-1}$ with $H_n := \sum_{1\le i\le q} R_n(i,\cdot)(2\mu_{n,i} + \rho_{n,i})$, and $r_n := \sum_{1\le i,j\le q} R_n(i,j)\mu_{n,i}\mu_{n,j}$ for $n \ge 0$. Hence we have $\rho_n := \rho_n^{(1)} + \rho_n^{(2)}$, where

$$\rho_n^{(1)} := \prod_{k=0}^{n-1}(H_k+\Gamma)\rho_0 + \sum_{k=1}^{n}\left(\prod_{j=k}^{n-1}(H_j+\Gamma)\right)\eta_k^{(1)},$$

$$\rho_n^{(2)} := \sum_{k=0}^{n-1}\left(\prod_{j=k+1}^{n-1}(H_j+\Gamma)\right)r_k.$$

As $\mu_n = O_{L^p}(m_n^{-1/2})$ and, by assumption, $\sum_n m_n^{-p/2} < \infty$, then $\mu_n = o_{\text{w.p.1}}(1)$ and thus, $\rho_n \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = o_{\text{w.p.1}}(1)$. Hence, $|H_n|\mathbb{1}_{\lim_n \tilde{S}_n = s^*} = o_{\text{w.p.1}}(1)$ and for any $\gamma < \tilde{\gamma} < 1$, $j \le n$, $|\prod_{k=j}^{n}(H_j+\Gamma)|\mathbb{1}_{\lim_n \tilde{S}_n = s^*} = O_{\text{w.p.1}}(\tilde{\gamma}^n)$. Along trajectories converging to s^*, the first term in $\rho_n^{(1)}$ is $O_{\text{w.p.1}}(1)O_{L^p}(\tilde{\gamma}^n)$ since, by M4, $\rho_0 \in L^p$. The first term in $\eta_n^{(1)}$ is only finitely often nonzero, and by M4, the second term in $\eta_n^{(1)}$ is bounded and the bound is inversely proportional to m_n. Thus, by choosing $\tilde{\gamma}^{-1} > \lim_n m_{n+1}/m_n$ and by applying Lemma 5,

$$\rho_n^{(1)}\mathbb{1}_{\lim_n \tilde{S}_n = s^*} = O_{\text{w.p.1}}(1)O_{L^p}(m_n^{-1}). \tag{37}$$

Similarly, as $r_n = O_{L^p}(m_n^{-1})$,

$$\rho_n^{(2)} \mathbb{1}_{\lim_n \tilde{S}_n = s^*} = O_{\text{w.p.1}}(1) O_{L^p}(m_n^{-1}) \tag{38}$$

and the proof of (36) is complete. □

7.2. *Proof of Lemma* 7.

LEMMA 15. *Let* $\{a_n\}$ *and* $\{b_n\}$, $b_n \neq 0$, *be two sequences such that*:

(i) *the power series* $f(x) := \sum_{n=1}^{\infty} a_n x^n$ *has a radius of convergence* r *and*

(ii) $\lim_{n\to\infty} b_{n+1}/b_n =: q$ *with* $|q| < r$. *Define* $c_n := \sum_{k\geq n} b_k a_{k-n}$. *Then* $\lim_{n\to\infty} c_n b_n^{-1} = f(q)$.

PROOF. By assumption, for any K and $\varepsilon > 0$, there exists N such that for all $n \geq N$, $|b_{n+K}/b_n - q^K| \leq \varepsilon$. In addition, there exist some positive constants A, ε such that for all $n, j \geq 0$, $b_{n+j}/b_n \leq A(q+\varepsilon)^j$,

$$\begin{aligned}&\left| b_n^{-1} \sum_{k\geq n} b_k a_{k-n} - \sum_{k\geq 0} q^k a_k \right| \\ &\qquad \leq \sum_{k=n}^{n+K} |b_k/b_n - q^{k-n}| a_{k-n} + \sum_{k\geq n+K} b_k/b_n a_{k-n} + \sum_{k\geq K} q^k a_k.\end{aligned}$$

Let $\varepsilon > 0$. Then there exists K such that the last two sums are upper bounded by ε. Now for those constants K, ε there exists N such that for $n \geq N$, the first sum is less than ε and the proof is completed. □

We now prove Lemma 7. We shall establish that for $m\gamma \neq 1$,

$$\begin{aligned}&(1-m\gamma)^r \left(\lim_n \xi_n^{(r)}\right)^r \\ &\qquad = 1 + m^{r/2} \sum_{l=0}^{r-1} \binom{r}{l} (-1)^{r-l} (m^{l-r/2}\gamma^{l-r} - 1)^{-1} \lim_n m_n^{r/2} \left(\sum_{k=0}^{n} m_k^{r/2}\right)^{-1}.\end{aligned} \tag{39}$$

If $m > 1$, then Lemma 5 implies that $\lim_n m_n^{r/2} (\sum_{k=0}^n m_k^{r/2})^{-1} = 1 - m^{-r/2}$. If $m = 1$, then $\lim_n m_n^{r/2} (\sum_{k=0}^n m_k^{r/2})^{-1} = 0$. In both cases, $\lim_n m_n^{r/2} \times (\sum_{k=0}^n m_k^{r/2})^{-1} = 1 - m^{-r/2}$. Thus Lemma 7 holds provided that (39) is established.

First case. $m\gamma < 1$. Define $S_n := \sum_{j\geq n} m_j \gamma^j$. Hence $S_n = \gamma^n \sum_{j\geq n} m_j \gamma^{j-n}$ and by applying Lemma 15, since $m < \gamma^{-1}$, it holds that

$$\lim_n m_n^{-1} \gamma^{-n} S_n = (1-m\gamma)^{-1}. \tag{40}$$

We write

$$\sum_{k=0}^{n} m_k^{-r/2}\left(\sum_{j=0}^{n-k} m_{j+k}\gamma^j\right)^r$$

$$= \sum_{k=0}^{n} m_k^{-r/2}\gamma^{-kr}(S_k - S_{n+1})^r$$

$$= \sum_{k=0}^{n} m_k^{-r/2}\gamma^{-kr}S_k^r + \sum_{l=0}^{r-1}\binom{r}{l}(-1)^{r-l}S_{n+1}^{r-l}\sum_{k=0}^{n} m_k^{-r/2}\gamma^{-kr}S_k^l$$

$$= \sum_{k=0}^{n} m_k^{r/2}(m_k^{-1}\gamma^{-k}S_k)^r$$

$$+ m_n^{r/2}\sum_{l=0}^{r-1}\binom{r}{l}(-1)^{r-l}\gamma^{r-l}(m_{n+1}/m_n)^{r-l}(m_{n+1}^{-1}\gamma^{-(n+1)}S_{n+1})^{r-l}\cdots$$

$$\times\left(m_n^{r/2-l}\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{(n-k)(r-l)}\right)\left(\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{k(l-r)}\right)^{-1}$$

$$\times\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{k(l-r)}(m_k^{-1}\gamma^{-k}S_k)^l.$$

By use of the Cesaro lemma and (40),

$$\lim_n\left(\sum_{k=0}^{n} m_k^{r/2}\right)^{-1}\sum_{k=0}^{n} m_k^{r/2}(m_k^{-1}\gamma^{-k}S_k)^r = (1-m\gamma)^{-r}.$$

In addition, for all $l\in\{0,\dots,r-1\}$, $(m\gamma)^{l-r}m^{r/2}>1$, showing that $\sum_{k=0}^{n} m_k^{l-r/2}\times\gamma^{k(l-r)}$ diverges to infinity. Then applying again the Cesaro lemma and (40),

$$\lim_n\left(\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{k(l-r)}\right)^{-1}\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{k(l-r)}(m_k^{-1}\gamma^{-k}S_k)^l = (1-m\gamma)^{-l}.$$

Finally, whereas $l<r$, $(m\gamma)^{r-l}m^{-r/2}<1$ and Lemma 5 implies that

$$\lim_n m_n^{r/2-l}\sum_{k=0}^{n} m_k^{l-r/2}\gamma^{(n-k)(r-l)} = (\gamma m)^{l-r}m^{r/2}(\gamma^{l-r}m^{l-r/2}-1)^{-1}.$$

Combining these limits gives (39).

Second case. $m\gamma>1$. Define $S_n := \sum_{j=0}^{n} m_j\gamma^j$. Hence $S_n=\gamma^n\sum_{j=0}^{n} m_j\times\gamma^{-(n-j)}$ and by applying Lemma 5, since $m^{-1}<\gamma$, it holds that

$$\lim_n m_n^{-1}\gamma^{-n}S_n = m\gamma(m\gamma-1)^{-1}. \tag{41}$$

We write, with the convention $S_{-1} := 0$,

$$\sum_{k=0}^{n} m_k^{-r/2} \left(\sum_{j=0}^{n-k} m_{j+k} \gamma^j \right)^r$$
$$= \sum_{k=0}^{n} m_k^{-r/2} \gamma^{-kr} (S_n - S_{k-1})^r$$
$$= (-1)^r \sum_{k=0}^{n} m_k^{-r/2} \gamma^{-kr} S_{k-1}^r + \sum_{l=0}^{r-1} \binom{r}{l} (-1)^l S_n^{r-l} \sum_{k=0}^{n} m_k^{-r/2} \gamma^{-kr} S_{k-1}^l$$
$$= (-1)^r \gamma^{-r} \sum_{k=0}^{n} m_k^{r/2} (m_{k-1}/m_k)^r (m_{k-1}^{-1} \gamma^{-(k-1)} S_{k-1})^r$$
$$+ m_n^{r/2} \sum_{l=0}^{r-1} \binom{r}{l} (-1)^l \gamma^{-l} (m_n^{-1} \gamma^{-n} S_n)^{r-l} m_n^{r/2-l}$$
$$\times \left(\sum_{k=0}^{n} m_k^{l-r/2} \gamma^{(r-l)(n-k)} \right) \left(\sum_{k=0}^{n} m_k^{l-r/2} \gamma^{k(l-r)} \right)^{-1}$$
$$\times \sum_{k=0}^{n} m_k^{l-r/2} \gamma^{k(l-r)} (m_k/m_{k-1})^{-l} (m_{k-1}^{-1} \gamma^{-(k-1)} S_{k-1})^l.$$

By use of the Cesaro Lemma and (41),

$$(-1)^r \gamma^{-r} \lim_n \left(\sum_{k=0}^{n} m_k^{r/2} \right)^{-1} \sum_{k=0}^{n} m_k^{r/2} (m_{k-1}/m_k)^r (m_{k-1}^{-1} \gamma^{-(k-1)} S_{k-1})^r$$
$$= (1 - m\gamma)^{-r}.$$

In addition, for all $l \in \{0, \dots, r-1\}$, $(m\gamma)^l (m\gamma^2)^{-r/2} > 1$, showing that $\sum_{k=0}^{n} m_k^{l-r/2} \gamma^{k(l-r)}$ diverges to infinity. Then, applying again the Cesaro lemma and (41),

$$\lim_n \left(\sum_{k=0}^{n} m_k^{l-r/2} \gamma^{k(l-r)} \right)^{-1} \sum_{k=0}^{n} m_k^{l-r/2} \gamma^{k(l-r)} (m_k/m_{k-1})^{-l} (m_{k-1}^{-1} \gamma^{-(k-1)} S_{k-1})^l$$
$$= \gamma^l (m\gamma - 1)^{-l}.$$

Finally, whereas $(m\gamma)^l (m\gamma^2)^{-r/2} > 1$, Lemma 5 implies that

$$\lim_n m_n^{r/2-l} \sum_{k=0}^{n} m_k^{l-r/2} \gamma^{(n-k)(r-l)} = (\gamma m)^{l-r} m^{r/2} (\gamma^{l-r} m^{l-r/2} - 1)^{-1}.$$

Combining these limits gives (39).

Acknowledgment. The authors would like to thank the anonymous referee for helpful comments.

REFERENCES

BISCARAT, J. (1994). Almost sure convergence of a class of stochastic algorithms. *Stochastic Process. Appl.* **50** 83–94.

BOOTH, J. and HOBERT, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 265–285.

BRANDIÈRE, O. (1998). The dynamic system method and the traps. *Adv. in Appl. Probab.* **30** 137–151.

BRÖCKER, T. (1975). *Differentiable Germs and Catastrophes*. Cambridge Univ. Press.

CAPPÉ, O., DOUCET, A., LAVIELLE, M. and MOULINES, E. (1999). Simulation-based methods for blind maximum-likelihood filter identification. *Signal Processing* **73** 3–25.

CELEUX, G. and DIEBOLT, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stochastics Stochastics Rep.* **41** 119–134.

CHAN, J. and KUK, A. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53** 86–97.

CHAN, K. and LEDOLTER, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.* **90** 242–252.

CHEN, H., GUO, L. and GAO, A. (1988). Convergence and robustness of the Robbins–Monro algorithm truncated at randomly varying bounds. *Stochastic Process. Appl.* **27** 217–231.

DALAL, S. and WEERAHANDI, S. (1992). Some approximations for the moments of a process used in diffusion of new products. *Statist. Probab. Lett.* **15** 181–189.

DALAL, S. and WEERAHANDI, S. (1995). Estimation of innovation diffusion models with application to a consumer durable. *Marketing Letters* **6** 123–136.

DELYON, B., LAVIELLE, M. and MOULINES, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.* **27** 94–128.

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

FORT, G., MOULINES, E., ROBERTS, G. O. and ROSENTHAL, J. S. (2003). On the geometric ergodicity of hybrid samplers. *J. Appl. Probab.* **40** 123–146.

GUO, S. and THOMPSON, E. (1991). Monte-Carlo estimation of variance component models for large complex pedigrees. *IMA Journal of Mathematics Applied in Medicine and Biology* **8** 171–189.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.

JAMSHIDIAN, M. and JENNRICH, R. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **59** 569–587.

MENG, X. and SCHILLING, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* **91** 1254–1267.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer, London.

PIERRE-LOTI-VIAUD, D. (1995). Random perturbations of recursive sequences with an application to an epidemic model. *J. Appl. Probab.* **32** 559–578.

PÓLYA, G. and SZEGÖ, G. (1976). *Problems and Theorems in Analysis* **2**. Springer, New York.

POLYAK, B. (1990). New stochastic approximation type procedures. *Automat. Remote Control* **51** 937–946.

SHAPIRO, A. and WARDI, Y. (1996). Convergence analysis of stochastic algorithms. *Math. Oper. Res.* **21** 615–628.

SHERMAN, R., HO, Y. and DALAL, S. (1999). Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. *Econom. J.* **2** 248–267.

TANNER, M. A. (1996). *Tools for Statistical Inference*, 3rd ed. Springer, New York.

WEI, G. and TANNER, M. (1990). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

WU, C. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11** 95–103.

ZEGER, S. L. (1988). A regression model for time series of counts. *Biometrika* **75** 621–629.

LMC-IMAG
51, RUE DES MATHÉMATIQUES
BP 53, 38041 GRENOBLE CEDEX 9
FRANCE
E-MAIL: Gersende.Fort@imag.fr

ENST
46, RUE BARRAULT
75634 PARIS CEDEX 13
FRANCE
E-MAIL: moulines@tsi.enst.fr