Statistical Science
1995, Vol. 10, No. 3, 231–253

Convergence Control Methods for Markov Chain Monte Carlo Algorithms

Christian P. Robert

Abstract. Markov chain Monte Carlo methods have been increasingly popular since their introduction by Gelfand and Smith. However, while the breadth and variety of Markov chain Monte Carlo applications are properly astounding, progress in the control of convergence for these algorithms has been slow, despite its relevance in practical implementations. We present here different approaches toward this goal based on functional and mixing theories, while paying particular attention to the central limit theorem and to the approximation of the limiting variance. Renewal theory in the spirit of Mykland, Tierney and Yu is presented as the most promising technique in this regard, and we illustrate its potential in several examples. In addition, we stress that many strong convergence properties can be derived from the study of simple subchains which are produced by Markov chain Monte Carlo algorithms, due to a *duality principle* obtained in Diebolt and Robert for mixture estimation. We show here the generality of this principle which applies, for instance, to most missing data models. A more empirical stopping rule for Markov chain Monte Carlo algorithms is related to the simultaneous convergence of different estimators of the quantity of interest. Besides the regular ergodic average, we propose the Rao–Blackwellized version as well as estimates based on importance sampling and trapezoidal approximations of the integrals.

Key words and phrases: Gibbs sampling, Metropolis algorithm, central limit theorem, asymptotic variance, renewal theory, duality principle, finite state Markov chains, missing data, ergodic theorem, Rao–Blackwellization, importance sampling, trapezoidal integration.

1. INTRODUCTION

Since the publication by Tanner and Wong (1987) and Gelfand and Smith (1990) of two seminal papers promoting the use of Markov chain Monte Carlo (MCMC) methods in statistical setups, there has been considerable interest in these simulation methods, which were known to physicists since Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953). While the literature on this topic has tremendously increased, as shown by the wide-ranging discussions in Gelman and Rubin (1992), Geyer (1992), Besag and Green (1993), Gilks et al. (1993), Smith and Roberts (1993), and Besag, Green, Higdon and Mengersen (1995), theoretical approaches to convergence control for these methods are comparatively quite limited.

Indeed, MCMC methods allow for the statistical treatment of models previously considered intractable, such as, for instance, mixture and classification models (Diebolt and Robert, 1994). Although a direct application of MCMC algorithms is most often justified, it is still necessary to assess the validity of these stochastic techniques—such as, for example, by checking that the posterior distributions are truly defined—and, more importantly, to guarantee the convergence of the proposed estimators to the true limit. The theoretical foundations for the applicability of MCMC techniques have been well charted in Tierney (1991, 1994), Geyer (1992), Schervish and Carlin (1992), Smith and Roberts (1993), Gelfand and Sahu (1994), Liu, Wong and Kong (1994, 1995), Chib and Greenberg

Professor Christian Robert is currently Head of the Statistics Laboratory at CREST-INSEE, on leave from the Mathematics Department, University of Rouen, France.

(1995), Mykland, Tierney and Yu (1995) and Polson (1995), among others. They state that irreducibility and the existence of a posterior probability distribution corresponding to the conditional distributions used in the algorithm are basically sufficient to ensure ergodicity of the Markov chain thus produced.

The issue addressed by this paper is another step in the theoretical study of MCMC methods. We propose several convergence criteria based on the properties of the simulated Markov chain in terms of speed of convergence and of accuracy of the approximation. This question has already been considered in the literature, either through theoretical probability and functional analysis or through more empirical measures, such as those used in engineering simulation. See, for instance, the reviews by Brooks and Roberts (1995) and Cowles and Carlin (1995) for the most practical tools. However, it is necessary to reconsider these approaches because, while probability theory provides tools quite helpful in the study of convergence for MCMC methods, it does not usually focus on the question of interest (to us).

Indeed, we are yet again faced with an inverse perspective quite common in statistics: the existence of a stationary measure or the ergodicity of the chain under study is rarely the problem in MCMC setups, while the global properties of the distribution of the chain at step n are not directly relevant for the study of the single *string* (or path) produced by an MCMC algorithm. From a probabilistic perspective, the study of Markov chains is usually concerned with the behavior of the continuum of possible chains produced by a Markov kernel, in the same way iid sampling probability theory describes the average behavior of a sample $x_1, \ldots, x_n$ generated from a distribution f. However, in pseudo-random generation as well as in statistics, we have to use a single chain/sample to derive convergence/accuracy properties.

A *lieu commun* in this area is that convergence of a Markov chain to its stationary distribution and the corresponding convergence of the empirical moments to the moments of this distribution can never be truly ascertained. In fact, it is always possible that the stationary distribution has such widely separated modes that jumps between these modes occur quite rarely, and a monitoring of the chain strongly hints at stationarity, although this chain is only exploring the neighborhood of a single mode. Gelman and Rubin (1992) and Besag et al. (1995), among others, have illustrated such cases. Nonetheless, it is necessary to produce indicators of convergence, which, although they are only partially adequate, help to control MCMC methods more rigorously than presently.

This is why, after a short review of the principal convergence results relevant for MCMC methods, Section 2 focuses on the central limit theorem and its assessment. This result is of major interest for convergence issues, because it provides a control for the Markov chain. We also note that the law of the iterated logarithm is often of marginal use in this framework. In particular, we discuss in Section 2.3 the relation between the central limit theorem and the mixing properties of the chain. Section 3 dwells on renewal theory to improve the range of applicability of the central limit theorem as well as the estimation of the asymptotic variance, while indicating through examples why this approach is also delicate to implement in practice, except in the important case of finite state spaces (Section 3.4). Section 4 reconsiders the previous methods in the light of the duality principle of Diebolt and Robert (1994), which extends the convergence properties of the simpler subchains involved in the MCMC process to the other components. Section 5 proposes convergence assessments of a more empirical and graphical nature; they rely on several estimates based on the Markov chain Monte Carlo sample and their monitoring until coincidence. In particular, we propose a benchmark estimate related to trapezoidal approximations of a given integral, derived from Yakowitz, Krimmel and Szidarovszky (1978).

2. CENTRAL LIMIT THEOREM AND MIXING PROPERTIES

2.1 Convergence Criteria

As stressed in the Introduction, we must try to separate as much as possible *functional* properties of the Markov chain and asymptotic properties of the sample $(x_1, \ldots, x_n)$ at hand. The first type of property is described through the transition probability of the Markov chain or the distribution of the chain at step n, both quantities being usually intractable. On the contrary, the properties of the chain provided by an MCMC device can be directly used for assessing convergence.

Consider, thus, a Markov chain (x_n) associated with a transition probability density $k(x_n|x_{n-1})$ which is usually unavailable in closed form. For instance, the transition kernel k for data augmentation (Tanner and Wong, 1987) can be decomposed as

$$k(x_n|x_{n-1}) = \int_{\mathscr{Y}} f_{X|Y}(x_n|y) f_{Y|X}(y|x_{n-1})\, dy; \tag{2.1}$$

for Gibbs sampling (Gelfand and Smith, 1990), it is

$$k(x_n^1, \ldots, x_n^p | x_{n-1}^1, \ldots, x_{n-1}^p) = f_1(x_n^1 | x_{n-1}^2, \ldots, x_{n-1}^p) \cdots f_n(x_n^p | x_n^1, \ldots, x_n^{p-1}). \tag{2.2}$$

In all cases of interest, the chain is known to have a stationary probability distribution $\tilde{\pi}$. This eliminates the irrelevant cases where the joint posterior distribution does not exist, although the conditional posterior distributions exist (see Casella and George, 1992, or Hobert and Casella, 1996), but, more importantly, it guarantees that the chain is ergodic when it is irreducible and thus that the (marginal) distribution of x_n, $\pi^n(\cdot|x_0)$, converges to the distribution $\tilde{\pi}(\cdot)$ in the sense of the total variation norm for almost every x_0:

$$||\pi^n(\cdot|x_0) - \tilde{\pi}||_{\text{TV}} = \sup_{A \in \mathscr{A}} |\pi^n(A|x_0) - \tilde{\pi}(A)| \to 0 \qquad \text{as } n \to \infty$$

(see Tierney, 1994, for definitions and discussions of different types of convergence). However, unless one uses the rather special setup of Tanner and Wong (1987), where the number of iid simulations from $\pi^n(\cdot|x_0)$ increases with n, the distribution $\pi^n(\cdot|x_0)$ is only sampled once by the MCMC algorithm and the corresponding simulation can be described more accurately by a simulation from $\pi(\cdot|x_{n-1})$ for conditioning reasons. This implies that precise convergence results like those of Rosenthal (1993) about the speed of the Metropolis algorithm in finite state spaces have limited applicability.

Ergodicity also has a corollary which has more immediate consequences in MCMC setups. In fact, under ergodicity, a law of large numbers often called the *ergodic theorem* applies, since the average

$$\frac{1}{N} \sum_{n=1}^{N} h(x_n) \tag{2.3}$$

converges to the theoretical mean $\mathbb{E}^{\tilde{\pi}}[h(x)]$ when $h \in \mathscr{L}_1(\tilde{\pi})$ and (2.3) provides a practical access to the characterization of the stationary probability measure. Gelfand and Smith (1990), Liu, Wong and Kong (1994, 1995) and Casella and Robert (1996) proposed a modification of (2.3) based on the Rao–Blackwell theorem which is shown to improve upon (2.3) in special setups. We stress the relevance of Rao–Blackwellization for convergence diagnoses in Sections 4 and 5.

Some additional functional properties of this chain also allow for a more precise description of the convergence properties of the chain, although they are often difficult to assess in MCMC setups. For instance, there may exist a constant $0 < \rho < 1$ such that convergence occurs at speed ρ^n, that is, such that there exists a constant C with

$$||\pi^n(\cdot|x_0) - \tilde{\pi}(\cdot)||_{\text{TV}} \leq C\rho^n.$$

Convergence to the stationary distribution is then said to be *geometric* and guarantees a similar speed for the convergence of the expectations, in the sense that, for every $h \in L_1(\tilde{\pi})$, there exists C_h such that

$$||\mathbb{E}^{\pi^n}[h(x)|x_0] - \mathbb{E}^{\tilde{\pi}}[h(x)]|| \leq C_h \rho^n. \tag{2.4}$$

Meyn and Tweedie (1994) derived some simple approximations for the rate ρ, based on the existence of a potential function V and of a small set R quite similar to the renewal set introduced in Section 3. Mengersen and Tweedie (1996) took advantage of these bounds to derive explicit rates in some particular Hastings–Metropolis setups; in particular, they showed that independent Hastings algorithms with large tails and random walk symmetric Metropolis algorithms are incompatible with geometric ergodicity.

However, note that, apart from the particular case of finite state chains, where ergodicity is equivalent to geometric ergodicity, this property can be difficult to assess. In addition, it is a property of the nth step distribution $\pi^n(\cdot|x_0)$ rather than of the chain at hand (x_n). It is therefore difficult to envision this characteristic as initiating a stopping rule or another convergence diagnosis, although it indicates the approximate speed of convergence. Nonetheless, Roberts and Tweedie (1994) and Gelman, Gilks and Roberts (1994) obtained some applications of geometric ergodicity for the proper acceptance rate of Metropolis algorithms.

Schervish and Carlin (1992) constructed some advanced theory on the Markov kernels as linear operators on the space of $\tilde{\pi}$-integrable functions and derived geometric convergence results under the Hilbert–Schmidt condition

$$\int k^2(x_n|x_{n-1})/(\tilde{\pi}(x_n)\tilde{\pi}(x_{n-1}))\, dx_n\, dx_{n-1} < \infty,$$

which is unfortunately too difficult to check in most cases. (See Liu, Wong and Kong, 1995, for related results.)

2.2 Central Limit Theorem

This result is of more direct interest for MCMC algorithms, since it characterizes the convergence of the average (2.3) the following way: when the central limit theorem applies, for every $h \in L_2(\tilde{\pi})$, there exists $0 < \sigma_h < +\infty$ such that

$$\frac{1}{\sqrt{N}} \sum_{n=1}^{N} (h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)]) \xrightarrow{\mathscr{L}} \mathscr{N}(0, \sigma_h^2). \tag{2.5}$$

Therefore, under this result, the variations of the average (2.3) around the limit $\mathbb{E}^{\tilde{\pi}}[h(x)]$ are asymptotically normal. If (2.5) holds, stability rates and convergence setups can be derived, provided a correct estimation of σ_h is available. First, as already mentioned in Geyer (1992) and shown in Kipnis and Varadhan (1986), the central limit theorem holds under irreducibility and reversibility of the Markov chain when

$$0 < \sigma_h^2 = \operatorname{var}(h(x_n)) + 2\sum_{t>0} \operatorname{cov}(h(x_n), h(x_{n+t})) < +\infty \tag{2.6}$$

with $x_n \sim \tilde{\pi}$. While irreducibility and reversibility are usually easy to assess in Markov chain Monte Carlo setups, the verification that $0 < \sigma_h^2 < \infty$ and the subsequent estimation of σ_h^2 are quite delicate steps. This issue is considered in Geweke (1992) and Geyer (1992), but treatments there are rather sketchy in our opinion. We relate mixing properties and finiteness of the asymptotic variance in Section 2.3, postponing the estimation of σ_h until Section 3.

Note that when the central limit theorem applies, the law of the iterated logarithm also holds, namely,

$$\limsup_{N\to\infty} \frac{\sum_{n=1}^N (h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)])}{\sqrt{2N\log\log(N)}} = \sigma_h$$

and

$$\liminf_{N\to\infty} \frac{\sum_{n=1}^N (h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)])}{\sqrt{2N\log\log(N)}} = -\sigma_h.$$

In other words, the sequence

$$\frac{\sum_{n=1}^N (h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)])}{\sqrt{2N\log\log(N)}} \tag{2.7}$$

reaches both extremities of $(-\sigma_h, \sigma_h)$. This property may be of interest in the setup of MCMC experiments since (a) simultaneous stabilization of the two ratios to the same value (in absolute value) is an additional indicator of the stationary regime and (b) it provides an alternative approach for estimating σ_h^2 when the asymptotic variance is finite. However, concerning (a), the criterion may be too conservative to be of practical use. For instance, Figure 1(b) shows the evolution of (2.7) in the case of iid observations from a $\mathcal{N}(0, 1)$ distribution, where the limit sup is 1, but is not attained after 1,000,000 iterations. For the same reason, the estimation of σ_h from (2.7) may require too many iterations to be effective in most setups. Moreover, since $\mathbb{E}^{\tilde{\pi}}[h(x)]$ is unknown, (2.7) should be replaced by

$$\frac{\sum_{n=1}^N \{h(x_n) - h(y_n)\}}{2\sqrt{2N\log\log(N)}},$$

where (y_n) is a sequence independent from (x_n), and the influence of the starting points x_0, y_0 may slow down convergence even more. [Note the similarity of (2.7) with Yu and Mykland's (1994) convergence criterion.]

Suppose the central limit theorem applies and an estimator $\hat{\sigma}_h$ of σ_h is available. Then, for N large enough,

$$z_N = \frac{1}{\sqrt{N}} \sum_{n=1}^N h(x_n)$$

should approximately behave like a normal $\mathcal{N}(\mathbb{E}^{\tilde{\pi}}[h(x)], \sigma_h^2/N)$ random variable. A first possible application of this property is to run T parallel independent chains (x_n^t) until the corresponding z_N^t are "sufficiently" normal. (For instance, one could impose that 95% of the z_N^t's are within $2\hat{\sigma}_h/\sqrt{N}$ of the overall mean.) This naive solution is rather costly, however, while being conservative since it requires most chains to converge before the stopping rule works. In addition, it does not take into account the side effects of initial values and of parallel runs, which are much criticized (see Geyer, 1992, and Brooks and Roberts, 1995, for instance). More advanced tools of probability, such as Berry–Essèen bounds (Feller, 1971) or large deviations (Malinovskii, 1987), are presumably necessary for a strict control of convergence via the central limit theorem, but these methods unfortunately involve the estimation of quantities similar to σ_h. We will see in Section 3 how renewal theory makes much more efficient use of the central limit theorem.

2.3 Mixing Properties

A study of the mixing properties of the Markov chain, that is, of the long term correlations between the x_n's, indicates how far from an iid sample (x_n) is while providing more explicit conditions on the function h for (2.5) to hold. In fact, some of these conditions can be checked in practice and this is why they are presented below.

(a) α-MIXING. This property is defined by the convergence of

$$\alpha(n) = \sup_{A,B} |P(x_n \in A,\ x_0 \in B) - \tilde{\pi}(x_n \in A)\tilde{\pi}(x_0 \in B)|$$

to 0 when n goes to infinity and $x_0 \sim \tilde{\pi}$. It defines a rather weak measure of asymptotic independence. Nonetheless, it may provide a basis for the applicability of the central limit theorem. As shown in Davydov (1973), if h is a measurable function such that $\mathbb{E}^{\tilde{\pi}}[|h(x)|^\gamma] < +\infty$ with $\gamma > 2$, a sufficient

condition for

$$S_n = \sum_{i=1}^{n} h(x_i)$$

to be asymptotically normal is that

$$\sum_n \alpha(n)^{(\gamma-2)/\gamma} < +\infty$$

(see also Doukhan, Massart and Rio, 1994, for similar conditions). A more amenable requirement is that $h \in L_2(\pi)$ satisfies

$$\limsup_n \sigma_n / \mathbb{E}^{\tilde{\pi}}[|S_n|] < \sqrt{\pi/2},$$

where σ_n^2 is the variance of S_n (Peligrad, 1986). Since the law of large numbers *theoretically* provides convergent approximations of σ_n and of $\mathbb{E}^{\tilde{\pi}}[|S_n|]$, it may be argued that this condition can be checked in practice. However, the *practical* estimation of σ_n is still an open question for most MCMC algorithms.

In our setup, α-mixing holds in great generality for the Markov chains induced by these methods, since every positive recurrent aperiodic Markov chain is α-mixing (Rosenblatt, 1971).

(b) β-MIXING. A stronger property than α-mixing, β-mixing is defined through the coefficient

$$\beta(n) = \sup_{(A_i)} \sup_{(B_j)} \sum_{i,j} \left| P(x_n \in A_i,\ x_0 \in B_j) - \tilde{\pi}(x_n \in A_i)\tilde{\pi}(x_0 \in B_j) \right|,$$

where the supremum is taken over all pairs of partitions (A_i) and (B_j) and $x_0 \sim \tilde{\pi}$. Under β-mixing, $\beta(n)$ converges to 0 as n goes to infinity. The β-mixing coefficient can also be written as (Davydov, 1973)

$$\beta(n) = \int ||\pi^n(\cdot|x_0) - \tilde{\pi}||_{\mathrm{TV}} \tilde{\pi}(x_0)\, dx_0,$$

although this does not really simplify the assessment of β-mixing or the computation of the coefficient $\beta(n)$. While a stronger property than α-mixing, β-mixing is not sufficient for the central limit theorem to apply, as shown by the following example.

EXAMPLE 2.1. Consider a density g and a function $0 < \rho < 1$ such that

$$\int \rho^{-1}(x) g(x)\, dx < \infty.$$

The transition

$$(2.8)\quad x_{n+1} = \begin{cases} x_n, & \text{with probability } 1 - \rho(x_n), \\ y \sim g, & \text{with probability } \rho(x_n) \end{cases}$$

is akin to the transition of the Metropolis algorithm and leads to the stationary distribution defined by

$$\tilde{\pi}(x) = \frac{\rho^{-1}(x) g(x)}{\int \rho^{-1}(u) g(u)\, du}.$$

For instance, in the particular case when g is a $\mathcal{B}e(\alpha + 1, 1)$ density ($\alpha < 1$) and $\rho(x) = x$, the Markov chain generated from (2.8) converges to a $\mathcal{B}e(\alpha, 1)$ distribution. Moreover, it follows from Doukhan, Massart and Rio (1994) that the chain (x_n) is β-mixing but that the deduced chain $(x_n^{1-\alpha})$ does not satisfy the central limit theorem and that

$$\limsup_{N\to\infty} \frac{\sum_{n=1}^{N} (x_n^{1-\alpha} - \alpha)}{\sqrt{2N \log\log(N)}} = +\infty,$$

with $\mathbb{E}^{\tilde{\pi}}[x_n^{1-\alpha}] = \alpha$. This divergence is quite surprising given that $y_n = x_n^{1-\alpha}$ has finite moments. In fact, the stationary distribution of (y_n) is a $\mathcal{B}e(\alpha/(1-\alpha), 1)$ distribution.

Figure 1(a) illustrates the behaviour of the sequence

$$\frac{\sum_{n=1}^{N} (x_n^{1-\alpha} - \alpha)}{\sqrt{2N \log\log(N)}}$$

for a sample of size 1,000,000 and $\alpha = 0.2$, and we contrast it to Figure 1(b), where a similar sequence is built for an iid sample of $\mathcal{N}(0, 1)$ random variables. Note the smoother path in (a) compared with the erratic behavior of the path in (b). It seems difficult to discriminate the applicability of the central limit theorem by considering only such graphs, although Yu and Mykland's (1994) cusum criterion is based on a similar evaluation. The incredibly slow convergence of the chain (x_n) is also illustrated by Figure 3 since it requires more than 1,000,000 iterations for the ergodic average to get close to the correct mean, $\alpha = 0.2$. [Robert, 1995, details the convergence properties of (2.8) and, while explaining why the convergence is so slow, proposes this algorithm as a benchmark for convergence control methods.]

Recall (see Meyn and Tweedie, 1993, and Tierney, 1994) that a *Harris recurrent chain* is such that the probability of an infinite number of returns to an arbitrary set A is 1 for $\tilde{\pi}(A) > 0$. These chains are, in addition, β-mixing when the density of the transition probability with respect to $\tilde{\pi}$ is positive (Davydov, 1973) and Tierney (1994) shows that, under fairly general conditions, Metropolis and Gibbs kernels produce Harris recurrent chains. See also the discussion by Chan and Geyer (1994) for a thorough treatment of Harris recurrence in the setup of

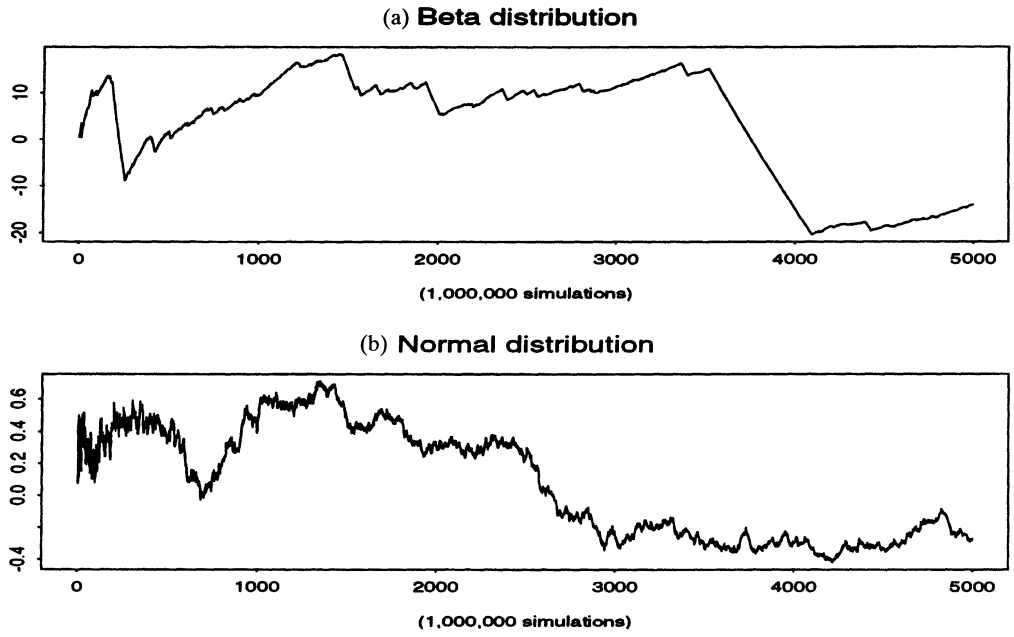

FIG. 1. *Law of the iterated logarithm evaluated for* (a) *a* $\mathscr{B}e(0.2, 1)$ *distribution simulated according to* (2.8); (b) *an iid sample from* $\mathscr{N}(0, 1)$. *The first series is divergent and the second series covers* $[-1, 1]$.

MCMC algorithms and for its connections with the central limit theorem.

(c) φ-MIXING. This mixing condition is sufficient to ensure by itself that the central limit theorem holds. It is defined through the mixing coefficient

$$\varphi(n) = \sup_{A, B} |\pi^n(x_n \in A |\ x_0 \in B) - \tilde{\pi}(x_n \in A)|,$$

which goes to 0 as n goes to infinity for φ-mixing chains. When φ-mixing holds, for every h in $L_2(\pi)$ such that $\mathbb{E}^{\tilde{\pi}}[h(X_0)] = 0$, the series

$$\sigma_h^2 = \mathbb{E}^{\tilde{\pi}}[h(X_0)^2] + 2\sum_{k=1}^{\infty} \mathbb{E}^{\tilde{\pi}}[h(X_0)h(X_k)]$$

is absolutely convergent and, if $\sigma_h > 0$, the central limit theorem applies to S_n, the limiting distribution being $\mathscr{N}(0, \sigma_h^2)$ (Billingsley, 1968). Moreover, checking φ-mixing is often straightforward: all finite and most compact state irreducible Markov chains are φ-mixing (Billingsley, 1968). It is also shown in Davydov (1973) that φ-mixing for Harris recurrent Markov chains is equivalent to *Döeblin irreducibility* when the kernel is strictly positive on the support of $\tilde{\pi}$. (See Meyn and Tweedie, 1993, for the relationship between Döeblin irreducibility and the existence of small sets underlying renewal theory.)

These mixing properties of Markov chains are thus well related with the central limit theorem and its applicability, but it can still be argued that they are too theoretical to be used in practical MCMC setups. We refute this potential criticism in Section 4 by introducing the *duality principle*, since some simple chains used in MCMC algorithms satisfy the above mixing properties. Meanwhile, Section 3 describes another promising approach to the assessment of the central limit theorem and to the estimation of σ_h^2.

3. RENEWAL THEORY

3.1 Preliminary Notions

As noted by Mykland, Tierney and Yu (1995), the renewal properties of the Markov chain under study can be used to assess convergence of the chain to the stationary distribution and to improve the estimation of the parameters of this distribution. The second aspect is more forcibly stressed by Mykland, Tierney and Yu (1995), but we want to present both appealing aspects of renewal theory because the improvement mentioned just above can be operated in a semiautomated manner, with no call to additional Metropolis steps, and it is particularly relevant when the duality principle can be used. The main appeal of renewal theory is that, when it applies, the study of the generic sums

$$S_N = \sum_{n=1}^{N} h(x_n)$$

can be simplified in a monitoring of iid random variables and a classical form of the central limit theorem then applies. This transformation to a simpler setting is actually done by decomposing S_N into a sum of iid random variables.

The condition for renewal theory to apply is that there exists a set A, a real $0 < \varepsilon < 1$ and a probability measure ν such that

$$P(x_{n+1} \in B | x_n) \geq \varepsilon \nu(B), \quad \forall x_n \in A, \ \forall B. \tag{3.1}$$

The set A is called *renewal set* (Asmussen, 1979) or *small set* (Meyn and Tweedie, 1993). When (3.1) holds for a triplet (A, ε, ν), the transition kernel of the chain (x_n) can be modified without change of stationary distribution. In fact, since

$$k(x_{n+1}|x_n) = \varepsilon\nu(x_{n+1}) + (1-\varepsilon)\frac{k(x_{n+1}|x_n) - \varepsilon\nu(x_{n+1})}{1-\varepsilon}$$

and since both terms of the mixture are positive when $x_n \in A$, we can generate x_{n+1} according to

$$x_{n+1} = \begin{cases} y_1 \sim \nu(y_1), & \text{with probability } \varepsilon, \\ y_2 \sim \dfrac{k(x_{n+1}|x_n) - \varepsilon\nu(x_{n+1})}{1-\varepsilon}, & \\ & \text{with probability } 1-\varepsilon, \end{cases} \tag{3.2}$$

when $x_n \in A$. The chain is not formally modified since we are marginally simulating from $k(x_{n+1}|x_n)$ at each step. However, if we take into account the uniform random variable u_n generated to choose between y_1 and y_2, the decomposition (3.2) introduces independent generations from a distribution ν when $x_n \in A$ and $u_n < \varepsilon$. We can then define a sequence of renewal times τ_t by

$$\tau_{t+1} = \inf\{n > \tau_t;\ x_n \in A \text{ and } u_n \leq \varepsilon\}.$$

The blocks $(x_{\tau_t+1}, \ldots, x_{\tau_{t+1}})$ are independent and the partial sums

$$S_t = \sum_{n=\tau_t+1}^{\tau_{t+1}} h(x_n)$$

are iid under stationarity. They thus satisfy the following limit theorem under usual regularity conditions:

LEMMA 3.1. *If* $\mathbb{E}[\tau_1] < \infty$ *and* $h \in L_1(\tilde{\pi})$, *the partial sums* S_t *satisfy:*

(i) $$\sum_{t=1}^{T} S_t/(\tau_{T+1} - \tau_1) \to \mathbb{E}^{\tilde{\pi}}[h(x)] \quad (a.s.)\ as\ T \to \infty;$$

(ii) $$\tau_T/T \to \mathbb{E}^{\tilde{\pi}}[\tau_2 - \tau_1] \quad (a.s.\ as\ T \to \infty) = \mu_A \quad (a.s.\ as\ T \to \infty).$$

Note that, since most MCMC algorithms produce Harris recurrent Markov chains, a finite average return time to A is usually guaranteed in most cases. Moreover, this renewal decomposition ensures the applicability of the central limit theorem for the original sum, under the conditions

$$\mathbb{E}^{\tilde{\pi}}[(\tau_{t+1} - \tau_t)^2] < \infty \quad \text{and} \quad \mathbb{E}^{\tilde{\pi}}\left[\left(\sum_{n=1}^{\tau_1} h(x_n)\right)^2\right] < \infty, \tag{3.3}$$

which imply $\sigma_h < \infty$ (see Meyn and Tweedie, 1993). Indeed, if we denote by $\lambda_t = (\tau_{t+1} - \tau_t)$ the excursion times and by T the number of renewal events before N, the normalized sum

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)]) = \frac{1}{\sqrt{N}}\left\{\sum_{n=1}^{\tau_1}(h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)]) + \sum_{t=1}^{T}(S_t - \lambda_t\mathbb{E}^{\tilde{\pi}}[h(x)]) + \sum_{n=\tau_T+1}^{N}(h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)])\right\}$$

is (a.s.) equivalent to

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{T}(S_t - \lambda_t\mathbb{E}^{\tilde{\pi}}[h(x)])$$

under the conditions (3.3) (since the first and the third terms converge a.s. to 0) and

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(S_t - \lambda_t\mathbb{E}^{\tilde{\pi}}[h(x)]) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \tilde{\sigma}_A^2)$$

by virtue of the usual central limit theorem, the asymptotic variance being indexed by the renewal set. Therefore, the central limit theorem truly applies to the sum of the $(h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)])$.

3.2 A Convergence Criterion

We now show how renewal theory provides an implementable estimation procedure for σ_h^2 and thus a convergence criterion for MCMC algorithms. Since the random variables $(S_t - \lambda_t\mathbb{E}^{\tilde{\pi}}[h(x)])$ are independent, the renewal variance $\tilde{\sigma}_A^2$ can be estimated by the usual sum of squares estimator

$$\frac{1}{T}\sum_{t=1}^{T}(S_t - \lambda_t\mathbb{E}^{\tilde{\pi}}[h(x)])^2$$

or, since the expectation $\mathbb{E}^{\tilde{\pi}}[h(x)]$ is unknown, by

$$\hat{\sigma}_A^2 = \frac{1}{T}\sum_{t=1}^{T}\left(S_t - \lambda_t \sum_{l=1}^{T}\frac{S_l}{N}\right)^2. \tag{3.4}$$

We then deduce the following invariance property.

PROPOSITION 3.2. *For every small set A such that* (3.3) *holds, the ratio*

$$\frac{\hat{\sigma}_A T}{N} \tag{3.5}$$

converges a.s. (in N) to σ_h^2.

PROOF. The result follows immediately from the a.s. convergence of $\hat{\sigma}_A^2$ to $\tilde{\sigma}_A^2$ and of (ii) in Lemma 3.1, as N/T converges a.s. to μ_A. Since

$$\frac{1}{\sqrt{N}}\sum_{n=1}^{N}(h(x_n) - \mathbb{E}^{\tilde{\pi}}[h(x)]) - \sqrt{\frac{T}{N}}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(S_t - \lambda_t \mathbb{E}^{\tilde{\pi}}[h(x)])$$

converges a.s. to 0 and the second term converges in distribution to $\mathcal{N}(0, \tilde{\sigma}_A^2/\mu_A)$, while the first term converges to $\mathcal{N}(0, \sigma_h^2)$ if $\sigma_h > 0$, the asymptotic invariance of the ratio $\hat{\sigma}_A T/N$ follows. □

That (3.5) is a convergent estimator of σ_h^2 is obviously an interesting feature, since it shows that renewal theory can lead to the estimation of the asymptotic variance, just as well as spectral theory or other time-series methods. However, for us the main incentive for using renewal theory is the asymptotic invariance of this ratio (3.5), because we can deduce a convergence criterion: *given several small sets A_i, wait until the ratios $\sigma_{A_i}^2 T_i/N$ have similar values.*

Although the theoretical basis of this method is quite sound, we are faced with two implementation caveats: first, the criterion is conservative, in the sense that it requires the slowest ratio to converge for the algorithm to stop. Second, as in other parallel methods, the dependence on the starting values is crucial, since close A_i's will lead to earlier terminations than far-spaced A_i's, while it is usually impossible to assess how close "close" is. However, we will introduce below a general class of models for which these drawbacks can be overcome.

3.3 Implementation of the Method

A first difficulty with the renewal convergence diagnosis is that small sets sets must be easily exhibited. Mykland, Tierney and Yu (1995) proposed a hybrid algorithm which overcomes this problem by adding an additional Metropolis step in the algorithm. This modification is the obvious solution when the transition kernel is too complex or too highly multidimensional to be examined in detail, but we want to point out the degree of generality of the renewal phenomenon to maintain that a modification of the algorithm is usually superfluous in theory, if not in practice. In fact, it follows from Asmussen (1979) that every ν-irreducible Markov chain allows for renewal. Since the Markov chains occurring in MCMC setups are generally $\tilde{\pi}$-irreducible, any probability measure ν equivalent to $\tilde{\pi}$ can thus be chosen for a renewal measure. In addition, a more precise result of Meyn and Tweedie (1993) states that every set E such that $\tilde{\pi}(E) > 0$ contains a small set A associated with $\tilde{\pi}$ and a corresponding bound ε. Therefore, from a theoretical point of view, renewal occurs for a wide range of models, even though the parameters A, ε and ν are not provided by the theory. Some settings allow for the whole space to be a small set, but the corresponding renewal rate may be too small to be of use in practice (see Robert, 1996, for examples).

In discrete cases, A can be selected as the collection of the most frequent states of the chain and ν derived as

$$\nu(E) = \inf_{x_n \in A} P(x_{n+1} \in E | x_n)$$

(see Section 3.4). In general, however, the derivation of (A, ε, ν) implies a more involved study of the particular Markov chain produced by the MCMC algorithm and of the corresponding kernel. Automated convergence diagnoses based on renewal theory then appear to remain out of reach, although the *duality principle* introduced in Section 4 shows why this conclusion can be attenuated.

Note also that (3.2) requires a simulation from

$$\frac{k(x_{n+1}|x_n) - \varepsilon\nu(x_{n+1})}{1-\varepsilon}, \tag{3.6}$$

but that k is usually unknown in a closed form. This can be achieved by simulating from $k(\cdot|x_n)$ until acceptance.

LEMMA 3.3. *Simulation from* (3.6) *can be done according to the following algorithm:*

1. *Simulate y from $k(\cdot|x_n)$;*
2. *Reject y with probability $\varepsilon\nu(y)/k(y|x_n)$.*

This lemma involves the computation of the ratio $\varepsilon\nu(y)/k(y|x_n)$ which can be approximated by regular Monte Carlo simulations. In fact, if the transition kernel is as in (2.1), $k(z|x_n)$ can be estimated

by

$$\frac{1}{M}\sum_{m=1}^{M} f_{X|Y}(z|y_m), \tag{3.7}$$

where the y_m's are iid from $f_{Y|X}(y|x_n)$, since (3.7) converges to $k(z|x_n)$ with M. The following example illustrates this approximation. In Metropolis setups, the transition kernel involves a Dirac mass, but, as pointed out by a referee, the lower bound on k derived from the continuous part of the transition kernel is sufficient for generation from (3.6).

EXAMPLE 3.1. Consider a normal prior, $\mathcal{N}(0, \sigma^2)$, on the location parameter θ of a Cauchy $\mathcal{C}(\theta, 1)$ distribution, with three observations x_1, x_2, x_3 from $\mathcal{C}(\theta, 1)$. The posterior distribution on θ is then

$$\begin{aligned}\pi(\theta|x_1, x_2, x_3) &\propto \{\exp(\theta^2/2\sigma^2)[1+(\theta-x_1)^2] \\ &\quad \cdot [1+(\theta-x_2)^2][1+(\theta-x_3)^2]\}^{-1}.\end{aligned} \tag{3.8}$$

A Gibbs sampler for the simulation from (3.8) is based on three artificial random variables η_1, η_2, η_3 such that (3.8) appears as the marginal distribution of

$$\begin{aligned}\pi(\theta, \eta_1, \eta_2, \eta_3|x_1, x_2, x_3) &\propto \exp(-\theta^2/2\sigma^2)\exp(-(1+(\theta-x_1)^2)\eta_1/2) \\ &\quad \cdot \exp(-(1+(\theta-x_2)^2)\eta_2/2) \\ &\quad \cdot \exp(-(1+(\theta-x_3)^2)\eta_3/2),\end{aligned} \tag{3.9}$$

since the conditional distributions derived from (3.9) are

$$\eta_i|(\theta, x_i) \sim Exp\left(\frac{1+(\theta-x_i)^2}{2}\right), \quad i = 1, 2, 3,$$

$$\theta|(\eta_i, x_i) \sim \mathcal{N}\left(\frac{\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3}{\eta_1+\eta_2+\eta_3+\sigma^{-2}}, \frac{1}{\eta_1+\eta_2+\eta_3+\sigma^{-2}}\right). \tag{3.10}$$

We denote $\tau^{-2} = \eta_1 + \eta_2 + \eta_3 + \sigma^{-2}$ and omit the dependence on (η_1, η_2, η_3).

The posterior distribution (3.8) is sometimes trimodal, depending on the values of x_1, x_2 and x_3, as shown by Figure 2. The gaps between the three peaks are so severe that they could jeopardize the actual convergence of the MCMC algorithm. However, a simulation based on 20,000 iterations of (3.10) shows that this is not the case, since the histogram in Figure 2 reproduces the shape of (3.8) quite accurately.

If the small set A is chosen of the form $[r_1, r_2]$ with $x_2 \in [r_1, r_2]$, $x_1 < r_1$ and $x_3 > r_2$ (assuming $x_1 < x_2 < x_3$), the bounds

$$\begin{aligned}&\rho_{11} = r_1 - x_1 < |\theta - x_1| < \rho_{12} = r_2 - x_1, \\ &0 < |\theta - x_2| < \rho_{22} = \max(r_2 - x_2, x_2 - r_1), \\ &\rho_{31} = x_3 - r_2 < |\theta - x_3| < \rho_{32} = x_3 - r_1\end{aligned}$$

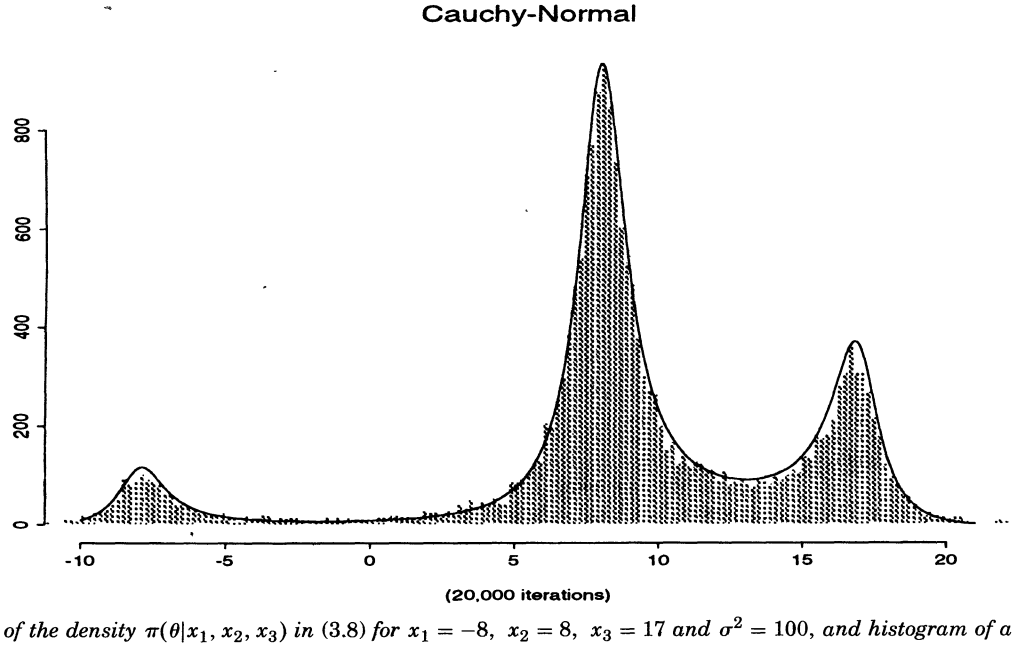

FIG. 2. *Graph of the density* $\pi(\theta|x_1, x_2, x_3)$ *in* (3.8) *for* $x_1 = -8$, $x_2 = 8$, $x_3 = 17$ *and* $\sigma^2 = 100$, *and histogram of a Gibbs sample of size* 20,000 *for this distribution.*

induce a minorizing probability measure ν and a corresponding constant ε. Indeed,

$$k(\theta'|\theta) \geq \int_{\mathbb{R}_+^3} \exp\left\{-\frac{(\theta' - \tau^2(\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3))^2}{2\tau^2}\right\} \cdot \tau \frac{1}{\sqrt{2\pi}} \frac{1+\rho_{11}^2}{2} \exp\left\{-\frac{(1+\rho_{12}^2)\eta_1}{2}\right\} \cdot \frac{1}{2} \exp\left\{-\frac{(1+\rho_{22}^2)\eta_2}{2}\right\} \frac{1+\rho_{31}^2}{2} \cdot \exp\left\{-\frac{(1+\rho_{32}^2)\eta_3}{2}\right\} d\eta_1\, d\eta_2\, d\eta_3$$

$$= \frac{1+\rho_{11}^2}{1+\rho_{12}^2} \frac{1}{1+\rho_{22}^2} \frac{1+\rho_{31}^2}{1+\rho_{32}^2} \nu(\theta') = \varepsilon\nu(\theta'),$$

where ν is the density of the marginal distribution (in θ) of

$$(\theta, \eta_1, \eta_2, \eta_3) \sim \mathcal{N}(\tau^2(\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3), \tau^2) \cdot Exp\left(\frac{1+\rho_{12}^2}{2}\right) Exp\left(\frac{1+\rho_{22}^2}{2}\right) Exp\left(\frac{1+\rho_{32}^2}{2}\right).$$

Therefore, given r_1 and r_2, ε and ν, a practical implementation of the renewal perturbation of the chain is possible. As mentioned earlier, a difficulty related to the simulation from (3.6) is that the ratio $k(\theta'|\theta)/\nu(\theta')$ is not available. We use the Monte Carlo approximation

$$\frac{k(\theta'|\theta)}{\nu(\theta')} \simeq \frac{\sum_{m=1}^M \varphi(\theta'|\eta_1^m, \eta_2^m, \eta_3^m)}{\sum_{m=1}^M \varphi(\theta'|\tilde{\eta}_1^m, \tilde{\eta}_2^m, \tilde{\eta}_3^m)},$$

where

$$\varphi(\theta'|\eta_1, \eta_2, \eta_3) = \tau^{-1} \exp\{-(\theta' - \tau^2(\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3))^2/2\tau^2\}$$

and

$$\eta_i^m \sim Exp\left(\frac{1+(\theta - x_m)^2}{2}\right),$$

$$\tilde{\eta}_i^m \sim Exp\left(\frac{1+\rho_{m2}^2}{2}\right), \quad i = 1, 2, 3.$$

We took $M = 50$ in our simulations. The η_i^m's have to be generated for each iteration from (3.6) since they depend on x_m, while the $\tilde{\eta}_i^m$'s can be simulated only once when initializing the Gibbs sampler.

As shown by Table 1, which describes the results of the simulations we conducted, the bound ε decreases quite slowly to 0 as ρ_{22} increases; the number of renewals in a sequence of Gibbs iterations is thus likely to be sufficiently high, although both quantities are not strongly connected. The average number of steps between two returns to A goes as low as 8.8 when $A = [x_2 - r, x_2 + r]$ and $r = 0.54$. Note the actual stabilization to $\sigma_h^2 \simeq 1150$ for most values of r. This large variance factor may be explained by the Cauchy tails of the posterior distribution as well as the iterative switching from one mode to another in the Gibbs algorithm.

TABLE 1

Renewal parameters when $A = [x_2 - r, x_2 + r]$ and $h(x) = x$ for $x_1 = -8$, $x_2 = 8$, $x_3 = 17$ and $\sigma^2 = 100$ (based on 1,000,000 simulations); $\bar{\tau}_A$ is the mean excursion time and $\hat{\sigma}_A^2$ is the estimate of σ_h^2 based on (3.5)

r	0.1	0.21	0.32	0.43	0.54	0.65	0.76	0.87	0.98	1.09
ε_A	0.92	0.83	0.73	0.63	0.53	0.45	0.38	0.31	0.26	0.22
$\bar{\tau}_A$	25.3	13.9	10.5	9.6	8.8	9.6	9.8	10.4	11.4	12.7
$\hat{\sigma}_A^2$	1135	1138	1162	1159	1162	1195	1199	1149	1109	1109

EXAMPLE 2.1 (Continued). For this very slowly converging chain, the natural renewal sets are $A_\varepsilon = [\varepsilon, 1]$ since the transition kernel is bounded from below by $\varepsilon(\alpha+1)x^\alpha$ when $x_n \in A_\varepsilon$. Moreover, the simulation from (3.6) is then straightforward and Table 2 shows the range of the estimates of σ_h^2 for $h(x) = x^{1-\alpha}$ when ε varies, after 5×10^7 iterations. The criterion thus indicates that convergence is not yet reached, which is indeed the case.

The simplicity of the derivation of A, ε and ν in Examples 2.1 and 3.1 may be misleading in the sense that A is chosen because of the shape of the posterior distribution of θ, which is not always available, especially in multidimensional setups. Mykland, Tierney and Yu (1995) exhibit general renewal features for Metropolis algorithms, but it seems quite difficult to envision an automated version of renewal control for MCMC algorithms. The improvement brought by MCMC methods in computation time and in the range of fields accessible to Bayesian analysis would thus be cancelled by either an expensive preliminary analysis required for a proper implementation of these methods or by a lack of control, which jeopardizes the validity of their output.

As mentioned above and as shown by Section 3.4, this negative perspective does not hold when we consider finite Markov chains. Due to the duality principle exposed in Section 4, the appeal of this

TABLE 2

Estimates of σ_h^2 for different small sets $[\varepsilon, 1]$

ε	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85
$\hat{\sigma}_\varepsilon^2$	22.26	41.55	27.41	41.17	37.58	31.41	41.97	41.64

method extends to more elaborate models and explains why we still reckon this approach to be a very powerful tool for the control of MCMC methods.

3.4 Renewal Control of Finite Markov Chains

Finite state space Markov chains are ideal settings for the application of the above technique. Consider a chain (x_n) with values in the finite space $\mathscr{X} = \{t_1, \ldots, t_m\}$ and transition matrix $P = (p_{ij})$. The cardinal m is often of the form k^p, in particular in missing data problems (see Section 4). We assume the chain (x_n) to be irreducible and aperiodic. Define $\tilde{\pi} = (\pi_1, \ldots, \pi_m)$ as the stationary distribution and take π_{i_0} as the probability of the most probable state i_0. Then, if A is $\{i_0\}$, renewal theory applies with $\varepsilon = 1$ and $\nu = (p_{i_0 1}, \ldots, p_{i_0 m})$. The "perturbation" (3.2) is then useless and the ratio $k(\theta'|\theta)/\nu(\theta')$ is equal to 1.

The consequences of this simplification are, however, far from trivial. Indeed, the sums $\sum_{w=1}^{W} h(x_w)$ can then be decomposed into iid sums

$$S_j = \sum_{w=\tau_j+1}^{\tau_{j+1}} h(x_w) = h(t_{i_0}) + \sum_{w=\tau_j+1}^{\tau_{j+1}-1} h(x_w), \quad j = 1, \ldots,$$

where

$$\tau_j = \inf\{w > \tau_{j-1}; x_\omega = t_{i_0}\}.$$

The variance σ_h^2 of the asymptotically normal expression

$$W^{-1/2} \sum_{w=1}^{W} (h(x_w) - \mathbb{E}^{\tilde{\pi}}[h(x)])$$

can therefore be estimated directly by (3.4) and (3.5). Moreover, a convergence criterion can be derived by considering other probable states $i_1, \ldots, i_c$ and checking convergence to the same value of the corresponding estimators (3.5) of σ_h^2.

EXAMPLE 3.2. Consider the special case $\mathscr{X} = \{0, 1, 2, 3\}$ and (x_n) with transition matrix

$$P = \begin{pmatrix} 0.26 & 0.04 & 0.08 & 0.62 \\ 0.05 & 0.24 & 0.03 & 0.68 \\ 0.11 & 0.10 & 0.08 & 0.71 \\ 0.08 & 0.04 & 0.09 & 0.79 \end{pmatrix}.$$

The stationary distribution is $\tilde{\pi} = (0.097, 0.056, 0.085, 0.762)$ and the corresponding mean is 2.507. If we use a simulation of this Markov chain, the four states can be chosen as renewal sets and an estimate of σ_h^2 can be constructed for each state, based on (3.3) and (3.4). Table 3 provides the different estimates of σ_h^2, for 5000 and 500,000 simulations from P, for $h(x) = x$. The larger simulation study clearly shows that convergence is achieved since the four estimates of σ_h^2 are equal up to the second decimal.

TABLE 3
Estimates of σ_h^2 for $h(x) = x$, based on renewal at i_0

n/i_0	0	1	2	3
5,000	1.19	1.29	1.26	1.21
500,000	1.344	1.335	1.340	1.343

In most practical setups, the most probable state i_0 is unknown beforehand, as are the other probable states $i_1, \ldots, i_c$. We suggest taking advantage of some "burn-in" initial iterations of the MCMC algorithm to derive these states or some approximations since, when the chain is close to stationarity, the most commonly sampled states are the most probable states for $\tilde{\pi}$. In some setups such as Ising models, the state space is too large for a single state to be probable enough, that is, to have a probability of occurrence larger than 0.01 or 0.005, say. In this case, the renewal set A can be selected as a union of states, $A = \{i_0, i_1, \ldots, i_r\}$, the renewal measure ν being defined by

$$\nu(i) \propto \min_{j \in A} p_{ji} \tag{3.11}$$

and the bound ε by

$$\varepsilon = \sum_{i=1}^{m} \min_{j \in A} p_{ji}. \tag{3.12}$$

If m is too large for the above distribution to be computed exactly, note that (3.12) is bounded from below by a similar sum on the most common states and that an additional Metropolis step can be used to simulate from (3.11).

4. THE DUALITY PRINCIPLE

Let us introduce the duality principle through the example by Diebolt and Robert (1994), which initiated this notion, as it provides a good insight into the theory.

EXAMPLE 4.1. Consider a two-component normal mixture distribution

$$p\mathscr{N}(\theta_1, \sigma_1^2) + (1-p)\mathscr{N}(\theta_2, \sigma_2^2), \tag{4.1}$$

with the conjugate prior distributions

$$p \sim \mathscr{B}e(1/2, 1/2),$$
$$\theta_i \sim \mathscr{N}(\xi_i, \sigma_i^2/n_i), \quad \sigma_i^2 \sim \mathscr{IG}(\nu_i/2, \omega_i^2/2), \quad i = 1, 2.$$

Given a sample $x_1, \ldots, x_n$ from (4.1), the posterior distribution appears as a sum of 2^n closed form terms from exponential families. It thus requires an MCMC approximation when n is larger than 30 (see Diebolt and Robert, 1994). The fruitful approach to the mixture problem is to perceive the model as a *missing data structure*, by introducing $z_1, \ldots, z_n$, indicators of the components from which the x_i's originated. The "completed model" stands as

$$z_i \sim \mathcal{B}(1, p), \quad x_i|z_i \sim \mathcal{N}(\theta_{2-z_i}, \sigma^2_{2-z_i}),$$

and the corresponding Gibbs implementation is to simulate iteratively the missing data and the parameters. Each simulation of the missing data provides two subsamples of sizes t and $n-t$ corresponding to each component and related averages $\bar{m}_1$ and $\bar{m}_2$, sums of squared errors s_1^2 and s_2^2. The two steps of a Gibbs iteration are then as follows:

1. Simulate

$$z_i \sim \mathcal{B}\Big(1, p\sigma_2 \exp(-(x_i-\theta_1)^2/2\sigma_1^2) \cdot \big[p\sigma_2 \exp(-(x_i-\theta_1)^2/2\sigma_1^2) + (1-p)\sigma_1 \cdot \exp(-(x_i-\theta_2)^2/2\sigma_2^2)\big]^{-1}\Big), \quad i = 1, \ldots, n.$$

2. Simulate

(i) $p \sim \mathcal{B}e(t+1, n-t+1)$;

(ii) $\sigma_1^2 \sim \mathcal{IG}\left(\frac{\nu_1+t}{2}, \frac{1}{2}\left(\omega_1^2 + s_1^2 + \frac{n_1 t}{n_1+t} \cdot \sum_i z_i(x_i-\xi_1)^2\right)\right)$;

(iii) $\sigma_2^2 \sim \mathcal{IG}\left(\frac{\nu_2+n-t}{2}, \frac{1}{2}\left(\omega_2^2 + s_2^2 + \frac{n_2(n-t)}{n_1+n-t} \cdot \sum_i (1-z_i)(x_i-\xi_2)^2\right)\right)$;

(iv) $\theta_1 \sim \mathcal{N}\left(\frac{n_1\xi_1 + t\bar{m}_1}{n_1+t}, \frac{\sigma_1^2}{n_1+t}\right)$;

(v) $\theta_2 \sim \mathcal{N}\left(\frac{n_2\xi_2 + (n-t)\bar{m}_2}{n_2+n-t}, \frac{\sigma_2^2}{n_2+n-t}\right)$.

In the general setting of this paper, the important aspect of the above algorithm is not the ergodic convergence to the stationary distribution $\tilde{\pi}$. It is rather that the parameter of interest $(p, \theta_1, \theta_2, \sigma_1, \sigma_2)$ is simulated (as a whole) conditionally on the z's, with $z = (z_1, \ldots, z_n)$ and therefore that *the finite state space Markov chain* (z^t) *determines the properties of the MCMC algorithm.* (In this case, the state space is of cardinality 2^n.)

Note that this property does not hold for alternative implementations of the Gibbs sampler. For instance, consider Mengersen and Robert's (1995) reparameterization

$$p\mathcal{N}(\mu, \tau^2) + (1-p)\mathcal{N}(\mu + \tau\theta, \tau^2\sigma^2),$$

with

$$\pi(p, \mu, \tau) = \tau^{-1}, \quad \sigma \sim \mathcal{U}_{[0,1]} \quad \text{and} \quad \theta \sim \mathcal{N}(0, \zeta^2).$$

This equivalent representation of (4.1) expresses the parameters of the second component as a local perturbation of an overall location–scale parameter (μ, τ) and is mainly of interest in noninformative settings since it allows for improper priors on (μ, τ). However, although it provides a higher efficiency in the Gibbs sampler, this perspective requires full conditional distributions and the parameter is not generated conditionally on z. In fact, step 2 is then:

2. Simulate

(i) $p \sim \mathcal{B}e(t+1, n-t+1)$;

(ii) $\sigma_1^{-2} \sim \mathcal{IG}\left(\frac{t+2}{2}, \frac{t(\bar{m}_1-\theta_1)^2 + s_1^2 + (\theta_2-\theta_1)^2\zeta^{-2}}{2}\right) \cdot \mathbb{I}_{[\sigma_2,\infty)}(\sigma_1)$;

(iii) $\sigma_2^{-2} \sim \mathcal{IG}\left(\frac{n-t-2}{2}, \frac{(n-t)(\bar{m}_2-\theta_2)^2 + s_2^2}{2}\right) \cdot \mathbb{I}_{[0,\sigma_1]}(\sigma_2)$;

(iv) $\theta_1 \sim \mathcal{N}\left(\frac{t\bar{m}_1 + \zeta^{-2}\theta_2}{t+\zeta^{-2}}, \frac{\sigma_1^2}{t+\zeta^{-2}}\right)$;

(v) $\theta_2 \sim \mathcal{N}\left(\frac{(n-t)\bar{m}_2 + \zeta^{-2}\theta_1\sigma_2^2/\sigma_1^2}{n-t+\zeta^{-2}\sigma_2^2/\sigma_1^2}, \frac{\sigma_2^2}{n-t+\zeta^{-2}\sigma_2^2/\sigma_1^2}\right)$,

when expressed in the parameterization of (4.1) (see also Robert and Titterington, 1996). Therefore, due to the dependence on the previous value of the parameter, it is not possible to use the finite state space chain (z_n) to create renewal sets. Moreover,

the subchain (z_n) cannot be considered independently from the parameter subchain since it is not a Markov chain.

In many MCMC setups similar to Example 4.1, the algorithm produces several chains in parallel. This is particularly true of *data augmentation* (Tanner and Wong, 1987), of *interleaving* Markov chains (Liu, Wong and Kong, 1994) and of general Gibbs sampling. In some cases, the chains of interest are not necessarily Markov chains, but the result below shows why this is not really a concern. Surprisingly enough, it says that it is not always appropriate to study directly the chain of of interest. More precisely, the *duality principle* leading to Theorems 4.1, 4.2 and 4.3 states that in cases where the chain (θ_n) is derived from a second chain (z_n) by simulation from $\pi(\theta|z)$, the properties of the chain (θ_n), whether it is a Markov chain or not, can be gathered from those of the chain (z_n). In this setup, z_n is simulated according to $f(z|\theta_{n-1}, z_{n-1})$.

THEOREM 4.1. *If the chain (z_n) is ergodic with stationary distribution $\tilde{f}$ (respectively geometrically ergodic with rate ϱ), the chain (θ_n) derived by $\theta_n \sim \pi(\theta|z_n)$ is ergodic (geometrically ergodic) for every conditional distribution $\pi(\cdot|z)$ and its stationary distribution is*

$$\tilde{\pi}(\theta) = \int \pi(\theta|z)\tilde{f}(z)\,dz.$$

Moreover, if (z_n) is φ-mixing, (θ_n) is also φ-mixing.

PROOF. The transition kernel associated to the chain (z_n) is

$$k(z'|z) = \int \pi(\theta|z)f(z'|\theta, z)\,d\theta.$$

If f^n is the marginal density of z_n at step n, $\pi^n(\theta) = \int \pi(\theta|z)f^n(z)\,dz$ is the marginal density of θ_n at step n and

$$\begin{aligned} &||\pi^n - \tilde{\pi}||_{\mathrm{TV}} \\ &\quad = \tfrac{1}{2}\left|\int_{\mathscr{Z}\times\Theta} \pi(\theta|z)(f^n(z) - \tilde{f}(z))\,dz\,d\theta\right| \\ &\quad \le ||f^n - \tilde{f}||_{\mathrm{TV}}. \end{aligned} \tag{4.2}$$

Therefore, (θ_n) converges to $\tilde{\pi}$ for every possible starting point and the chain is ergodic when (z_n) is ergodic. The same transfer applies for geometric ergodicity. Note that the inequalities

$$||f^{n+1} - \tilde{f}||_{\mathrm{TV}} \le ||\pi^n - \tilde{\pi}||_{\mathrm{TV}} \le ||f^n - \tilde{f}||_{\mathrm{TV}}$$

imply that the same geometric rate ϱ applies to both chains.

Moreover, if φ-mixing holds for the chain (z_n), there exists a finite measure μ such that

$$|f^n(z) - \tilde{f}(z)| \le \varphi(n)\mu(z)$$

and

$$\begin{aligned} |\pi^n(\theta) - \tilde{\pi}(\theta)| &\le \int_{\mathscr{Z}} \pi(\theta|z)|f^n(z) - \tilde{f}(z)|\,dz \\ &\le \varphi(n)\int_{\mathscr{Z}} \pi(\theta|z)\mu(z)\,dz = \varphi(n)\tilde{\mu}(\theta). \end{aligned}$$

The measure $\tilde{\mu}$ is finite since

$$\int_\Theta \tilde{\mu}(\theta)\,d\theta = \int_{\mathscr{Z}} \mu(z)\,dz < \infty. \qquad \square$$

In the special case when (θ_n) is a Markov chain, for example, when $z_n \sim f(z_n|\theta_{n-1})$, which corresponds to data augmentation, α-mixing and β-mixing properties also transfer from (z_n) to (θ_n).

THEOREM 4.2. *If the chain (z_n) is α-mixing (respectively β-mixing), the chain (θ_n) is also α-mixing (β-mixing).*

PROOF. Consider the following representation of the α-mixing coefficients

$$\begin{aligned} &\alpha_\theta(n) \\ &\quad = \sup_{||h||_\infty<1} \int_\Theta \left|\int_\Theta h(\theta)(\pi^n(\theta|\theta_0) - \tilde{\pi}(\theta))\,d\theta\right| \tilde{\pi}(\theta_0)\,d\theta_0. \end{aligned}$$

Then

$$\begin{aligned} \alpha_\theta(n) &\le \sup_{||h||_\infty<1} \int_\Theta \left|\int_{\mathscr{Z}}\int_\Theta h(\theta)\pi(\theta|z)\,d\theta \right. \\ &\qquad \left. \cdot (f^n(z|\theta_0) - \tilde{f}(z))\,dz\right| \tilde{\pi}(\theta_0)\,d\theta_0 \\ &\le \sup_{||g||_\infty<1} \int_\Theta \int_{\mathscr{Z}} \left|\int_{\mathscr{Z}} g(z)(f^{n-1}(z|z_1) - \tilde{f}(z))\,dz\right| \\ &\qquad \cdot f(z_1|\theta_0)\,dz_1\tilde{\pi}(\theta_0)\,d\theta_0 \\ &= \sup_{||g||_\infty<1} \int_{\mathscr{Z}} \left|\int_{\mathscr{Z}} g(z)(f^{n-1}(z|z_1) - \tilde{f}(z))\,dz\right| \\ &\qquad \cdot \tilde{f}(z_1)\,dz_1 \\ &= \alpha_z(n-1). \end{aligned}$$

Similarly, since (Davydov, 1973)

$$\beta_\theta(n) = \int_\Theta \int_\Theta |\pi^n(\theta|\theta_0) - \tilde{\pi}(\theta)|\,d\theta\,\tilde{\pi}(\theta_0)\,d\theta_0,$$

we get

$$\begin{aligned}\beta_\theta(n) &\le \int_\Theta \int_{\mathscr{Z}} |f^n(z|\theta_0) - \tilde{f}(z)|\, dz\, \tilde{\pi}(\theta_0)\, d\theta_0 \\ &\le \int_\Theta \int_{\mathscr{Z}} \int_{\mathscr{Z}} |f^n(z|z_1) - \tilde{f}(z)|\, dz \\ &\qquad \cdot f(z_1|\theta_0)\tilde{\pi}(\theta_0)\, d\theta_0 \\ &= \int_{\mathscr{Z}} \int_{\mathscr{Z}} |f^{n-1}(z|z_0) - \tilde{f}(z)|\, dz\, \tilde{f}(z_0)\, dz_0 \\ &= \beta_z(n-1)\end{aligned}$$

and these inequalities complete the proof of Theorem 4.2. □

This correspondence between mixing properties of both chains is obviously of interest, considering the developments of Section 2, since it allows for the assessment of the central limit theorem at little expense. In fact, the sufficient conditions described in Section 2.3 have only to be checked for the chain (z_n) for the central limit theorem to apply to the chain (θ_n). Since (z_n) is usually a finite state space Markov chain, such checking is often straightforward. For instance, when the state space of (z_n) is finite, it follows from Billingsley (1968) that (z_n) is geometrically ergodic and even φ-mixing, under irreducibility and aperiodicity of the kernel.

Note also that this straightforward verification of the central limit theorem conditions is particularly compelling because of the Rao–Blackwell theorem. In fact, as suggested by Gelfand and Smith (1990), it is sometimes preferable to consider the expected sums

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{E}^{\pi^n}[h(\theta)|z_n] = \frac{1}{N}\sum_{n=1}^{N} \tilde{h}(z_n), \tag{4.3}$$

rather than the direct average $\sum_{n=1}^{N} h(\theta_n)/N$, since the integration leading to (4.3) reduces the variance of the estimate. [Liu, Wong and Kong (1994) give some sufficient conditions for this improvement to hold for every convex loss function.] Therefore, when Rao–Blackwellization is justified theoretically and when $\tilde{h}$ can be written explicitly, the convergence of (4.3) to the expected value $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ can be directly controlled by the central limit theorem because the estimates (4.3) only depend on the chain (z_n).

When Rao–Blackwellization does not apply, usual averages can still be directed by the chain (z_n), as shown by the following result:

THEOREM 4.3. *If (z_n) is geometrically convergent with compact state space and rate ϱ, for every $h \in L_2(\tilde{\pi})$, there exists C_h such that*

$$||\, \mathbb{E}^{\pi^n}[h(\theta)] - \mathbb{E}^{\tilde{\pi}}[h(\theta)]\,||_2 < C_h \varrho^n.$$

PROOF. Without loss of generality, consider the case when h is a real-valued function. Then

$$\begin{aligned}&\left(\mathbb{E}^{\pi^n}[h(\theta)] - \mathbb{E}^{\tilde{\pi}}[h(\theta)]\right)^2 \\ &\quad = \left(\int h(\theta)(\pi^n(\theta) - \tilde{\pi}(\theta))\, d\theta\right)^2 \\ &\quad = \left(\int\int h(\theta)\pi(\theta|z)\, d\theta (f^n(z) - \tilde{f}(z))\, dz\right)^2 \\ &\quad \le \max_z \left(\mathbb{E}^{\pi}[h(\theta)|z]^2\right) ||f^n - \tilde{f}||_1^2 < C_h^2 \varrho^{2n}.\end{aligned}$$ □

This result can be related to those of Liu, Wong and Kong (1994), who showed that an *interleaving* property of the MCMC chains, corresponding basically to data augmentation setups, allows for the application of Rao–Blackwell theorem, but also for monotone decrease to 0 of the covariances $\operatorname{cov}(h(\theta_n), h(\theta_{n+m}))$ (in m) and for geometric convergence of the empirical moments. Our approach is more general in the sense that it puts no restriction on the way (z_n) is generated, but conversely it does require a preliminary study of this chain to certify that the central limit theorem applies.

Renewal theory can take advantage of the duality principle as well since, when (z_n) has a finite support, the renewal set A can be reduced to a single point (atom), at least theoretically (see Section 3.4 for extensions to cases when no atom has a high enough probability of return). For instance, the set A for Example 4.1 can be constructed by allocating each observation x_i to its most probable component, that is, in practice, to the mode of the values taken by z_i in the "burn-in" sequence. When the excursion times τ_t are too large, it makes sense to remove the unstable observations from the definition of A, although this really complicates the derivation of ν.

Example 4.1 was instrumental in the derivation of the duality principle, but it is far from being the only setup where the duality principle applies with practical consequences. For instance, *grouping* (or *data coarsening* as in Heitjan and Rubin, 1991) is another type of missing data structure where the duality principle can strengthen the convergence study.

EXAMPLE 4.2. Consider the multinomial grouped model of Tanner and Wong (1987) and Gelfand and Smith (1990):

$$x \sim \mathscr{M}_5(a_1\mu + b_1, a_2\mu + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1-\mu-\eta)),$$

with

$$0 \le a_1 + a_2 = a_3 + a_4 = 1 - \sum_{i=1}^{4} b_i = c \le 1,$$

$$0 \le \mu, \eta \le 1,$$

and the a_i's and b_i's are known. This model is actually a grouping of

$$y \sim \mathcal{M}_9(a_1\mu, b_1, a_2\mu, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1-\mu-\eta))$$

into

$$x_1 = y_1 + y_2, \quad x_2 = y_3 + y_4, \quad x_3 = y_5 + y_6, \quad x_4 = y_7 + y_8, \quad x_5 = y_9.$$

When $\pi(\mu, \eta)$ is the Dirichlet $\mathcal{D}(1/2, 1/2, 1/2)$ distribution, the posterior distribution is not available in closed form, but only through a Gibbs algorithm completing the data as follows:

1. Simulate $z = (z_1, z_2, z_3, z_4) = (y_1, y_3, y_5, y_7)$ by

$$z_i \sim \mathcal{B}\left(x_i, \frac{a_i\mu}{a_i\mu + b_i}\right), \quad i = 1, 2,$$
$$z_i \sim \mathcal{B}\left(x_i, \frac{a_i\eta}{a_i\eta + b_i}\right), \quad i = 3, 4.$$

2. Simulate

$$(\mu, \eta) \sim \mathcal{D}(1/2+z_1+z_2, 1/2+z_3+z_4, 1/2+x_5).$$

The chain (z_n) is then generated on a finite state space of cardinal $x_1 \times x_2 \times x_3 \times x_4$, which should be of size moderate enough to allow for the selection of a single atom as renewal set A. The finiteness of the state space also guarantees that the central limit theorem applies for both chains.

For instance, take $a = (0.1, 0.14, 0.7, 0.9)$, $b = (0.17, 0.24, 0.19, 0.20)$ and $x_1 = 4$, $x_2 = 15$, $x_3 = 12$, $x_4 = 7$, $x_5 = 4$. The simulation according to steps 1 and 2 then gives $(0, 1, 0, 0)$ as the most likely state for $(z_1, \ldots, z_4)$ and the average excursion time is 27.1. The second most likely state is $(0, 2, 0, 0)$ with an average excursion time 28.1. To compare the performances of the variance estimators with a less frequent state, we also consider $(1, 1, 0, 0)$, which has an average excursion time of 49.2. Table 4 provides the Gibbs and variance estimates for three functions of interest,

$$h_1(\mu, \eta) = \mu - \eta, \quad h_2(\mu, \eta) = \mathbb{I}_{\mu>\eta},$$
$$h_3(\mu, \eta) = \frac{\mu}{1-\mu-\eta}$$

and for 50,000 iterations. The Gibbs estimate of $h_3(\mu, \eta)$ is based on a Rao–Blackwellized version

$$\frac{1}{N}\sum_{n=1}^{N} \frac{0.5 + z_1 + z_2}{x_5 - 0.5}, \tag{4.4}$$

due to substantial gains of stability in both the estimate and the corresponding variance. Note the influence of the less frequent state on the estimation of σ_3^2.

TABLE 4

Gibbs estimates of posterior expectations and asymptotic variances for three functions of interest, based on three renewal sets for the multinomial grouped model (50,000 *iterations*)

i	$\mathbb{E}^{\hat{\pi}}[h_i(\mu, \eta)\|x]$	$\hat{\sigma}_i^2(1)$	$\hat{\sigma}_i^2(2)$	$\hat{\sigma}_i^2(3)$
h_1	0.0005	0.758	0.720	0.789
h_2	0.496	1.24	1.21	1.25
h_3	0.739	1.45	1.41	1.67

EXAMPLE 4.3. Consider p random variables $y_1, \ldots, y_p \sim Exp(\theta)$ which are grouped into classes according to binary random variables $g_i \sim \mathcal{B}(1, \Phi(\gamma_1 - \gamma_2 y_i))$ as

$$x_i = \begin{cases} [y_i/a], & \text{if } g_i = 0, \\ [y_i/b], & \text{if } g_i = 1, \end{cases} \quad 1 \le i \le p,$$

where a, b, γ_1 and γ_2 are known, and Φ is the normal cdf. Heitjan and Rubin (1991) provide a justification for this model through round-up errors in surveys. The observations x_i can be completed by the missing data (y_i, g_i) and

$$f(y_i, g_i|x_i, \theta) \propto \theta e^{-\theta y_i}\{\mathbb{I}_{[ax_i, a(x_i+1)]}(y_i)\mathbb{I}_{g_i=0}[1 - \Phi(\gamma_1 - \gamma_2 y_i)] + \mathbb{I}_{[bx_i, b(x_i+1)]}(y_i)\mathbb{I}_{g_i=1}\Phi(\gamma_1 - \gamma_2 y_i)\}.$$

If the prior distribution on θ is a $\mathcal{G}a(\alpha, \beta)$ distribution, a Gibbs algorithm for the simulation of the posterior distribution of θ is to consider the Markov chain (z_n, θ_n) with the following transition steps:

1. Simulate $z_n = (y_1^n, g_1^n, \ldots, y_p^n, g_p^n)$ by

$$g_i^n \sim \mathcal{B}(1, \Phi(\gamma_1 - \gamma_2 y_i^{n-1}),$$
$$y_i^n | g_i^n \sim \theta_{n-1}\exp(-\theta_{n-1}y) \cdot \{\mathbb{I}_{[ax_i, a(x_i+1)]}(y_i)\mathbb{I}_{g_i^n=0} + \mathbb{I}_{[bx_i, b(x_i+1)]}(y_i)\mathbb{I}_{g_i^n=1}\}, \quad 1 \le i \le p.$$

2. Simulate

$$\theta_n \sim \mathcal{G}a\left(\alpha + p, \beta + \sum_{i=1}^{p} y_n^i\right).$$

In this case, the missing data (y_i, g_i) $(1 \le i \le p)$ have a compact support and the chain z_n is φ-mixing. Therefore, the central limit theorem also applies.

This example shows that the duality principle applies in a wider context than just data augmentation, that is, when $z_n \sim f(z|\theta_{n-1})$ and $\theta_n \sim \pi(\theta|z_n)$. In fact, in Example 4.3, z_n is generated from a distribution of the form $f(z|\theta_n, z_{n-1})$, because the joint distribution of (y_i, g_i) is decomposed into two full conditional distributions for simulation reasons. A similar example is given in Robert, Celeux and Diebolt (1993) for *hidden Markov chains*. This model encompasses mixture models, but allows for a possible Markov dependence between the observations, $x_1, \ldots, x_T$, which can be described at the missing data level. The simulation of this missing data then gets too time-consuming to be operated directly and this imposes the following Gibbs decomposition:

[1.1] Simulate $z_{1n}|z_{2(n-1)}, \ldots, z_{T(n-1)}, \theta_{n-1}$;
...
[1.T] Simulate $z_{Tn}|z_{1n}, \ldots, z_{(T-1)n}, \theta_{n-1}$.

This type of decomposition implies that (θ_n) is not a Markov chain, but the finiteness of the state space of $z = (z_1, \ldots, z_T)$ and the irreducibility of the Markov chain (z_n) ensure that the central limit theorem holds in this setup. Other examples of a similar nature can be found in capture–recapture (George and Robert, 1992; Dupuis, 1995) and in image analysis and spatial statistics (Besag and Green, 1993); this variety of examples shows that the duality principle extends further than missing data structures. Another important application of the duality principle is the setting of *deconvolution problems*, where complex expressions involving recurrent sums can be simplified to usual densities by a call to artificial indicator variables. (See Robert, 1994, Example 1.16, for an illustration in the case of a nonparametric mixture of geometric random variables.)

5. CONVERGENCE MONITORING VIA MULTIPLE USES OF THE GIBBS SAMPLE

The previous developments focus on the central limit theorem, both in terms of assessment—which is simplified when some duality principle applies—and of estimation of the asymptotic variance—which can be brought back to an iid setup if some renewal features of the chain can be exhibited. Although relatively straightforward, these assessments still require a minute examination of the MCMC algorithm and of the resulting chain. Moreover, they do not always hold, as shown by Example 3.1. This section considers the convergence assessment from an alternative perspective, by proposing a rudimentary and informal stopping rule which operates in more general settings, but can be embedded in the previous machinery whenever it operates.

Given a MCMC sample $\theta_1, \ldots, \theta_N$ and a function of interest h, it is usually possible to compare the average (2.3) with other estimates of $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$. For instance, if there exists a dual chain $z_1, \ldots, z_N$ and if the conditional expectation $\mathbb{E}^{\tilde{\pi}}[h(\theta)|z]$ can be computed, the *Rao–Blackwellized average* (4.3) also converges to $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$. In other cases, for example, for the Metropolis algorithm, it is possible to condition on the previous value of the chain, θ_{n-1}, that is, to replace $h(\theta_n)$ in (2.3) by $\mathbb{E}[h(\theta_n)|\theta_{n-1}]$. Note that Rao–Blackwellization is not limited to exponential families and natural parameters since Casella and Robert (1996) have shown that it is always possible to compute a Rao–Blackwellized version of (2.3) for a Metropolis algorithm by integrating out the uniform random variables simulated at each step. A second alternative to (2.3) is to use classical Monte Carlo estimates based on $\theta_1, \ldots, \theta_N$. Gelfand and Sahu (1994) suggested *accept–reject* algorithms, but this approach requires discarding some of the θ_i's, while computing a maximum density ratio; we consider the standard technique based on *importance sampling*. In fact, as already mentioned in the literature (Rubin, 1987; Gelfand and Smith, 1990), this classical Monte Carlo method may improve upon the estimate (2.3). If the density $\tilde{\pi}$ is known up to a constant, the weighted sum

$$\text{(5.1)} \qquad \sum_{n=1}^{N} \omega_n h(\theta_n) \quad \text{with} \quad \omega_n \propto \frac{\tilde{\pi}(\theta_n)}{\pi^*(\theta_n)},$$

also converges to $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ when the θ_i's are simulated according to π^*, which may depend on the previous θ_n or z_n. Note in particular that the terms in (5.1) are uncorrelated since

$$\begin{aligned}
&\mathbb{E}^{\pi^*}[\omega_n h(\theta_n)\omega_m h(\theta_m)] \\
&\quad = \int h(\theta_n)\tilde{\pi}(\theta_n)h(\theta_m)\tilde{\pi}(\theta_m)\, d\theta_n\, d\theta_m \\
&\quad = \mathbb{E}^{\tilde{\pi}}[h(\theta_n)]\mathbb{E}^{\tilde{\pi}}[h(\theta_m)] \\
&\quad = \mathbb{E}^{\pi^*}[\omega_n h(\theta_n)]\mathbb{E}^{\pi^*}[\omega_n h(\theta_m)],
\end{aligned}$$

whatever the correlation induced by π^*. In Gibbs setups, π^* is given by (2.2), while $\tilde{\pi}$ is known up to a normalizing constant, from the Hammersley–Clifford theorem (see Besag, 1974, 1994). When the normalizing constant is unknown, the weights can be normalized by $\sum_n \omega_n = 1$. Due to the strong law of large numbers, (5.1) converges to $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ even when the variance is infinite. However, we will see

below that importance sampling does not perform well and should not be used as a convergence criterion in the latter case.

The third alternative we consider in this section is based on the trapezoidal approximation to

$$\int h(\theta)\tilde{\pi}(\theta)\,d\theta;$$

that is,

$$\sum_{n=1}^{N-1}(\theta_{(n+1)}-\theta_{(n)})h(\theta_{(n)})\tilde{\pi}(\theta_{(n)}),$$

where $\theta_{(i)}$ denotes the ith order statistic of the sample $\theta_1,\ldots,\theta_N$. This alternative is called *weighted Monte Carlo integration* in Yakowitz, Krimmel and Szidarovszky (1978). When the normalizing factor of $\tilde{\pi}$ is missing, a manageable version is to use

$$\frac{\sum_{n=1}^{N-1}(\theta_{(n+1)}-\theta_{(n)})h(\theta_{(n)})\tilde{\pi}(\theta_{(n)})}{\sum_{n=1}^{N-1}(\theta_{(n+1)}-\theta_{(n)})\tilde{\pi}(\theta_{(n)})}, \tag{5.2}$$

which converge to $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ when N goes to infinity (Philippe, 1996). Note that this option only applies in one dimension, because multidimensional extensions do not perform well (see Example 4.2).

These different approaches provide four possible estimates of $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ denoted by δ_e, δ_{rb}, δ_{is} and δ_w for ergodic, Rao–Blackwell, importance sampling and weighted Monte Carlo, respectively. A straightforward stopping rule is then to monitor the simultaneous convergence of the four estimates to the same quantity. This naive implementation is quite conservative since these estimates may be converging to $\mathbb{E}^{\tilde{\pi}}[h(\theta)]$ at very different speeds. Nonetheless, a theoretical comparison between them often depends on the setting [existence of a manageable Rao–Blackwellized version, finiteness of the variance factor $\mathbb{E}^{\tilde{\pi}}[h^2(\theta)\tilde{\pi}(\theta)/\pi^*(\theta)]$, behavior of the extreme order statistics $(\theta_{(1)}-\theta_{(0)})$ and $(\theta_{(N)}-\theta_{(N-1)})$, etc.]. Given that we are looking for general and robust stopping rules, it seems relevant to monitor the different convergence paths and wait until simultaneous convergence of δ_e and δ_w for instance. Obviously, this rule is not foolproof since two estimates can stabilize while the algorithm is still in the neighborhood of a mode of the posterior distribution. The same criticism applies to most stopping rules, though, and it is important to realize that alternatives to the usual empirical mean are usually available and sometimes perform better. We now examine through a few examples how this rudimentary comparison performs.

EXAMPLE 2.1 (Continued). Since x_n is generated according to

$$x_{n+1}=\begin{cases} x_n, & \text{with probability } 1-x_n,\\ y\sim\mathcal{B}e(\alpha+1,1), & \text{otherwise,}\end{cases} \tag{5.3}$$

we have

$$\begin{aligned}\mathbb{E}[x_{n+1}^{1-\alpha}|x_n] &= (1-x_n)x_n^{1-\alpha}+x_n\mathbb{E}[y^{1-\alpha}]\\ &= (1-x_n)x_n^{1-\alpha}+x_n(\alpha+1)\int_0^1 y^{1-\alpha+\alpha}\,dy\\ &= (1-x_n)x_n^{1-\alpha}+x_n(\alpha+1)/2,\end{aligned}$$

which leads to the following Rao–Blackwellized estimate of $\mathbb{E}^{\tilde{\pi}}[x^{1-\alpha}]$:

$$\delta_{rb}=(1/N)\sum_{n=1}^{N}\left\{(1-x_n)x_n^{1-\alpha}+x_n(\alpha+1)/2\right\}.$$

If we assume that the y's in (5.3) are generated at each iteration, the importance sampling estimate is based on a sample $y_1,\ldots,y_N$ of iid $\mathcal{B}e(\alpha+1,1)$ observations, instead of the original sample $x_1,\ldots,x_N$. In this case,

$$\delta_{is}=\sum_{n=1}^{N}y_n^{-\alpha}\Big/\sum_{n=1}^{N}y_n^{-1}$$

since the weight (5.1) satisfies $\omega_n\propto y_n^{\alpha-1}/y_n^{\alpha}=y_n^{-1}$. Note that δ_w can be constructed on either the x_n's sample or the y_n's sample. We select the second approach, with

$$\begin{aligned}\delta_w &= \sum_{n=1}^{N-1}(y_{(n+1)}-y_{(n)})\Big/\sum_{n=1}^{N-1}(y_{(n+1)}-y_{(n)})y_{(n)}^{\alpha-1}\\ &= (y_{(N)}-y_{(0)})\Big/\sum_{n=1}^{N-1}(y_{(n+1)}-y_{(n)})y_{(n)}^{\alpha-1},\end{aligned}$$

because the Metropolis sample leads to inefficiency, given that it induces null terms $(x_{(n+1)}-x_{(n)})$ in (5.2). (Note that the convergence of δ_w to the true value does not require the y_n's to be generated according to the correct distribution.) Figure 3 gives an illustration of the convergence paths of δ_e, δ_{rb}, δ_{is} and δ_w.

Given the artificial aspect of this example, the various estimates δ_{rb}, δ_{is} and δ_w are somehow contrived, but the comparison allowed by Figure 3 is still interesting. Two usual features are that the two estimates δ_e and δ_{rb} are quite similar almost from the start and that they are more unstable than the weighted Monte Carlo estimate δ_w. In general, δ_w provide a benchmark whenever available. In addition, when comparing with the two examples below (see Figures 4 and 5), the graph associated with

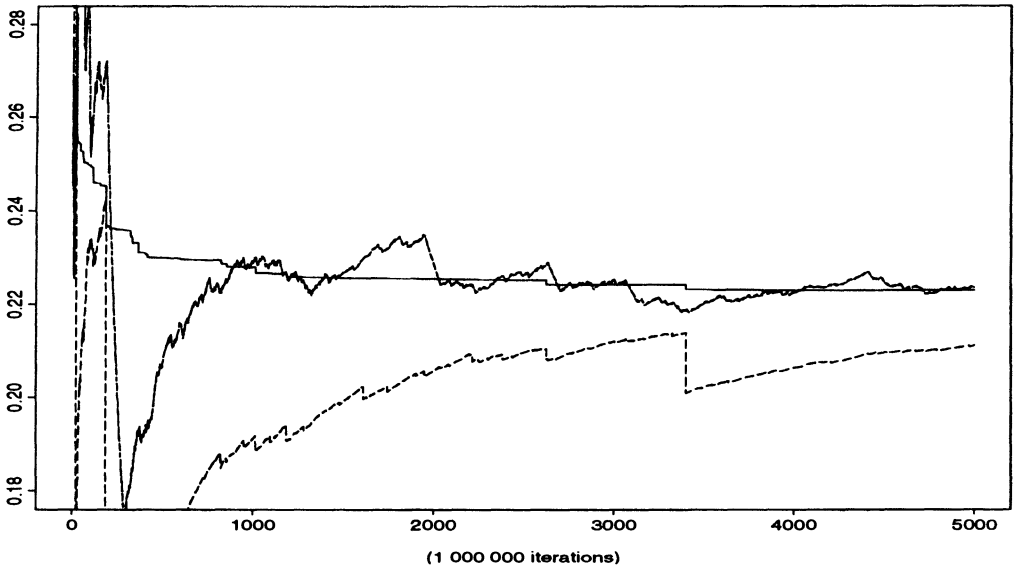

FIG. 3. *Convergence paths for different estimators of* $\mathbb{E}^{\tilde{\pi}}[x^{1-\alpha}]$: δ_e (*dots*), δ_{rb} (*dashes*), δ_{is} (*long dashes*) *and* δ_w (*plain*). *In this graph,* δ_e *and* δ_{rb} *are indistinguishable from the start.*

δ_w shows a singular lack of stability and thus indicates that convergence is not yet achieved. In this regard, the criterion is successful since all three estimates fail to provide a proper approximation of the true value, $\alpha = 0.2$ after one million iterations. The importance sampling estimate δ_{is} gets closer to 0.2, although it exhibits some important jumps due to huge values of the weights ω_n which are of such magnitude that the shape of the cumulated curve for δ_{is} implies that convergence is not yet attained for this estimate. While one should keep in mind this example was chosen for its pathological features (Robert, 1995, shows that 250 million iterations are necessary to achieve a correct evaluation of α by δ_e), its setting is rather favorable to importance sampling since the y_n's are simulated independently.

EXAMPLE 3.1 (Continued). In the more realistic setup of (3.7), the Rao–Blackwellized estimator of $\exp(-\theta/\sigma)$ is

$$\delta_{rb} = \frac{1}{N}\sum_{n=1}^{N} \exp\left\{\tau^2\left[-(\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3) + 1/2\right]\right\}$$

as shown by (3.9). Similarly, the importance sampling estimate is

$$\delta_{is} = \sum_{n=1}^{N} \omega_n \exp(-\theta_n/\sigma),$$

with

$$\omega_n \propto \tau[\exp\{-\theta_n^2/2\sigma^2 + (\theta_n - \tau^2(\eta_1 x_1 + \eta_2 x_2 + \eta_3 x_3))^2/2\tau^2\}] \cdot \left[\prod_{i=1}^{3}(1 + (x_i - \theta_n)^2)\right]^{-1},$$

while the weighted Monte Carlo estimate is

$$\delta_w = \left[\sum_{n=1}^{N-1}(\theta_{(n+1)} - \theta_{(n)})\exp(-\theta_{(n)}/\sigma - \theta_{(n)}^2/2\sigma^2) \cdot \prod_{i=1}^{3}[1 + (x_i - \theta_{(n)})^2]^{-1}\right] \cdot \left[\sum_{n=1}^{N-1}(\theta_{(n+1)} - \theta_{(n)})\exp(-\theta_{(n)}^2/2\sigma^2) \cdot \prod_{i=1}^{3}[1 + (x_i - \theta_{(n)})^2]^{-1}\right]^{-1}.$$

Figure 4 leads to different conclusions than the previous example, although δ_e and δ_{rb} are again quite indistinguishable. They both converge to the value obtained via δ_w, which is stable almost from the start, while δ_{is} is not converging to this quantity at the same speed, for reasons related to the lack of variance of the weights ω_n. It may well be that a phenomenon similar to Example 2.1 occurs in this case, namely, that δ_e, δ_{rb} and δ_w all fall far from the mark, while δ_{is} indicates the exact value, but the

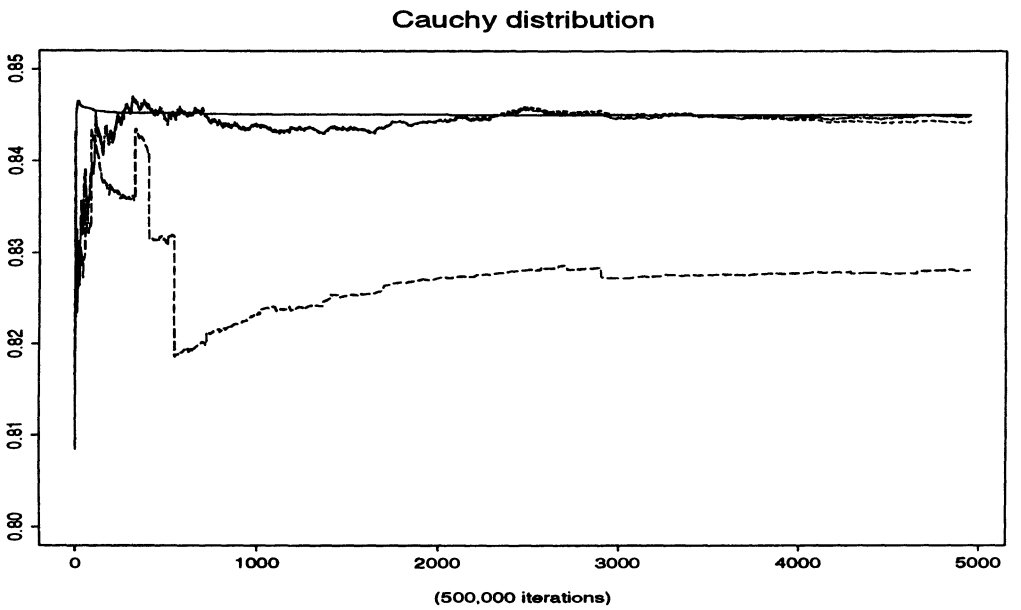

FIG. 4. *Convergence paths for different estimators of* $\mathbb{E}^{\tilde{\pi}}[\exp(-\theta/\sigma)]$: δ_e (*dots*), δ_{rb} (*dashes*), δ_{is}, (*long dashes*) *and* δ_w (*plain*). *In this graph, the estimates* δ_e *and* δ_{rb} *cannot be distinguished.*

similarity of the three estimates over a large number of iterations and the stability of their convergence path are quite reassuring factors to argue that the Bayes estimate of $\exp(-\theta/\sigma)$ should be close to 0.845.

EXAMPLE 4.2 (Continued). In this case, the function of interest is the odds ratio $h_3(\mu, \eta) = \mu/(1-\mu-\eta)$. While δ_{rb} is given by (4.4), the weights in δ_{is} are

$$\begin{aligned}\omega_n &\propto \left[(a_1\mu_n + b_1)^{x_1}(a_2\mu_n + b_2)^{x_2}(a_3\eta_n + b_3)^{x_3}\right.\\ &\qquad \left.\cdot (a_4\eta_n + b_4)^{x_4}(1-\mu_n-\eta_n)^{x_5-1/2}\right]\\ &\quad \cdot \left[\mu_n^{1/2}\eta_n^{1/2}(1-\mu_n-\eta_n)^{1/2}\mu_n^{z_1+z_2-1/2}\right.\\ &\qquad \left.\cdot \eta_n^{z_3+z_4-1/2}(1-\mu_n-\eta_n)^{x_5-1/2}\right]^{-1}\\ &= \left[(a_1\mu_n + b_1)^{x_1}(a_2\mu_n + b_2)^{x_2}\right.\\ &\quad \left.\cdot (a_3\eta_n + b_3)^{x_3}(a_4\eta_n + b_4)^{x_4}\right] \cdot \left[\mu_n^{z_1+z_2}\eta_n^{z_3+z_4}\right]^{-1}.\end{aligned}$$

The weighted Monte Carlo estimate cannot be constructed as above since we are in a bidimensional setup. A first solution is to extend (5.2) by considering trapezoidal approximations on squares $[\mu_{(n)}, \mu_{(n+1)}] \times [\eta_{(m)}, \eta_{(m+1)}]$, but Yakowitz, Krimmel and Szidarovszky (1978) have shown that the multidimensional versions of (5.2) are less efficient. In our case, it is actually possible to derive the marginal distribution of $\xi = h_3(\mu, \eta)$ by integrating out

$$\begin{aligned}\tilde{\pi}(\mu, \eta) &\propto \mu_n^{-1/2}\eta_n^{-1/2}(a_1\mu_n + b_1)^{x_1}(a_2\mu_n + b_2)^{x_2}\\ &\quad \cdot (a_3\eta_n + b_3)^{x_3}(a_4\eta_n + b_4)^{x_4}\\ &\quad \cdot (1-\mu_n-\eta_n)^{x_5-1/2}.\end{aligned}$$

Indeed,

$$\begin{aligned}\tilde{\pi}(\xi) &\propto \int_0^1 \left[\frac{\xi(1-\eta)}{1+\xi}\right]^{-1/2} \eta_n^{-1/2}\left(a_1\frac{\xi(1-\eta)}{1+\xi} + b_1\right)^{x_1}\\ &\quad \cdot \left(a_2\frac{\xi(1-\eta)}{1+\xi} + b_2\right)^{x_2}\\ &\quad \cdot (a_3\eta_n + b_3)^{x_3}(a_4\eta_n + b_4)^{x_4}\\ &\quad \cdot \left(\frac{1-\eta}{1+\xi}\right)^{x_5-1/2} \frac{1-\eta}{(1+\xi)^2}\, d\eta\\ &= \frac{\xi^{-1/2}}{(1+\xi)^{x_5+1}} \int_0^1 \eta_n^{-1/2}\left(a_1\frac{\xi(1-\eta)}{1+\xi} + b_1\right)^{x_1}\\ &\quad \cdot \left(a_2\frac{\xi(1-\eta)}{1+\xi} + b_2\right)^{x_2}\\ &\quad \cdot (a_3\eta_n + b_3)^{x_3}(a_4\eta_n + b_4)^{x_4}\\ &\quad \cdot (1-\eta)^{x_5}\, d\eta\end{aligned}$$

$$= \frac{\xi^{-1/2}}{(1+\xi)^{x_5+1}} \int_0^1 \sum_{i_1=1}^{x_1} \sum_{i_2=1}^{x_2} \sum_{i_3=1}^{x_3} \sum_{i_4=1}^{x_4} \binom{x_1}{i_1}\binom{x_2}{i_2} \cdot \binom{x_3}{i_3}\binom{x_4}{i_4} a_1^{i_1} b_1^{x_1-i_1} a_2^{i_2} b_2^{x_2-i_2} \cdot a_3^{i_3} b_3^{x_3-i_3} a_4^{i_4} b_4^{x_4-i_4} \left(\frac{\xi}{1+\xi}\right)^{i_1+i_2} \cdot (1-\eta)^{i_1+i_2+x_5} \eta^{i_3+i_4-1/2}\, d\eta,$$

and this density can be expressed as a polynomial in $\xi/(1+\xi)$, namely,

$$\sum_{j=0}^{x_1+x_2} \varpi_j \frac{\xi^{j-1/2}}{(1+\xi)^{j+x_5+1}} \tag{5.4}$$

with

$$\varpi_j \propto \sum_{i_1=0\vee j-x_2}^{j\wedge x_1} \binom{x_1}{i_1} a_1^{i_1} b_1^{x_1-i_1} \binom{x_2}{j-i_1} a_2^{j-i_1} b_2^{x_2-j-i_1} \cdot \sum_{k=1}^{x_3+x_4} \int_0^1 \eta^{k-1/2}(1-\eta)^{j+x_5} d\eta \cdot \sum_{i_3=0\vee k-x_4}^{k\wedge x_3} \binom{x_3}{i_3} a_3^{i_3} b_3^{x_3-i_3} \binom{x_4}{k-i_3} a_4^{k-i_3} b_4^{x_4-k-i_3}.$$

Since

$$\int_0^1 \eta^{k-1/2}(1-\eta)^{j+x_5} d\eta = \frac{\Gamma(k+1/2)\Gamma(j+x_5+1)}{\Gamma(j+x_5+k+3/2)},$$

the weights ϖ_j in (5.4) are proportional to

$$\sum_{i_1=0\vee j-x_2}^{j\wedge x_1} \binom{x_1}{i_1} \left(\frac{a_1}{b_1}\right)^{i_1} \binom{x_2}{j-i_1} \left(\frac{a_2}{b_2}\right)^{j-i_1} \cdot \sum_{k=1}^{x_3+x_4} \frac{\Gamma(k+1/2)\Gamma(j+x_5+1)}{\Gamma(j+x_5+k+3/2)} \cdot \sum_{i_3=0\vee k-x_4}^{k\wedge x_3} \binom{x_3}{i_3} \left(\frac{a_3}{b_3}\right)^{i_3} \binom{x_4}{k-i_3} \left(\frac{a_4}{b_4}\right)^{k-i_3}. \tag{5.5}$$

Figure 5 illustrates a possible behavior of δ_{is} when the weights have infinite variance. Although unbiased, as shown by the few intersections with the other paths, δ_{is} fluctuates too widely to be of any use in a convergence diagnosis or even in the estimation of $\mathbb{E}^{\tilde{\pi}}[\mu/(1-\mu-\eta)]$. Its erratic path and the huge discrepancies with the three other paths are enough to identify its lack of relevance on the spot, that is, during the simulation. The three other estimates converge to the true value, 0.747, which can be computed analytically from the weights (5.5). Note again the higher efficiency of δ_w which converges to the true value much faster than the two other estimates. This efficiency is obviously re-

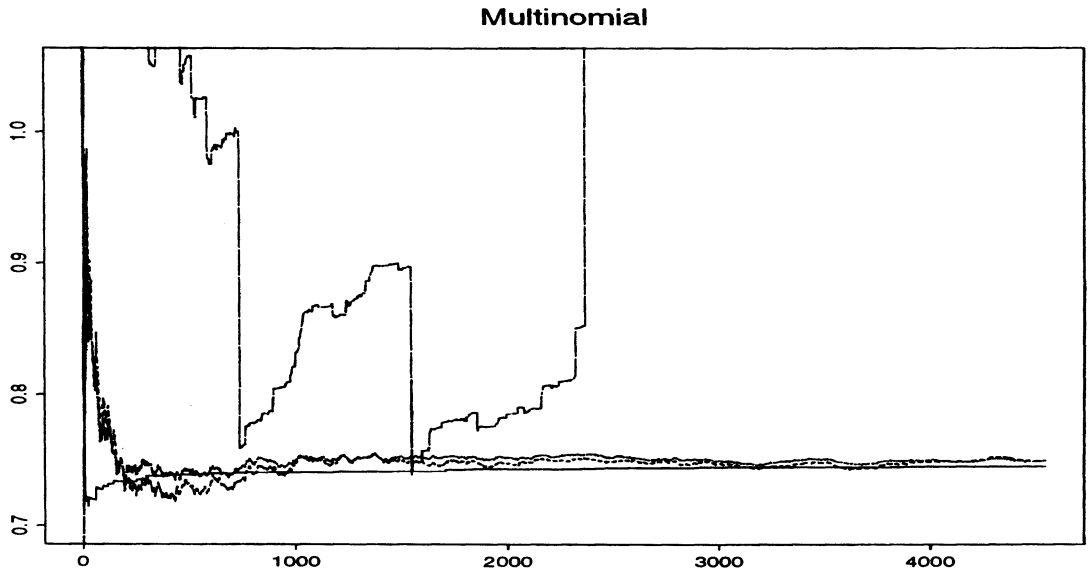

FIG. 5. *Convergence paths for different estimators of* $\mathbb{E}^{\tilde{\pi}}[\mu/(1-\mu-\eta)]$: δ_e *(dots),* δ_{rb} *(dashes),* δ_{is}*, (dots and dashes) and* δ_w *(plain). Both* δ_e *and* δ_{rb} *follow a similar path around the almost constant* δ_w*, while* δ_{is} *does not fit in the scale of the graph after* 24,000 *iterations. The curves are based on* 50,000 *iterations.*

lated to the extensive use of $\tilde{\pi}$, which is not always available.

6. CONCLUSION

This paper covers several topics pertaining to the improvement of Markov chain Monte Carlo techniques in terms of convergence control. The developments around the mixing properties of the chain are instrumental in assessing that the central limit theorem actually applies, but more advanced tools are necessary to approximate the limiting variance factor. For instance, renewal theory seems quite appropriate in this regard, since it allows for a classical iid setup and for regular estimators of the asymptotic variance. Further developments along this line are still necessary since the determination of the factors involved in the renewal process is quite problem dependent. Mykland, Tierney and Yu (1995) overcame that drawback by modifying the original kernel, but it may be possible to produce automated renewal versions for the original algorithms.

The duality principle appears as an important step toward a necessary automation, since finite state space and other simple Markov chains provide manageable settings for the derivation of renewal sets, while the dual structure motivating the principle occurs in a wide range of settings. Unfortunately, although it applies to a much more general setup than just data augmentation, there is no immediate extension of this principle to hierarchical simulation structures such as those appearing for Gibbs sampling, since every sub-chain generated by the algorithm is not a Markov chain per se. It goes without saying that a technique should not be recommended only because theoretical results are at our disposal. However, strongly correlated structures such as those appearing in highly dimensional Gibbs algorithms are usually slower to converge than faster mixing techniques like some alternative Metropolis algorithms where the duality principle may apply (Geyer, 1992; Besag et al., 1995) and theoretical evaluations are urgently needed to back up this intuition.

When renewal theory applies in an MCMC setup, the central limit theorem or even the law of the iterated logarithm are some tools available to control convergence. Further investigation should not be dismissed, however, because the applicability of these results is not always guaranteed and also because they assume that stationarity is already attained. While this assumption has little bearing in low-dimensional problems, it seems more difficult to retain it in large dimensions, and the future direction for research should be the incorporation in the convergence diagnosis of the study of multiple Markov chains generated from the same MCMC algorithm. As *splitting* is the technique behind renewal theory, *coupling* (see Lindvall, 1992, and Meyn and Tweedie, 1993) should presumably be used for the control of multiple run MCMC methods because the interchange of the various chains running in parallel should drastically increase the mixing of the chains, that is, the lack of dependence on the starting points of the chain. However, coupling is usually associated with the existence of a small set A, where the coupling process occurs (see Asmussen, 1979, and Meyn and Tweedie, 1993) and should thus be preceded by a determination of A, as in renewal theory. See Johnson (1994) and Propp and Wilson (1995) for some steps in this direction, where discrete models are again easier to manage.

The last section, although more empirical and less theoretical than the previous ones, also has some possible bearing on the control of MCMC methods. In its current version, the method is mainly graphical and does not allow for easy extensions to multidimensional setups. Moreover, the stopping rule is subjective, given that some estimates like δ_{is} are often eliminated for erratic behavior. The main lesson in our examples is thus that weighted Monte Carlo estimates should be used as a benchmark whenever the distribution $\tilde{\pi}$ is available in closed form, since they are more stable and accurate than the ergodic or Rao–Blackwellized averages, while these two estimates are too similar to control convergence. On the contrary, importance sampling estimates have shown a strong propensity to err far away from the true value of the posterior expectation and one should exercise great caution when using them. Some developments are nonetheless possible toward a stabilization of the importance sampling estimates (mixed proposal distributions, trimming, accept–reject, ...) as well as a generalization of weighted Monte Carlo estimates to more general setting (estimated Rao–Blackwellization, iterative marginalization, ...), and these topics currently under study should obviously benefit the control of MCMC methods and therefore their dissemination to more complex settings.

ACKNOWLEDGMENTS

The author is grateful to M. Broniatowski, F. Charlot and J. Diebolt for helpful discussions and to K. Mengersen and R. Tweedie for discussions as well as for their hospitality in Fort Collins, Colorado, in July 1993. Comments from an Editor and a referee were instrumental in clarifying style, exposition and focus.

REFERENCES

ASMUSSEN, S. (1979). *Applied Probability and Queues*. Wiley, New York.

BESAG, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–326.

BESAG, J. (1994). Discussion of "Markov chains for exploring posterior distributions" by L. Tierney *Ann. Statist.* **22** 1734–1741.

BESAG, J. and GREEN, P. J. (1993). Spatial statistics and Bayesian computation. *J. Roy. Statist. Soc. Ser. B* **55** 25–38.

BESAG, J., GREEN, P. J., HIGDON, D. and MENGERSEN, K. L. M. (1995). Bayesian computation and stochastic systems, (with discussion). *Statist. Sci.* **10** 3–66.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

BROOKS, S. and ROBERTS, G. O. (1995). Diagnosing convergence of Markov chain Monte-Carlo algorithms. Technical Report 95-12, Statistical Laboratory, Univ. Cambridge.

CASELLA, G. and GEORGE, E. I. (1992). An introduction to Gibbs sampling. *Amer. Statist.* **46** 167–174.

CASELLA, G. and ROBERT, C. P. (1996). Rao–Blackwellisation of sampling schemes. *Biometrika* **83**. To appear.

CHAN, K. S. and GEYER, C. J. (1994). Discussion of "Markov chains for exploring posterior distributions" by L. Tierney *Ann. Statist.* **22** 1747–1758.

CHIB, S. and GREENBERG, E. (1995). Understanding the Metropolis–Hastings algorithm. *Amer. Statist.* **49** 327–335.

COWLES, M. K. and CARLIN, B. P. (1995). Markov chain Monte-Carlo convergence diagnostics: a comparative study. Technical Report, Univ. Minnesota.

DAVYDOV, Y. A. (1973). Mixing conditions for Markov chains. *Theory Probab. Appl.* **28** 312–328.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Roy. Statist. Soc. Ser. B* **56** 363–375.

DOUKHAN, P., MASSART, P. and RIO, E. (1994). The functional central limit theorem for strongly mixing processes. *Ann. Inst. H. Poincaré Probab. Statist.* **30** 63–82.

DUPUIS, J. A. (1995). Bayesian estimation of movement probabilities in open populations using hidden Markov chains. *Biometrika* **82** 761–772.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**. Wiley, New York.

GELFAND, A. E. and SAHU, S. K. (1994). On Markov chain Monte-Carlo acceleration. *Journal of Computational and Graphical Statistics* **3** 261–276.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GELMAN, A., GILKS, W. R. and ROBERTS, G. O. (1994). Efficient Metropolis jumping rules. Research Report 94–10, Statistics Laboratory, Univ. Cambridge.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7** 457–511.

GEORGE, E. I. and ROBERT, C. P. (1992). Calculating Bayes estimates for capture–recapture models. *Biometrika* **79** 677–683.

GEWEKE, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics 4* (J. O. Berger, J. M. Bernardo, A. P. Dawid and A. F. M. Smith, eds.) 169–194. Oxford Univ. Press.

GEYER, C. J. (1992). Practical Markov chain Monte-Carlo (with discussion). *Statist. Sci.* **7** 473–511.

GILKS, W., CLAYTON, D. G., SPIEGELHALTER, D. I., BEST, N. G., SHARPLES, L. D. and KIRBY, A. J. (1993). Modelling complexity: applications of Gibbs sampling in medicine (with discussion). *J. Roy. Statist. Soc. Ser. B* **55** 39–52.

HEITJAN, D. F. and RUBIN, D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19** 2244–2253.

HOBERT, J. P. and CASELLA, G. (1996). Gibbs sampling with improper distributions. *J. Amer. Statist. Assoc.* To appear.

JOHNSON, V. E. (1994). Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. ISDS Working Paper 94–07, Duke Univ.

KIPNIS, C. and VARADHAN, S. R. (1986). Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Comm. Math. Phys.* **104** 1–19.

LINDVALL, T. (1992). *Lectures on Coupling Theory*. Wiley, New York.

LIU, J. S., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and sampling schemes. *Biometrika* **81** 27–40.

LIU, J. S., WONG, W. H. and KONG, A. (1995). Correlation structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57** 157–169.

MALINOVSKII, V. K. (1987). Limit theorems for Harris Markov chains. *Theory Probab. Appl.* **31** 269–285.

MENGERSEN, K. L. and ROBERT, C. P. (1995). Testing for mixtures: a Bayesian entropic approach (with discussion). In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 255–276. Oxford Univ. Press.

MENGERSEN, K. L. and TWEEDIE, R. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.* **24** 101–121.

METROPOLIS, N., ROSENBLUTH, A. W., ROSENBLUTH, M. N., TELLER, A. H. and TELLER, E. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21** 1087–1092.

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

MEYN, S. P. and TWEEDIE, R. L. (1994). Computable bounds for convergence rates of Markov chains. *Ann. Appl. Probab.* **4** 124–148.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–241.

PELIGRAD, M. (1986). Recent advances in the central limit theorem and its weak invariance principle for mixing sequences of random variables. In *Dependence in Probability and Statistics* (E. Ebberlein and M. Taqqu, eds.) 192–223. Birkhäuser, Boston.

PHILIPPE, A. (1996). Processing simulation output by Riemann sums. Technical Report, 96–02, Univ. Rouen.

POLSON, N. G. (1995). Convergence of Markov chain Monte-Carlo algorithms. In *Bayesian Statistics 5* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford Univ. Press.

PROPP and WILSON (1995). Exact sampling with coupled Markov chains and applications to statistical mechanics. Technical Report, Dept. Mathematics, MIT.

ROBERT, C. P. (1994). *The Bayesian Choice: A Decision-Theoretic Motivation*. Springer, New York.

ROBERT, C. P. (1995). A pathological MCMC algorithm and its use as a benchmark for convergence assessment techniques. Doctoral work, CREST, INSEE, Paris.

ROBERT, C. P. (1996). *Méthodes de Simulation en Statistique: Une Introduction aux Méthodes de Monte-Carlo par Chaînes*

de Markov. Economica, Paris. To appear.

ROBERT, C. P., CELEUX, G. and DIEBOLT, J. (1993). Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statist. Probab. Lett.* **16** 77–83.

ROBERT, C. P. and TITTERINGTON, M. (1996). Reparametrising schemes for hidden Markov models and their application for maximum likelihood estimation. Technical Report, Dept. Statistics, Univ. Glasgow.

ROBERTS, G. O. and TWEEDIE, R. L. (1994) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. Research Report 94–9, Statistics Laboratory, Cambridge Univ.

ROSENBLATT, M. (1971). *Markov Processes. Structure and Asymptotic Behavior.* Springer, New York.

ROSENTHAL, J. S. (1993). Rates of convergence for data augmentation on finite sample spaces. *Ann. Appl. Probab.* **3** 819–839.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

SCHERVISH, M. J. and CARLIN, B. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **2** 111–122.

SMITH, A. F. M. and ROBERTS, G. O. (1993). Bayesian computation via Gibbs and related Markov Chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **55** 3–24.

TANNER, M. and WONG, W. (1987). The calculation of posterior distributions (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

TIERNEY, L. (1991). Markov chains for exploring posterior distributions. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (E. Kerimidas, ed.) 563–570. Interface Foundation of North America, Fairfax Station, VA.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

YAKOWITZ, S., KRIMMEL, J. E. and SZIDAROVSZKY, F. (1978). Weighted Monte Carlo integration. *SIAM J. Numer. Anal.* **15** 1289–1300.

YU, B. and MYKLAND, P. (1994). Looking at Markov samplers through cusum path plots: a simple diagnostic idea. Technical Report 9413, Dept. Statistics, Univ. California, Berkeley.