CHRISTIAN P. ROBERT

Université de Rouen and Université Pierre et Marie Curie

Statistics is concerned not only with the derivation of optimal procedures but also with the computation of these procedures. It is thus quite timely that *The Annals of Statistics* address this issue and the editors are to be congratulated for opening a tribune for a discussion about the pros and cons of Markov chain Monte Carlo methods. Moreover, most papers in the recent literature on Gibbs sampling have mainly focused on implementation aspects and on the width of the application range, with only marginal attention to probabilistic justifications and convergence problems. Tierney's paper is thus most welcomed since it highlights the probabilistic foundations of Markov chain Monte Carlo methods, exposes the various types of convergence and more generally introduces us to tools we were maybe hardly aware of but which are of major interest for the sound use, improvement and control of Gibbs simulation techniques.

Nonetheless, further progress is still necessary in this direction, as statistics is once more in an *inverse position* in regards to probability. Indeed, in most Markov chain Monte Carlo settings, we are faced with Markov chains which are ergodic with a known invariant probability distribution and we want to assess the rate of convergence and the behavior of the simulated sequence, while classical Markov chain results are usually dealing with the derivation of recurrence, ergodicity and convergence properties in terms of the transition kernel. Therefore, there is a need for adapting Markov chain theory to the specific setting of Markov chain Monte Carlo simulation.

Another imperative in the development of convergence diagnoses is *simplicity*. In fact, Markov chain Monte Carlo methods have been introduced in Tanner and Wong (1987) and Gelfand and Smith (1990) as powerful alternatives to numerical integration or optimization techniques and to analytical approximations such as Laplace's. Therefore, there is a danger in the current trend, namely an increased complexity of the proposed implementations of Markov chain Monte Carlo methods, since the additional efficiency is counterbalanced by a difficulty of implementation and involved preliminary studies. And, while Tierney's review quite usefully recalls the convergence results available in the literature, it also exposes the lack of more automatic characterizations of the properties of simulated Markov chains, which would be necessary for a safer and wider application of Gibbs methods. For instance, the verification of the minorization condition in Section 3.2 may involve a complex study of the Markov chain, which is prohibitive for most users. The following comments will thus focus on convergence issues in this spirit, that is, aiming at a limited involvement of the user.

1. Mixing and the central limit theorem. Tierney mentions *uniform ergodicity* and *ergodicity of degree* 2 as sufficient conditions for the CLT to

apply. Unfortunately, these two properties are quite difficult to check without a detailed study of the particular chain at hand. However, there exist alternative conditions for the CLT to apply, based on the *mixing* properties of the chain, that is, on the degree of dependency between the X_n's.

1.1 A first mixing property, *α-mixing*, that is, the fact that

$$\alpha(n) = \sup_{A,B} |P(X_n \in A, X_0 \in B) - P(X_n \in A)P(X_0 \in B)|$$

goes to 0 as n goes to ∞, is one of the weakest measures of asymptotic independence. However, it may be enough for normal convergence: if f is a measurable function such that $\mathbb{E}^\pi[|f(X)|^\gamma] < +\infty$, $\gamma > 2$, a sufficient condition for the sum of the $f(X_n)$'s, S_n, to be asymptotically normal is that

$$\sum_n \alpha(n)^{(\gamma-2)/\gamma} < +\infty \tag{1.1}$$

[see Davydov (1973)]. A somewhat more manageable condition is recalled in Peligrad (1986), namely that when f is $L^2(\pi)$, it is sufficient that

$$\limsup_n \sigma_n / \mathbb{E}[|S_n|] < \sqrt{\pi/2}, \tag{1.2}$$

where σ_n^2 is the variance of S_n; in fact, the law of large numbers provides convergent approximations of σ_n and of $\mathbb{E}[|S_n|]$ and therefore allows for an "on-line" verification of (1.2) as the Markov chain (X_n) gets simulated.

Obviously, α-mixing itself has to be established for the Markov chain at hand, but it actually holds for most Markov chains induced by Gibbs sampling and other Markov chain Monte Carlo methods, since every positive recurrent aperiodic Markov chain is α-mixing [Rosenblatt (1971), page 200]. In addition, α-mixing is induced by other kinds of mixing, like β-,ϱ-,φ- and ψ-mixing (see below). Moreover, the last three types of mixing are necessarily exponential when they hold for Markov chains [Bradley (1986), page 175]; this implies that (1.1) is necessarily satisfied whatever γ is. Another sufficient condition for α-mixing is the *Doeblin condition* [Revuz (1975)], equivalent to φ-mixing under Harris recurrence and satisfied when the transition operator has a strictly positive density with respect to the invariant probability π [Davydov (1973)].

1.2 The β-mixing condition is rather awkward to express (and even more to check directly) since it involves

$$\beta(n) = \sup_{(A_i)} \sup_{(B_j)} \sum_{i,j} |P(X_n \in A_i, X_0 \in B_j) - P(X_n \in A_i)P(X_0 \in B_j)|,$$

where the supremum is taken over all couples of partitions; under β-mixing, $\beta(n)$ converges to 0 as n goes to ∞. However, a Markov chain is β-mixing when it is Harris recurrent and the density of the Markov transition probability

with respect to the invariant probability π is positive [Davydov (1973)]. Berbee (1979), page 102, also shows that φ-*mixing* implies β-mixing. The interest of this type of mixing appears in Section 2 below in connection with coupling theory.

1.3 A mixing condition which is enough by itself to ensure the CLT to hold is φ-*mixing*, that is, the case when

$$\varphi(n) = \sup_{A,B} \left| P\left(X_n \in A \,|\, X_0 \in B\right) - P(X_n \in A) \right|$$

goes to 0 as n goes to ∞. In fact, for every f in $L^2(\pi)$ such that $\mathbb{E}[f(X_0)] = 0$, the series

$$\sigma_f^2 = \mathbb{E}\left[f(X_0)^2\right] + 2\sum_{k=1}^{\infty} \mathbb{E}\left[f(X_0)f(X_k)\right]$$

is absolutely convergent and, if $\sigma_f > 0$, the CLT applies to the average of the $f(X_n)$'s, the limiting distribution being $\mathcal{N}(0, \sigma_f^2)$ [Billingsley (1968), Theorem 20.1]. Therefore, a simple monitoring of the estimators of σ_f^2 can show whether the CLT applies or not [see Geyer (1991)]. And the assessment of φ-mixing is often straightforward: all finite and most compact state irreducible Markov chains are φ-mixing [Billingsley (1968)] as well as Doeblin irreducible Markov chains [Davydov (1973)].

1.4 Another type of mixing which is of interest is ϱ-*mixing*, defined as the convergence of

$$\varrho(n) = \sup_{f,g \in L^2(\pi)} \operatorname{Corr}\left(f(X_n), g(X_0)\right)$$

to 0 when n goes to ∞. Indeed, in this case, the CLT holds under ϱ-mixing either if $f \in L^2(\pi)$ and $\sigma_n^2 = \mathbb{E}[|f(X_1) + \cdots + f(X_n)|^2] \approx n\sigma^2$ or if $\mathbb{E}[|f(X_1)|^l] < \infty$ for $l > 2$ and σ_n goes to ∞ [Rosenblatt (1971)].

2. Renewal processes and i.i.d. replacements. Another attack on the dependency problem is to try to replace the Markov chain produced by the sampling method with an "equivalent independent sequence," instead of checking the above mixing conditions. This technique is sometimes called *Berstein's method* in the literature [Peligrad (1986)].

2.1. A first approach is to introduce a sequence of *hitting times* (τ_j), defined by the successive return of the Markov chain to a particular set A. The Markov sequence $X_n, \ldots, X_{n+m}$ is then divided into blocks associated with these hitting times, $Y_j = (X_{\tau_j+1}, \ldots, X_{\tau_{j+1}})$. Under some conditions on A, it can be shown that the renewal sequences Y_j are independent or rather 2-independent (namely that the sequences Y_{2j} are independent) and identically distributed. For instance, this is the case when A is an *atom*, that is, when the transition probability

satisfies $P(x, \cdot) = \nu$ for every x in A, where ν is a probability measure [see Charlot (1991) and Meyn and Tweedie (1992)]. And atoms do exist for all Harris recurrent Markov chains [Revuz (1975)]. More generally, such settings can be created by perturbating the initial chain when $X_n \in A$ in the following way:

$$X_{n+1} = \begin{cases} Y \sim \nu(y), & \text{with probability } \alpha, \\ Z \sim \dfrac{P(X_n, z) - \alpha\nu(z)}{1 - \alpha}, & \text{otherwise.} \end{cases}$$

In fact, as shown in Asmussen (1979), every *ν-irreducible Markov chain* allows for the existence of $r \in \mathbb{N}$, $0 < \varepsilon < 1$, a set A and a probability measure ν such that

$$P^r(x, E) \geq \varepsilon\nu(E)$$

for every $x \in A$ and every set E, that is, a *minorization condition* as in Tierney's paper. The introduction of independent sequences derived from the original chain is obviously of capital interest since the averages

$$\overline{X}_j = \sum_{n=\tau_j+1}^{\tau_{j+1}} X_n$$

are also independent and identically distributed, thus leading to a CLT when the associated variance is finite. Malinovskii (1986, 1989) presents some detailed central limit approximations and large deviations results for the sums S_n. Once again, the main problem with this technique is to determine the values of A, ν and r.

2.2 Another technique is to replace the stationary chain (X_n) by another sequence (Y_n) such that the Y_i's, $i \geq n$, are independent of the X_i's, $i \leq n$, with the same distribution and there exists a.s. an n_0 such that $X_i = Y_i$ for $i \geq n_0$. This replacement is known as "coupling theory" and is particularly interesting under β-mixing since, as shown in Berbee (1979), page 106, it is equivalent to β-mixing. Moreover, there exists a version of (Y_n) such that

$$P(\exists k \geq n \text{ such that } X_k \neq Y_k) = \beta(n).$$

Now, defining $\beta = \beta(1)$, this result implies that, for a β-mixing chain with rate β, a sample of size N can be somehow replaced by an i.i.d. sample of size $(1 - \beta)N$. This solves at once the CLT and the "batch size" question, the latter being repeatedly considered in the Gibbs sampling literature [see, e.g., Geyer (1991), Raftery and Lewis (1992) or Tierney's paper]. Indeed, it is then sufficient to eliminate $\beta\%$ of the chain to get a subsequence which behaves on the average like an i.i.d. chain. However, a quite difficult and still open question is to estimate β.

3. The duality principle. When dealing with some Markov chain Monte-Carlo approaches in finite mixture estimation, Diebolt and Robert (1994) obtain convergence results and CLTs by using a *duality principle* [see also Diebolt and Robert (1992)]. This principle is actually applicable in most *missing data* settings and works for cases when the chain of interest (X_n) is associated with a secondary (or dual) chain (z_n) such that X_n is generated according to a conditional distribution $f(x \mid z_n)$. We assume in addition that the dual chain (z_n) has an invariant probability distribution $g(z)$ such that the distribution of interest, π, is the marginal distribution of the invariant probability distribution of (X_n, z_n), namely $\pi(X_n, z_n) = f(X_n \mid z_n)g(z_n)$. The duality principle borrows strength from the simplest chain (z_n) in the following sense.

THEOREM 1. *If the Markov chain (z_n) is ergodic (resp. geometrically ergodic with rate ϱ), the chain (X_n) is also ergodic (resp. geometrically ergodic with rate ϱ). Moreover, if the support of z_n is compact and if (z_n) is φ-mixing (resp. ϱ-mixing), X_n is also φ-mixing (resp. ϱ-mixing).*

In many settings, the chain (z_n) can be easily studied and it is straightforward to derive its convergence and mixing properties. For instance, in many *missing data problems*, z_n has a finite state space and is automatically φ-mixing under ergodicity and irreducibility. Tanner and Wong (1987) provide several examples of such cases, including mixture estimation, censoring and missing value problems. See also Heitjan and Rubin (1991) and Gelfand, Smith and Lee (1992) for additional examples. Robert, Celeux and Diebolt (1993) take advantage of the duality principle to establish convergence results for Gibbs estimation of *hidden Markov chain models.* In addition, the rate of φ-mixing, that is, ϱ such that $\varphi(n) \leq C\varrho^n$, can be estimated for finite state space Markov chains by

$$\varrho \leq 1 - \min_{ij} p_{ij}$$

if this quantity is not 1 (otherwise, p_{ij} should be replaced by p_{ij}^t with t large enough). This can be of major interest in the control of the convergence of the Gibbs sampler. As a last remark, note that the duality principle can be related to the *interleaving property* introduced in Liu, Wong and Kong (1991) which justifies in particular Rao-Blackwellization.

Acknowledgment. The author is grateful to François Charlot and Jean Diebolt for helpful comments.

REFERENCES

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

BERBEE, H. (1979). *Random Walks with Stationary Increments and Renewal Theory. Math. Centre Tract* **112**. Math. Centrum, Amsterdam.

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

BRADLEY, R. C. (1986). Basic properties of strong mixing conditions. In *Dependence in Probability and Statistics* (E. Ebberlein and M. Taqqu, eds.) 165–192. Birkhäuser, Boston.

CHARLOT, F. (1991). Régénération, chaînes de Markov et files d'attente. *Séminaire de Mathématique de Rouen* **2** 143–154.

DAVYDOV, Y. A. (1973). Mixing conditions for Markov chains. *Theory Probab. Appl.* **18** 312–328.

DIEBOLT, J. and ROBERT, C. P. (1992). Discussion of Smith and Roberts and Besag and Green's papers. *J. Roy. Statist. Soc. Ser. B* To appear.

DIEBOLT, J. and ROBERT, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Roy. Statist. Soc. Ser. B.* **56** 363–376.

GELFAND, A. E., SMITH, A. F. M. and LEE, T. M. (1992). Bayesian analysis of constrained parameter and truncated data problems. *J. Amer. Statist. Assoc.* **87** 523–532.

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

MALINOVSKII, V. K. (1986). Limit theorems for Harris Markov chains. I. *Theory Probab. Appl.* **31** 269–285.

MALINOVSKII, V. K. (1989). Limit theorems for Harris Markov chains. II. *Theory Probab. Appl.* **34** 252–265.

MEYN, S. P. and TWEEDIE, R. L. (1992). Stability of Markovian processes. I. *Adv. in Appl. Probab.* **24** 542–574.

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.

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

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

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

LABORATOIRE DE STATISTIQUE THÉORETIQUE
ET APPLIQUÉE
CNRS–URA 1321
UNIVERSITÉ PIERRE ET MARIE CURIE,
PARIS VI
TOUR 45–55, BOÎTE 158
4, PLACE JUSSIEU
75252 PARIS CEDEX 05
FRANCE

KUNG SIK CHAN[1] AND CHARLES J. GEYER[2]

University of Iowa and University of Minnesota

We congratulate Luke Tierney for this paper, which even before its appearance has done a valuable service in clarifying both theory and practice in this important area. For example, the discussion of combining strategies in Section 2.4 helped researchers break away from pure Gibbs sampling in 1991; it was, for example, part of the reasoning that lead to the "Metropolis-coupled" scheme of Geyer (1991) mentioned at the end of Section 2.3.3.

Harris Recurrence. The discussion of Harris recurrence in Section 3.1 has been very helpful. Harris recurrence essentially says that there is no measure-

[1]Supported in part by NSF Grant DMS-91-18626.
[2]Supported in part by NSF Grant DMS-90-07833.