Comment

Nicholas G. Polson

The goal of Markov chain Monte Carlo (MCMC) algorithms is to provide answers in multidimensional statistical models that are computationally quicker than other techniques such as importance sampling or numerical integration. However, not all MCMC algorithms lead to procedures that are more efficient than these other techniques. Both Geyer, and Gelman and Rubin propose empirical approaches to assess when a particular MCMC procedure is useful. Unfortunately, I am skeptical about the potential for any empirical diagnostics in the MCMC setting.

Two desirable properties for a MCMC procedure are: (1) to provide estimators that, with arbitrarily high probability, approximate the quantity of interest to any specified level of accuracy and (2) to perform this task "quickly," say in polynomial time. Unfortunately, the two procedures presented cannot establish these properties. Geyer proposes one long run of the chain together with a time-series analysis of the output, whereas Gelman and Rubin propose multiple runs using an overdispersed initial distribution together with a diagnostic approach to stop the chain. I will show that when (1) and (2) are satisfied there is no need to "diagnose" convergence or perform a time-series analysis of the chain. One long run of the chain is sufficient and run lengths can be bounded using the second eigenvalue of the Markov chain. These bounds are stronger than those obtained using central limit theorem arguments. On the other hand, when (2) does not hold, it is theoretically unclear whether the output has any informational content and whether diagnostics, multiple runs or time-series analyses of the chain can help solve (1). In this discussion I will focus on techniques for checking (2) for any MCMC procedure.

In the following I will discuss several topics related to properties (1) and (2): fast convergence of the chain, selection of the "burn-in" period, how long to run the chain and caveats associated with a purely diagnostic approach. To introduce notation and fix ideas, consider a time reversible ergodic Markov chain defined on a finite state space, V. Let π be the distribution that we wish to sample from and h: $V \to \Re$ be the functional whose expectation under π, $E_\pi(h)$, is the quantity of interest. Imagine V, as the computer does, to be a fine discretisation of the k dimensional parameter space. I define "quick" algorithms as those that satisfy (1) in $o(|V|)$ rather than $O(|V|)$ operations [importance sampling and numerical integration are $O(|V|)$]. This criterion is central to the issue of why randomised algorithms can be more powerful than deterministic algorithms. Basically, MCMC algorithms meeting this criterion are superior to numerical integration strategies and are called provably convergent.

Let P denote the transition matrix of the chain designed with π as the unique stationary distribution. The efficiency of the algorithm, as we will see, depends on the rate of convergence of the chain, which in turn depends crucially on the second eigenvalue, λ_1. One long run of the chain turns out to be sufficient to generate samples from π and to draw inferences about $E_\pi(h)$. I will show that there is no need to "diagnose" convergence or perform a time-series analysis of the chain *as long as* an a priori bound on λ_1 is available.

To assess the efficiency of a MCMC procedure one may proceed as follows: by Perron-Frobenius theory, the L^1 distance between the distribution after t steps of the chain, $P^t(\varphi)$, and the stationary distribution, $\pi(\varphi)$, is geometrically bounded as

$$\sum_{\varphi} | P^t(\varphi) - \pi(\varphi) | \leq \frac{\lambda_1^t}{\sqrt{\pi(\varphi_0)}},$$

where φ_0 is the initial starting point of the chain and the negative eigenvalues are assumed to be bounded below by $-\lambda_1$ [see, e.g., Diaconis and Stroock (1991)]. Therefore, one can achieve a desired level of sampling accuracy, ε, by running the chain for $T = \log(\varepsilon\sqrt{\pi(\varphi_0)})/\log(1/\lambda_1)$ steps. At first sight this is appealing and looks straightforward to implement and there is no need to "diagnose" convergence from the realised chain. However, for the algorithm to be computationally efficient we need T to be small relative to $|V|$. More precisely we need $T = o(|V|)$. Notice that since $|V|$ is exponential (in k), this essentially requires T to be polynomial (in k).

Demonstrating that a MCMC algorithm is provably (polynomial time) convergent can be a difficult problem. Several papers describe techniques for checking fast convergence of a MCMC algorithm. Applegate, Kannan and Polson (1990) provide a bound for T for Gibbs and Metropolis algorithms by using the notion of conductance to obtain a bound for λ_1 and hence T. Conductance (Sinclair and Jerrum, 1989) is widely used in computer science and has the following intuitive definition: the chain will converge rapidly if the escape probability for each subset S of states is high, as measured by $\Sigma_{\theta\in S,\varphi\notin S}\, \pi(\theta)P(\theta,\varphi)/\Sigma_{\theta\in S}\, \pi(\theta)$, where $P(\theta,\varphi)$

Nicholas G. Polson is Assistant Professor, Graduate School of Business, University of Chicago, Chicago, Illinois 60637.

is the transition probability of going from θ to φ. The conductance of the whole chain, Φ, is the minimum over all subsets such that $\pi(S) \leq 1/2$. Conductance provides an upper bound on the second eigenvalue due to the inequality $\lambda_1 \leq 1 - \Phi^2$. Diaconis and Stroock (1991) discuss other ways of finding tighter bounds for λ_1; for example, using the Poincaré inequality.

MCMC algorithms that are not provably convergent are problematic due to slow convergence. Heuristically, slow convergence will occur if there exists a subset of states with poor conductance as the chain can get stuck in this set for long periods. If this is the case, then *any* diagnostic procedure will give an overconfident assessment of convergence as a set with poor conductance can be "hidden" from the chain, such as, in the witch's hat distribution. The problem can be more serious for multimodal distributions. Here convergence can "appear" to be very quick when in fact the chain mixes poorly. Multiple-run procedures might alleviate this problem, but proper application requires substantial knowledge of the underlying distribution, such as the location of modes and low conductance sets. The analyses of Geyer, and Gelman and Rubin on chains with slow convergence are largely unexplored. Both of these issues are areas for further research.

The time-to-stationarity, T, for any joint distribution π can be bounded by supposing that the experimenter has specified a measure of local curvature, a, and a global curvature measure β of log-concavity of π. Then, given ε, the chain is run for

$$O\left(k^2 d^2 a^2 e^{2\beta}\left(\log\left(\frac{a^k}{\varepsilon}\right) + \log\left(\frac{a^k}{\pi(\varphi_0)}\right)\right)\right)$$

steps, where V is assumed to be contained in a cube of size d and $k = \dim(\theta)$ (Applegate, Kannan and Polson, 1990). For log-concave distributions, corresponding to $\beta = 0$, the Gibbs and Metropolis algorithms satisfy (2); they perform (1) in $o(|V|)$ steps and as such are (random) polynomial time convergent. This is why randomized algorithms like Gibbs sampling are powerful computational tools in high dimensions, whereas other Monte Carlo procedures or numerical integration are prone to the curse of dimensionality.

I now turn to the problem of using MCMC algorithms to draw inferences for $E_\pi(h)$ once provable convergence is established. Specifically, consider the class of estimators of $E_\pi(h)$ are constructed by deleting the first N_0 draws and then averaging the next N_1 values of $h(\cdot)$ at the Markov chain draws, $\theta^{(i)}$, where $N_0 \leq i \leq N_0 + N_1$. That is,

$$\hat{\theta}_{N_0,N_1} = \frac{1}{N_1}\sum_{i=N_0+1}^{N_0+N_1} h(\theta^{(i)}).$$

Notice that, as Geyer points out, it is inefficient not to take every draw. An analysis based on Aldous (1987) provides a bound on the mean-square error (MSE) of $\hat{\theta}_{N_0,N_1}$ as follows:

$$MSE(\hat{\theta}_{N_0,N_1}) \leq \frac{C(\lambda_1,N_1)\sigma^2}{N_1}\left(1 + \frac{\lambda_1^{N_0}}{\sqrt{\pi(\varphi_0)\pi^*}}\right),$$

where $\sigma^2 = Var_\pi(h)$ and $\pi^* = \min \pi$. The constant $C(\lambda_1, N_1) \leq 2$ if $N_1 \gg 1/\log(1/\lambda_1)$, and $MSE(\hat{\theta}_{N_0,N_1})$ can be made arbitrarily small *as long as*

$$N_0 \gg \frac{\log(1/\sqrt{\pi(\varphi_0)\pi^*})}{\log(1/\lambda_1)} \quad \text{and} \quad N_1 \gg \frac{1}{\sigma^2 \log(1/\lambda_1)}.$$

The bound on N_0 is related to that on T and is more stringent than that on N_1. Therefore, when there is fast convergence, the experimenter can guarantee provable estimates of $E_\pi(h)$ in $o(|V|)$ steps without diagnostics or time-series analysis. The current practice for selecting "burn-in" values using *ad hoc* rules or diagnostics tend to give values for N_0 that are extremely small and misleading. This is of a more serious nature than indicated by Geyer.

The above reasoning also provides the following guidelines: N_0 reduces bias in $\hat{\theta}_{N_0,N_1}$, N_1 decreases the MSE($\hat{\theta}$) and it suggests adopting a starting point near the mode. An additional consequence is that N_1 depends on λ_1 (and in implementation on the bound for λ_1), rather than on an estimated value from the observed chain that might underestimate N_1. When the chain converges quickly, the bound is $o(|V|)$. In contrast, assessing N_1 purely using a central limit theorem argument produces on a bound of order $O(|V|)$ (that is, exponential in k) (Aldous, 1987). Therefore, the central limit theorem analysis provides a bound no better than those in importance sampling in terms of computational complexity.

Two related issues concerning λ_1 should be noted. First, as a caveat, eigenvalue bounds can be pessimistic, and the actual convergence of the L^1 distance is not always an exponential decay (as suggested by the upper bound) but rather drops sharply after a suitable number of steps (Diaconis, 1988). Second, one practical solution to poor convergence properties of a given chain is to introduce auxiliary variables. For example, the Swendsen-Wang algorithm for the Ising model can exhibit polynomial time convergence (Jerrum and Sinclair, 1990), whereas ordinary Gibbs sampling procedures can be nonpolynomial time convergent. For the latter, any diagnostic procedure or multiple-run procedure to assess convergence is prone to failure.

Hopefully, future research will develop methodologies that combine the procedures presented by Geyer, and Gelman and Rubin with the theoretical considerations discussed here.