Monte Carlo methods: an overview [1]

Peter J. Green, University of Bristol

It is often the case that a researcher, seeking to quantify the uncertainty in some system, has to resort to simulation. Thus, for example, in operational research, simulation may be used to study the behaviour of a queuing system, or in statistics it may be used to assess the performance of a proposed new estimator. Although snide remarks of the "last resort of a scoundrel" kind are sometimes made by theoretically-oriented mathematicians about simulation, it is a well-understood and respectable approach, at least when simulation experiments are well designed and analysed. A good modern treatment of the subject can be found in Ripley (1987). The main attraction of simulation over analytic or conventional numerical methods is its flexibility and adaptability: for the queue, an analytic answer may be available for a standard case, and limit theorems may provide heavy traffic approximations, but for (essentially) exact solutions for arbitrary functionals of the process, or to deal with non-Poisson arrivals, or transient behaviour, simulation would probably be needed.

Viewed somewhat abstractly, simulation is concerned with estimating expectations with respect to a prescribed multivariate distribution, which we shall denote by $\{\pi(x) : x \in \mathcal{X}\}$. (It will be immaterial here whether the distribution is discrete or continuous: to fix ideas, let us assume discrete.) The expectations of interest might be probabilities, or vectors of probabilities, and hence histograms can be estimated. The distribution may be "prescribed" in various ways, often not explicitly as a properly normalised algebraic expression. Ordinary simulation deals with the case where, having factorised $\pi(x)$ sequentially as

$$\pi(x) = \prod_{i=1}^{n} \pi(x_i \mid x_{<i})$$

(where $x_{<i} = \{x_j \ : \ j < i\}$), the conditional densities $\pi(x_i \mid x_{<i})$ are available for simulation; it may have been necessary to re-order the components $x_1, x_2 \ldots x_n$ first. This factorisation reduces multivariate simulation to a series of univariate simulations, which are the basic tasks that random number generators perform. But what can we do if there is *no* re-ordering for which the one-sided conditional densities $\pi(x_i \mid x_{<i})$ are available, in this sense? It is routinely the case in statistical physics, spatial statistics, statistical image analysis and general Bayesian inference (to cite some key application areas, listed in historical order) that it is the *undirectional* conditional densities $\pi(x_i|x_{-i})$, where $x_{-i} = \{x_j : j \neq i\}$ that are available. These are usually known as *full conditionals* or *local characteristics*; the construction from these of the one-sided conditionals usually requires computation of a difficulty approaching that of the original problem.

This is the generic problem addressed by Markov chain Monte Carlo (MCMC) methods, and this is the explanation for the dramatic recent development of such methods in complex stochastic systems of the kinds just mentioned. Instead of producing a sequence of independent realisations x from $\pi(x)$, a MCMC method produces a single realisation $x^{(1)}, x^{(2)} \ldots x^{(m)}$ of a Markov chain whose limiting distribution is $\pi(x)$. At the price of the serial dependence in the realisation is achieved the freedom of being able to work only with the full conditionals, and thus to be able to update the variables in turn rather than all at once.

The issues that will be mentioned in this overview are the construction of MCMC methods, the assessment of simulation errors, some of the strategies for combating difficulties such as

[1] Invited contribution to Proceedings of the IMA Conference on Complex Stochastic Systems and Engineering, Leeds, September 1993.

excessive autocorrelation, and some of the novel types of inference and types of application for which MCMC can be used.

A major related area that will not be covered is the role played by simulation in optimisation. In broad terms, this role can be subdivided into that of dealing with problems where the objective function involves intractable integrals, and that of controlling stochastic search. Under the first of these headings, we note that most MCMC methods can be adapted to perform simulated annealing (as in Geman and Geman, 1984, for example), while Geyer and Thompson (1992) developed the application of MCMC to complicated likelihood calculations with unavailable normalising constants and/or missing data. One example of optimisation that uses stochastic search is the genetic algorithm, discussed in the abstract in this volume by Jennison, Franconi and Sheehan.

Construction

To obtain a Markov chain P with a specified "target" distribution π as its limiting distribution, we need first π to be an equilibrium distribution for P, and secondly P to be irreducible and aperiodic. In practice the first of these requirements is usually met by the stronger assumption of *detailed balance*

$$\pi(x)\ P(x \to x') = \pi(x')\ P(x' \to x)$$

for all x, x', with which it is usually easier to work. The two standard methods for updating just some of the variables in x, x_A say, in accordance with detailed balance both produce transition kernels $P_A(x \to x')$ say. The Gibbs sampler (Geman and Geman, 1984; from Grenander, 1983, and known earlier by other names in statistical physics) replaces x_A by new value(s) drawn from the full conditional $\pi(x_A \mid x_{-A})$. The Metropolis-Hastings method (Metropolis *et al*, 1953; Hastings, 1970) is a two-stage approach, which first draws a proposal x'_A from a rather arbitrary distribution $q(x'_A; x)$ indexed by the current state x, then accepts it with probability

$$\alpha = \min\left\{1,\ \frac{\pi(x'_A \mid x_{-A})\ q(x_A; x')}{\pi(x_A \mid x_{-A})\ q(x'_A; x)}\right\}.$$

It is easy to verify that Gibbs is a special case of Hastings, and that Hastings satisfies detailed balance.

Unless A is the whole of $\{1, 2 \dots n\}$, P_A cannot be irreducible; there are various ways of combining P_A for different subsets A covering $\{1, 2 \dots n\}$ that potentially provide irreducibility and aperiodicity including mixing, cycling, random permutations, etc. See Tierney (1994) and Besag, Green *et al*, (1994, Section 2) for thorough discussion of methods of construction for MCMC.

Assessing and reducing Monte Carlo variances

The obvious estimate of $E_\pi(f) = \sum_x f(x)\pi(x)$ is the empirical average $\bar{f}_m = \frac{1}{m}\sum_{t=1}^m f(x^{(t)})$. In practice, the chain is usually run for some time (the *burn-in* period) before collecting such statistics. The arguments against sub-sampling the realisation to reduce dependence, and in favour of one long run rather than several short ones, are well put by Geyer (1993). The only argument for sub-sampling, unless f is expensive to compute, applies when it is desired to store the entire realisation to allow later decision about which expectations are to be estimated. In this case the appropriate frequency of sub-sampling is determined solely by the time and space limitations.

There is no point to the assessment of the variance of $\bar{f}_m$ unless it is first established that the bias is negligible, that is, that $f(x^{(t)})$ has effectively converged. Diagnostics for convergence are being developed, but in a sense there is an unattainable goal here, as by definition diagnostics cannot comment on parts of the distribution $\pi(x)$ "not yet seen". It is important to realise that the power of MCMC computation does not obviate the need for a thorough qualitative understanding of the target distribution π. This understanding is a prerequisite for design of reliable MCMC procedures, and it is only in this context that diagnostics can be trusted.

Given convergence, estimates of the variance of $\bar{f}_m$ can be obtained by blocking, by time series methods (both requiring only stationarity, not the Markov property, and avoiding explicit estimation of autocorrelations) or from a consideration of convexity properties of the autocorrelation function. See Hastings (1970), Green and Han (1992) and Geyer (1993), respectively. Green and Han (1992) includes a general discussion of convergence and efficiency in terms of the spectrum of P, and proposes methods analogous to the use of antithetic variables for exploiting the eigenstructure to reduce $\text{var}(\bar{f}_m)$.

A very appealing route to avoiding the quantification of serial dependence in the chain is *regenerative simulation* (Mykland, Tierney and Yu, 1993). By exploiting a technique due to Nummelin, it is possible to split a Markov chain realisation into i.i.d. *tours* of random length. Expectations can then be estimated by averaging across tours, and variances of such estimates obtained by ratio estimation formulae. The attraction of this approach is presently limited somewhat by the awkwardness of finding splits in a complicated Markov chain. Further, of course, many of the concerns raised above remain. Regeneration will not help the chain visit the entire state space freely; it just gives a different basis for variance estimation and some useful diagnostics if it does.

Returning to the central question of convergence, there has been a good deal of work in recent years in the probability and theoretical computer science literatures on analytic techniques for establishing bounds on rates of convergence. Some very subtle and elegant approaches have been discovered, typified by the article by Jerrum in this volume, and quite a lot is now known rigorously, at least for certain highly-symmetric combinatorial problems. However, it seems fair to say that present techniques fall short of providing good non-pessimistic estimates of rates for the chains currently used in Bayesian computation. Further, the criteria for convergence addressed in current theoretical work, total variation norm for example, seem more stringent than necessary to evaluate the convergence of one or two scalar functionals of interest.

Strategies for improving convergence and efficiency

The canonical example for illustrating the problems that can arise with single variable updating in MCMC is a mixture of two bivariate normal distributions, displaced diagonally from each other. As the distance between the two components increases, swaps from one to the other become increasingly rare, if they occur at all in the finite realisation observed. Yet mathematically the Gibbs sampler and other standard single-variable updating schemes are irreducible.

The example is a prototype for many of the problems where MCMC works poorly. Multimodality can occur in Bayesian inference whenever prior and likelihood are strong but in disagreement (although one must distinguish here between peaks in density and modes containing substantial probability). In spatial statistics and statistical physics, MCMC has difficulty in the presence of strong spatial interaction: in the latter it is associated with phase transition.

Two general strategies, both originating in statistical physics and sometimes used together offer some prospect of being developed to combat such problem – multigrid and auxiliary variables. With multigrid methods (see Sokal, 1989, for example) the size of the sets of variables that are simultaneously updated is varied systematically from small to large and back again, usually according to a recursive 'V' or 'W' schedule. (The letter pictures the graph of the size of these updating sets against time.) In the spatial context, these sets of variables would be physically contiguous areas of the field of view.

In auxiliary variables methods (Swendsen and Wang, 1987; Edwards and Sokal, 1988; Besag and Green, 1993), the original variables x are augmented with additional variables u, with $\pi(u \mid x)$ specified. Simulation proceeds usually by updating x and u alternately, by Gibbs sampling or otherwise, and, with careful choice, the result mixes much more rapidly. The marginal for x is unchanged, and hence so is the extraction of information from the simulation. Auxiliary variables and multigrid can be used in combination, as in a block-restricted version of Swendsen-Wang with which the author has experimented, and in the stochastically-blocked version of Kandel, Domany and Brandt (1989).

Novel applications and opportunities for inference

Some of the most innovative opportunities created by MCMC-based computation are only available if the whole realisation, or a large sub-sample from it, can be stored for later post-processing. It is not always possible to identify in advance which probabilities and expectations are of interest. Good examples arise in ranking and selection problems. In Bayesian analysis of variety trials, for example (Besag and Higdon, 1994; Besag, Green *et al*, 1994), it may be of interest to identify that smallest subset of the varieties that includes the best-yielding variety with specified posterior probability. Such a subset would be very awkward or impossible to compute, even under distributional approximations, by conventional computational methods, but given a simulated sample, only routine enumeration is involved.

Another interesting example is that of the estimation of simultaneous credibility bounds valid uniformly for a whole function, surface or image. Again, see Besag, Green *et al*, (1994) for details.

When the target distribution does not have a density with respect to a simple dominating measure in a parameter space of fixed dimension, it is not clear how to define MCMC methods. Grenander and Miller (1994) proposed a class of *jump-diffusion* processes to deal with this problem, especially as arising in vision applications. Green (1994), in work now being further developed to deal with Bayesian inference with step-functions (such as in change-point problems and in image segmentation), generalised such processes to a class making *reversible jumps* under exact detailed balance, in discrete time.

Finally, a topic of considerable current interest is *simulated tempering*. As originally conceived by Marinari and Parisi (1992), this addresses the problem of dealing with a parameterised target distribution $\pi(x|\beta)$, for a parameter value β at which standard MCMC methods produce highly correlated realisations. The idea is to mix over a range of β, by treating β as another dynamic variable, and to design a Markov chain aimed at the target $\sum \pi(x|\beta)\pi_0(\beta)$. Here $\pi_0(\beta)$ is arbitrary, but usually taken as uniform on a finite set. Properties of the original target distribution can then easily be obtained by conditioning. The idea is applied to a genetics problem by Geyer and Thompson (1994). From a statistical perspective, this is an entirely natural hierarchical Bayes procedure, with $\pi_0(\beta)$ the hyperprior for β, no longer arbitrary. In spatial problems, $\pi(x|\beta)$ often involves an unknown normalising constant, depending on β, which is in principle needed to conduct the joint simulation of x and β. However, unpublished work by the author demonstrates that such normalising constants can be learnt on the fly during the simulation.

References

Besag, J. E. and Green, P. J. (1993). Spatial statistics and Bayesian computation (with Discussion). *Journal of the Royal Statistical Society* B, **55**, 25–37.

Besag, J. E. and Higdon, D. M. (1993). Bayesian inference for agricultural field experiments. *Bulletin of the International Statistical Institute*, **LV**, 121–136.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1994). Spatial statistics, image analysis and Bayesian inference, *University of Bristol Mathematics Research Report S-93-04*, tentatively accepted by *Statistical Science.*

Edwards, R. G. and Sokal, A. D. (1988). Generalization of the Fortuin-Kasteleyn-Swendsen-Wang representation and Monte Carlo algorithm. *Physical Review* D, **38**, 2009–2012.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geyer, C. J. (1993). Practical Markov chain Monte Carlo (with Discussion). *Statistical Science*, **8**, 473–483.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion), *J. Roy. Statist. Soc.* (B), **54**, 657–699.

Geyer, C. J. and Thompson, E. A. (1994). *Annealing Markov chain Monte Carlo with applications to ancestral inference.* Technical report, School of Statistics, University of Minnesota.

Green, P. J. (1994). Contribution to the discussion of paper by Grenander and Miller (1994). *Journal of the Royal Statistical Society* B, **56**, 000–000.

Green, P. J. and Han, X-L. (1992). Metropolis methods, Gaussian proposals, and antithetic variables. In *Stochastic models, statistical methods and algorithms in image analysis* (P. Barone, A. Frigessi and M. Piccioni, eds.), *Lecture Notes in Statistics*, **74**, 142–164. Berlin: Springer.

Grenander, U. (1983). *Tutorial in pattern theory.* Division of Applied Mathematics, Brown University, Providence.

Grenander, U. and Miller, M. (1994). Representations of knowledge in complex systems (with Discussion). *Journal of the Royal Statistical Society*, B, **56**, 000–000. To appear.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Kandel, D., Domany, E. and Brandt, A. (1989). Simulations without critical slowing down: Ising and three-state Potts models. *Physical Review* B, **40**, 330–344.

Marinari, E. and Parisi, G. (1992). Simulated tempering: a new Monte Carlo Scheme. *Europhys. Lett.*, **19** 451–458.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Mykland, P., Tierney, L. and Yu, B. (1993). Regeneration in Markov chain samplers. *Journal of the American Statistical Association.* To appear.

Ripley, B. D. (1987). *Stochastic Simulation.* New York: Wiley.

Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande.* Lausanne.

Swendsen, R. H. and Wang, J. S. (1987). Nonuniversal critical dynamics in Monte Carlo simulations. *Physical Review Letters*, **58**, 86–88.

Tierney, L. (1994). Markov chains for exploring posterior distributions, *Annals of Statistics.* To appear.