Iterative and Non-iterative Simulation Algorithms

Andrew Gelman*
Department of Statistics
University of California
Berkeley, CA 94720

Abstract

The Gibbs sampler, Metropolis' algorithm, and similar iterative simulation methods are related to rejection sampling and importance sampling, two methods which have been traditionally thought of as non-iterative. We explore connections between importance sampling, iterative simulation, and importance-weighted resampling (SIR), and present new algorithms that combine aspects of importance sampling, Metropolis' algorithm, and the Gibbs sampler.

1. Introduction

Currently, one of the most active topics in statistical computation is inference from iterative simulation, especially the Metropolis algorithm and the Gibbs sampler (Metropolis and Ulam, 1949; Metropolis et al., 1953; Hastings, 1970; Geman and Geman, 1984; Gelfand et al., 1990). (The Gibbs sampler is in fact a special case of the generalized Metropolis algorithm; see Section 4.3 below.) The essential idea of iterative simulation is to draw values of a random variable x from a sequence of distributions that converge, as iterations continue, to the desired *target distribution* of x. For inference about x, iterative simulation is typically less efficient than direct simulation, which is simply drawing from the target distribution, but iterative simulation is applicable in a much wider range of cases, as current statistical literature makes abundantly clear (see, e.g., Smith and Roberts, 1993, Besag and Green, 1993, and Gilks et al., 1993).

This paper presents iterative simulation methods as an outgrowth of the non-iterative methods of rejection and importance sampling, both of which use simulation to correct an approximation of the target distribution.[1] The connection between iterative and non-iterative simulation is of interest for three reasons. First, a unified treatment is appealing, and highlights the similar problems faced by users of all of these methods. Second, the general formulation suggests potentially useful new methods, such as the iterative importance resampling of Section 3.3 and the Metropolis-approximate Gibbs sampler of Section 4.4. Third, noniterative simulation can be important for obtaining starting distributions for iterative algorithms, as discussed in Gelman and Rubin (1993).

*To appear in *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface*. Thanks to Donald B. Rubin for helpful comments and the National Science Foundation for financial support.

[1]Another related and important topic, which we will not discuss here, is analytic approximations to integrals; see Tierney and Kadane (1986) and Morris (1988).

2. Normal-based Inference

2.1. Modes, Standard Errors, and the Normal Approximation

A point estimate and its associated standard error (or, more generally, its variance-covariance matrix), are motivated, explicitly or implicitly, by the normal approximation. Typically, the mean of the normal approximation is set equal to the mode (i.e., the maximum likelihood estimate or the posterior mode), and the inverse variance matrix is approximated by the negative of the second derivative matrix of the log posterior distribution at the mode. Computing these can be difficult in highly multivariate problems. Just finding the mode can require iteration, with Newton's method and EM (Dempster, Laird, and Rubin, 1977) being popular choices for common statistical models. Estimates of the "variance matrix" can be computed by analytic differentiation, numerical differention, or combined methods such as SEM (Meng and Rubin, 1991).

2.2. Approximation Using a Mixture of Normals

When the distribution is multimodal, it is necessary to run an iterative mode-finder several times, starting from different points, in an attempt to find all the modes. This strategy is also sensible and commonly used if the distribution is complicated enough that it *may* be multimodal. Once all the modes are found (possibly a difficult task) and the second derivative matrix calculated at each mode, the target distribution can be approximated by a mixture of k multivariate normals, each with its

own mode μ_k and variance matrix Σ_k; that is, the target density $P(x)$ is approximated by

$$\tilde{P}(x) = \sum_{k=1}^{K} \frac{\omega_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right),$$

where K is the number of modes, d is the dimension of x, and ω_k is the mass of the k-th component of the multivariate normal mixture. The masses ω_k can be calculated by equating the approximate density $\tilde{P}$ to the exact density P at the k modes, so that $\tilde{P}(\mu_k) = P(\mu_k)$, for $k = 1, \ldots, K$. Assuming the modes are well-separated, this implies that for each k, the mass ω_k is roughly proportional to $|\Sigma_k|^{1/2} P(\mu_k)$.

2.3. Student-t Approximation

In general, one can replace the normal approximation by a multivariate t, with the same center and scale, but wider tails. A mixture of Student-t densities with η degrees of freedom has density

$$\tilde{P}(x) \propto \sum_{k=1}^{K} \frac{\omega_k}{|\Sigma_k|^{1/2}} \left(\eta + (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right)^{-(d+\eta)/2} \tag{2.1}$$

and can be simulated by first drawing from the normal mixture of Section 2.2 and then dividing the sampled vector by a χ^2_η random deviate, divided by η. Because of its wide tails (and that it can be easily simulated and its density function is easy to calculate), the multivariate t will turn out to be useful as a starting distribution for the exact simulation methods described below.

3. Using Analytic Approximations and Importance Weights to Obtain Exact Simulations

3.1. Rejection Sampling

A simple way to draw samples from a target distribution P, using an approximate starting distribution P_0, is *rejection sampling*, which requires the ability to calculate $P(x)/P_0(x)$, up to a proportionality constant, for all x. We will label $w(x) \propto P(x)/P_0(x)$, the *importance ratio* of x. In addition, rejection sampling requires a known constant M that is no less than $\sup w(x)$. The algorithm proceeds in two steps:

1. Sample x at random from $P_0(x)$.
2. With probability $\frac{w(x)}{M}$, *reject* x and return to step 1; otherwise, keep x.

An accepted x has the correct distribution $P(x)$; that is, the conditional distribution of drawn x, given it is accepted, is $P(x)$.

The above steps can be repeated to obtain additional independent samples from P. Rejection sampling cannot be used if no finite value of M exists, which will happen when P_0 has lighter tails than P, as when the support of P_0 is smaller than the support of P. (Hence the use of a multivariate t, instead of a normal, for a starting ditsribution.) In practice, when P_0 is not a good approximation to P, the required M will be so large that almost all samples obtained in step 1 will be rejected in step 2. The virtue of rejection sampling as an iterative simulation method is that it is self-monitoring—if the simulation is not working, you will know it, because no simulated draws will be accepted.

3.2. Approximate Rejection Sampling Using Importance Ratios

When no bound on $w(x)$ is known, rejection sampling is impossible. However, one can still draw samples $x^{(1)}, \ldots, x^{(N)}$ from $P_0(x)$, and calculate their importance ratios, $w(x^{(l)})$, for $l = 1, \ldots, N$, known only up to a proportionality constant. The method of *importance weighting* seeks to adjust the sampling by using $w(x^{(l)})$ to weight each random sample, $x^{(l)}$. Instead of discarding samples, those values $x^{(l)}$ with low importance ratios are just downweighted.

For any finite N, importance weighting gives only approximate results; it can thus be thought of as an iterative simulation method, improving as N increases. For importance weighting to be effective, the starting distribution P_0 should cover the target distribution P, in the sense that the importance ratios should not get too high. Even if importance ratios are unbounded, the method can still be useful—in contrast to rejection sampling—but the large values should be rare with respect to the target distribution.

Importance weights can be used to get a sequence of draws that approximately follow the target distribution by using the method of *importance resampling* (called SIR for "sampling-importance resampling" in Rubin, 1987, 1988). If N draws from the approximate starting distribution P_0 have been created, a sample of $n < N$ draws from a better approximation can be simulated as follows.

1. Sample a value x from the set $\{x^{(1)}, \ldots, x^{(N)}\}$, where the probability of sampling each $x^{(l)}$ is proportional to the weight, $w(x^{(l)})$.
2. Sample a second value using the same procedure, but excluding the already-sampled value from the set.
3. Repeatedly sample without replacement $n-2$ more times.

3.3. Iterative Importance Resampling

For any fixed N, importance resampling yields draws from an approximation to the target distribution. We can allow the approximation to improve in a smooth way as $n \to \infty$ by simply increasing N as n increases, that is, by expanding the pool of candidates $(x^{(1)}, \ldots, x^{(N)})$ as more values are subsampled (without replacement) with probabilities proportional to their importance ratios; for example, N could increase as the square of n.

The simulation procedure thus becomes *iterative*: at each time $t = 1, 2, \ldots$, a single draw x_t is taken from the set $(x^{(1)}, \ldots, x^{(N)})$, with probabilities of sampling proportional to importance weights. The set $(x^{(1)}, \ldots, x^{(N)})$ is created by supplementing the set of previously unsampled draws at time $t-1$ with new independent draws from the approximate distribution P_0. (We use the notation t for samples taken in succession, as opposed to n, the number of values in the final sample.)

If the importance ratios $P_0(x)/P(x)$ are bounded, the distribution of the samples x_t converge to the target distribution as $t \to \infty$.[2] The importance ratios will be bounded if the starting distribution P_0 has support at all the modes of interest in the target distribution P and has at least as heavy tails. We then say that the starting distribution is *overdispersed*, which is desirable.

The sequence $x_1, x_2, \ldots$ can be thought of as (dependent) draws from successively improving approximate distributions $P_1, P_2, \ldots$ that form a transition from the starting distribution P_0 toward the target distribution P. This is a conceptual improvement upon the basic attack of importance resampling, which provided no intermediate steps between the starting and target distributions. An obvious limitation, however, is that for all t, the supplemental draws are from P_0, which may be a much less accurate approximation to P than that afforded by P_{t-1}. Section 4 reviews Markov chain methods, which modify the drawing distribution as t increases. Related ideas connecting importance weighting to iterative simulation appear in Tanner and Wong (1987), Gelfand and Smith (1990), and Kong, Liu, and Wong (1991).

[2] With unbounded importance ratios, the simulations may still converge to the target distribution. In general, the distributions of the resampled draws depends on the rate of increase of the population sample size N. Determining the necessary and sufficient conditions for convergence of importance resampling is a difficult problem not addressed in this paper.

4. A Review of Markov Chain Methods for Exact Simulation

4.1. Why is Markov Chain Simulation Needed?

Markov chain methods are especially desirable when no starting distribution is available that is accurate enough to produce useful importance weights. If the starting distribution is not close, the importance weights will be so variable that, for reasonable values of n and N, the set of draws from importance resampling will be a poor approximation to the target distribution. In order to correct the defects of the drawing distribution, P_0, we must rely on a very large N.

In contrast, with any starting distribution that even loosely covers the target distribution, the steps of a Markov chain simulation directly improve the approximate distributions from which samples are drawn. Thus, the distributions, used for taking each draw, themselves converge to P as t increases. In a wide range of practical cases, it turns out that the iterations of a Markov chain simulation allow accurate inference from starting distributions that are much too vague for useful results from rejection or importance resampling. See Tierney (1991) for a unifying overview of many Markov simulation methods and Gelfand and Smith (1990) for an example in which importance resampling compares unfavorably to the Gibbs sampler.

4.2. The Method of Metropolis and its Generalizations

Given a target distribution $P(x)$, the generalized Metropolis algorithm (Hastings, 1970) draws a sequence of random points $(x^{(1)}, x^{(2)}, \ldots)$ whose distributions converge to the target distribution. The sequence $(x^{(t)})$ may be considered a random walk whose stationary distribution is $P(x)$. The algorithm proceeds as follows:

1. Draw a starting point $x^{(0)}$, for which $P(x^{(0)}) > 0$, from a *starting distribution* $P_0(x)$.
2. For $t = 1, 2, \ldots$:
   (a) At iteration t, take as input the point $x^{(t-1)}$.
   (b) Sample a candidate point $\tilde{x}$ from a *jumping distribution* at time t, $J_t(\tilde{x}|x^{(t-1)})$.
   (c) Calculate the ratio of importance ratios,

$$r = \frac{P(\tilde{x})}{P(x^{(t-1)})} \frac{J_t(x^{(t-1)}|\tilde{x})}{J_t(\tilde{x}|x^{(t-1)})}.$$

   (r is always defined, because a jump from $x^{(t-1)}$ to $\tilde{x}$ can only occur if both $P(x^{(t-1)})$ and $J_t(\tilde{x}|x^{(t-1)})$ are nonzero.)

(d) Set

$$x^{(t)} = \begin{cases} \tilde{x} & \text{with probability } \min(r,1) \\ x^{(t-1)} & \text{otherwise.} \end{cases}$$

This method requires the calculation of the relative importance ratios $P(x)/J_t(x|x')$ for all x, x', and an ability to draw x from the jumping distribution $J_t(x|x')$ for all x' and t.

The proof that the iteration converges to the target distribution has two steps: first, it is shown that the simulated sequence $(x^{(t)})$ is a Markov chain with a unique stationary distribution, and second, it is shown that the stationary distribution equals the target distribution. The first step of the proof holds if the Markov chain is irreducible, aperiodic, and not transient (see, e.g., Feller, 1968). Except for trivial exceptions, the latter two conditions hold for a random walk on any proper distribution, and irreducibility holds as long as the random walk has a positive probability of eventually reaching any state from any other state; that is, the jumping distributions J_t must be able to eventually jump to all states with positive probability.

To see that the target distribution is the stationary distribution of the Markov chain generated by the generalized Metropolis algorithm, consider starting the algorithm at time $t-1$ with a draw $x^{(t-1)}$ from the target distribution $P(x)$. Now consider any two points y and z with positive probability under P, labeled so that $P(z)J_t(y|z) \geq P(y)J_t(z|y)$. The unconditional probability of a transition from y to z is

$$\Pr(x^{(t-1)}=y, x^{(t)}=z) = P(y)J_t(z|y),$$

and the unconditional probability of a transition from z to y is

$$\begin{aligned} \Pr(x^{(t-1)}=z, x^{(t)}=y) &= P(z)J_t(y|z)\frac{P(y)}{P(z)}\frac{J_t(z|y)}{J_t(y|z)} \\ &= P(y)J_t(z|y), \end{aligned}$$

which is the same as the probability of a transition from y to z. Since their joint distribution is exchangeable, $x^{(t)}$ and $x^{(t-1)}$ have the same marginal distributions, and so P is the stationary distribution of the Markov chain.

The method of Metropolis et al. (1953) is the same as that described above, with the restrictions that the jumping distribution be symmetric and not depend on t: $J_t(y|z) = J_t(z|y) = J_0(z|y)$ for any y, z.[3]

[3]Barker (1965) suggests a method identical to Metropolis', except that the switching probability at each step is changed from $\min(r,1)$ to $\frac{r}{1+r} = \frac{P(\tilde{x})}{P(\tilde{x})+P(x^{(t-1)})}$. Alternatively, Barker's method may be considered a generalized Metropolis algorithm in which the jumping distribution $J(y|z)$ is replaced by

$$J[\text{Barker}](y|z) = J(y|z)\frac{P(y)}{P(y)+P(z)} \text{ for all } y \neq z.$$

See also Hastings (1970) for further discussion of Barker's, Metropolis', and related algorithms.

4.3. Gibbs Sampling

Geman and Geman (1984) introduced "Gibbs sampling," a procedure for simulating a multivariate probability distribution $P(x) = P(x_1, \ldots, x_d)$, by performing a random walk on the vector $x = (x_1, \ldots, x_d)$, altering one, possibly vector, component x_i at a time. At iteration t, an ordering of the d components of x is chosen and, in turn, each $x_i^{(t)}$ is sampled from the conditional distribution given that all the other components remained fixed:

$$P(x_i|x_{-i}^{(t-1)}),$$

where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$. Each of these d steps can be interpreted as one iteration of the generalized Metropolis algorithm, with the following jumping distribution, which only jumps along the i-th component, and does so with the conditional probability of x_i given $x_{-i} = x_{-i}^{(t-1)}$ obtained from the target distribution:

$$J_{it}[\text{Gibbs}](\tilde{x}|x^{(t-1)}) = \begin{cases} P(\tilde{x}_i \,|\, x_{-i}^{(t-1)}) & \text{if } \tilde{x}_{-i} = x_{-i}^{(t-1)} \\ 0 & \text{otherwise.} \end{cases}$$

Under this jumping distribution, the ratio of importance ratios is

$$\begin{aligned} r &= \frac{P(\tilde{x})}{P(x^{(t-1)})}\frac{J_{it}(x^{(t-1)}|\tilde{x})}{J_{it}(\tilde{x}|x^{(t-1)})} \\ &= \frac{P(\tilde{x})}{P(x^{(t-1)})}\frac{P(x_i^{(t-1)}|x_{-i}^{(t-1)})}{P(\tilde{x}_i|x_{-i}^{(t-1)})} \\ &\equiv 1, \end{aligned}$$

and so jumps always occur, as prescribed by the Gibbs sampler. Obviously, as described, the Gibbs sampler requires the ability to draw from the conditional distributions derived from the target distribution.

Usually, one iteration of the Gibbs sampler is defined as above, to include all d Metropolis steps corresponding to the d components of x, thereby updating all of x at each iteration. It is possible, however, to define Gibbs sampling without the restriction that each component be updated in each iteration.

4.4. Gibbs Sampling with Approximations

For some problems, sampling from some, or all, of the conditional distributions $P(x_i|x_{-i})$ is impossible, and one must resort to approximations $g(x_i|x_{-i})$. Trying

to perform the Gibbs sampler directly, using the conditional distributions g instead of P, will not work. The generalized Metropolis algorithm, however, is suited for the task. As in the Gibbs sampler, we must choose an ordering for altering the d elements of x; the jumping function at the i-th Metropolis step at iteration t is then

$$J_{it}(\tilde{x}|x^{(t-1)}) = \begin{cases} g(\tilde{x}_i|x_{-i}^{(t-1)}) & \text{if } \tilde{x}_{-i} = x_{-i}^{(t-1)} \\ 0 & \text{otherwise,} \end{cases}$$

and the ratio of importance ratios is

$$\begin{aligned} r &= \frac{P(\tilde{x})}{P(x^{(t-1)})} \frac{J_{it}(x^{(t-1)}|\tilde{x})}{J_{it}(\tilde{x}|x^{(t-1)})} \\ &= \frac{P(\tilde{x}_i|x_{-i}^{(t-1)})}{P(x_i^{(t-1)}|x_{-i}^{(t-1)})} \frac{g(x_i^{(t-1)}|x_{-i}^{(t-1)})}{g(\tilde{x}_i|x_{-i}^{(t-1)})}, \end{aligned}$$

which is identically equal to 1 only if $g(x|x_{-i}) \equiv P(x|x_{-i})$. If g is only an approximation, the Metropolis step will have a positive probability of not jumping.

5. Discussion

A large and expanding family of iterative and non-iterative simulation algorithms exist for approximating a target distribution using samples from a starting distribution. Despite the non-iterative appearance of rejection and importance sampling, all the available methods (except for direct simulation) yield exact simulations of the target distribution only in the limit that the number of samples $n \to \infty$. (Rejection sampling, however, has the advantage that once samples have been obtained, they are known to follow the target distribution.) In every approach, the starting distribution is key; an overdispersed start has long been recognized as necessary for rejection and importance sampling, and more recently been advocated for Markov chain simulation (Gelman and Rubin, 1992, 1993).

Monitoring convergence of all these methods (except for rejection sampling) can be difficult for any of these algorithms in practice. Gelman and Rubin (1993) present one approach based on performing multiple independent simulation runs which, while designed for iterative simulation methods such as the Gibbs sampler, might also be useful for inference from importance sampling.

6. References

Barker, A. A. (1965), "Monte Carlo calculations of radial distribution functions for a proton-electron plasma", *Aust. J. Phys.*, **18**, 119–133.

Besag, J., and Green, P. J. (1993), "Spatial statistics and Bayesian computation", *J. Roy. Stat. Soc. B*, **55**, to appear.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *J. Roy. Stat. Soc. B*, **39**, 1–38.

Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*. New York: Wiley.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-based approaches to calculating marginal densities", *J. Amer. Stat. Assoc.*, **85**, 398–409.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990), "Illustration of Bayesian inference in normal data models using Gibbs sampling", *J. Amer. Stat. Assoc.*, **85**, 398–409.

Gelman, A., and Rubin, D. B. (1992), "A single series from the Gibbs sampler provides a false sense of security", in *Bayesian Statistics 4*, ed. J. Bernardo, Oxford University Press.

Gelman, A., and Rubin, D. B. (1993), "Inference from iterative simulation using multiple sequences", *Stat. Sci.*, to appear.

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

Gilks, W. R., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., McNeil, A. J., Sharples, L. D., and Kirby, A. J. (1993), "Modelling complexity: applications of Gibbs sampling in medicine", *J. Roy. Stat. Soc. B*, **55**, to appear.

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

Kong, A., Liu, J., and Wong, W. H. (1991), "Sequential imputations and Bayesian missing data problems", Technical Report #321, Department of Statistics, University of Chicago.

Meng, X. L., and Rubin, D. B. (1991), "Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm", *J. Amer. Stat. Assoc.*, **86**, 899–909.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines", *J. Chem. Phys.*, **21**, 1087–1092.

Metropolis, N., and Ulam, S. (1949), "The Monte Carlo method", *J. Amer. Stat. Assoc.*, **44**, 335–341.

Morris, C. N. (1988), "Approximating posterior distributions and posterior moments", in *Bayesian Statistics 3*, 327–344, ed. J. Bernardo, Oxford University Press.

Rubin, D. B. (1987), "A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm", *J. Amer. Stat. Assoc.*, **82**, 543–546.

Rubin, D. B. (1988), "Using the SIR algorithm to simulate posterior distributions", in *Bayesian Statistics 3*, 395–402, ed. J. Bernardo, Oxford University Press.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods", *J. Roy. Stat. Soc. B*, **55**, to appear.

Tanner, M. A., and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation (with discussion)", *J. Amer. Stat. Assoc.*, **82**, 528–550.

Tierney, L. (1991), "Exploring posterior distributions using Markov chains", *Proceedings of the Interface Conference*.

Tierney, L., and Kadane, J. B. (1986), "Accurate approximations for posterior moments and marginal densities", *J. Amer. Stat. Assoc.*, **81**, 82–86.