A Single Series from the Gibbs Sampler Provides a False Sense of Security*

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

Donald B. Rubin
Department of Statistics
Harvard University
Cambridge, MA 02138

September 25, 1991

Abstract

The Gibbs sampler can be very useful for simulating multivariate distributions, but naive use of it can give misleading—falsely precise—answers. An example with the Ising lattice model demonstrates that it is generally impossibile to assess convergence of a Gibbs sampler from a single sample series. This conclusion also applies to other iterative simulation methods such as the Metropolis algorithm.

Keywords: Bayesian inference, iterative simulation, Ising model, Metropolis algorithm, random walk.

1 Introduction

Bayesian inference is becoming more common in applied statistical work, partly to make use of the flexible modeling that occures when treating all unknowns as random variables, but also because of the increasing availability of powerful computing environments. It is now often possible to obtain inferences using simulation—that is, to summarize a "target" posterior distribution by a sample of random draws from it, rather than by analytic calculations.

*To appear in *Bayesian Statistics 4*, J. Bernardo, ed., Oxford University Press (1992). This work was partially supported by the National Science Foundation and AT&T Bell Laboratories.

For many problems, direct simulation of a multivariate distribution is impossible, but one can simulate a random walk through the parameter space whose stationary distribution is the target distribution. Such a random walk may be considered an iterative Markov process that changes its distribution at every iteration, converging from the incorrect starting distribution to the target distribution as the number of iterations increases.

Such iterative simulation techniques have been used in physics since Metropolis et al. (1953) for studying Boltzmann distributions (also called Gibbs models) from statistical mechanics, with applications including solid state physics and the kinetic theory of gases. Geman and Geman (1984) applied the physical lattice models of statistical mechanics, and the associated computational techniques of iterative simulation, to image analysis, and coined the term "Gibbs sampler" for a particular technique. Since then, the Gibbs sampler has been applied with increasing frequency in a variety statistical estimation problems, often to probability distributions with no connection to Gibbs models; Gelfand et al. (1990) provide a review.

Although the Gibbs sampler is becoming popular, it can be easily misused relative to direct simulation, because in practice, a finite number of iterations must be used to estimate the target distribution, and thus the simulated random variables are, in general, never from the desired target distribution. Various suggestions, some of which are quite sophisticated, have appeared in the statistical literature and elsewhere for judging convergence using one iteratively simulated sequence (e.g., Ripley, 1987; Geweke, 1991; Raftery and Lewis, 1991). Here we present a simple but striking example of the problems that can arise when trying to assess convergence from one observed series. (We use the terms "series" and "sequence" interchangeably.)

This example suggests that a far more generally successful approach is based on simulating multiple independent sequences. The suggestion to

use multiple sequences with iterative simulation is not new (e.g., Fosdick, 1959), but no general method has appeared for drawing inferences about a target distribution from multiple series of finite length. Subsequent work (Gelman and Rubin, 1991) uses simple statistics to obtain valid conservative inferences for a large class of target distributions, including but not limited to those distributions that can be handled by one sequence.

2 Ising model

The Ising model, described in detail in Kinderman and Snell (1980) and Pickard (1987), is a family of probability distributions defined on a lattice $Y = (Y_1, \ldots, Y_k)$ of binary variables: $Y_i = \pm 1$. It is a particularly appropriate model to illustrate potential problems of judging convergence with one iteratively simulated sequence since much of the understanding of it comes from computer simulation (e.g., Fosdick, 1959; Ehrman et al., 1960). The Ising model was originally used by physicists to idealize the magnetic behavior of solid iron: each component Y_i is the magnetic field ("up" or "down") of a dipole at a site of a crystal lattice. From the laws of statistical mechanics, Y is assigned the Boltzmann distribution:

$$P(Y) \propto \exp(-\beta U(Y)), \tag{1}$$

where the "inverse temperature" β is a scalar quantity, assumed known, and the "potential energy" U is a known scalar function that reflects the attraction of dipoles of like sign in the lattice—more likely states have lower energy. The Ising energy function is simple in that only attractions between nearest neighbors contribute to the total energy:

$$U(Y) = -\sum_{i,j} \delta_{ij} Y_i Y_j \tag{2}$$

where $\delta_{ij} = 1$ if i and j are "nearest neighbors" in the lattice and $\delta_{ij} = 0$ otherwise; in a two-dimensional lattice, each site has four neighbors, except for

edge sites, which have three neighbors, and corner sites, with two neighbors each. The Ising model is commonly summarized by the "nearest-neighbor correlation" $\rho(Y)$:

$$\rho(Y) = \frac{\sum_{i,j} \delta_{ij} Y_i Y_j}{\sum_{i,j} \delta_{ij}}.$$

We take the distribution of $\rho(Y)$ as the target distribution, where Y is defined on a two-dimensional 100 × 100 lattice, and has distribution (1) with k = 10,000 and β fixed at 0.5.

3 Iterative simulation with the Gibbs sampler

For all but minscule lattices, it is difficult analytically to calculate summaries, such as the mean or variance of $\rho(Y)$; integrating out k lattice parameters involves adding 2^k terms. Direct simulation of the model is essentially impossible except in the one-dimensional case, for which the lattice parameters can be simulated in order. However, it is easy to iteratvely simulate the Ising distribution of Y, and thus of $\rho(Y)$, using the Gibbs sampler.

Given a multivariate target distribution $P(Y) = P(Y_1, \ldots, Y_k)$, the Gibbs sampler simulates a sequence of random vectors $(Y^{(1)}, Y^{(2)}, \ldots)$ whose distributions converge to the target distribution. The sequence $(Y^{(t)})$ may be considered a random walk whose stationary distribution is $P(Y)$. The Gibbs sampler proceeds as follows:

1. Choose a starting point $Y^{(0)} = (Y_1^{(0)}, \ldots, Y_k^{(0)})$ for which $P(Y^{(0)}) > 0$.

2. For $t = 1, 2, \ldots$:
   For $i = 1, \ldots, k$:
   Sample $Y_i^{(t)}$ from the conditional distribution:

   $$P(Y_i \mid Y_j = Y_j^{(t-1)}, \text{ for all } j \neq i),$$

   thereby altering one component of Y at a time; each *iteration* of the Gibbs sampler alters all k components of Y.

The proof that the Gibbs sampler converges to the target distribution has two steps: first, it is shown that the simulated sequence $(Y^{(t)})$ 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). 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, a condition satisfied by the Ising model.

To see that the target distribution is the stationary distribution of the Markov chain generated by the Gibbs sampler, consider starting the Gibbs sampler with a draw from $P(Y)$. Updating any component Y_i moves us to a new distribution with density,

$$P(Y_i|Y_j, \text{all } j \neq i)\, P(Y_j, \text{all } j \neq i),$$

which is the same as $P(Y)$.

4 Results of iterative simulation

For the Ising model, each step of the Gibbs sampler can alter all k lattice sites in order. An obvious way to start the iterative simulation is by setting each site to ± 1 at random. For the Ising model on a 100×100 lattice with $\beta = 0.5$, theoretical calculations (Pickard, 1987) show that the marginal distribution of $\rho(Y)$ is approximately Gaussian with mean nearly 0.9 and standard deviation about 0.01.

Figure 1 shows the values of $\rho(Y^{(t)})$, for $t = 1$ to 2000, obtained by the Gibbs sampler with a random start, so that $\rho(Y^{(0)}) \approx 0$; the first few values are not displayed in order to improve resolution on the graph. The series seems to have "converged to stationarity" after the thousand or so steps required to free itself from the initial state. Now look at Figure 2,

Figure 1: 2000 steps, starting at $\rho = 0$

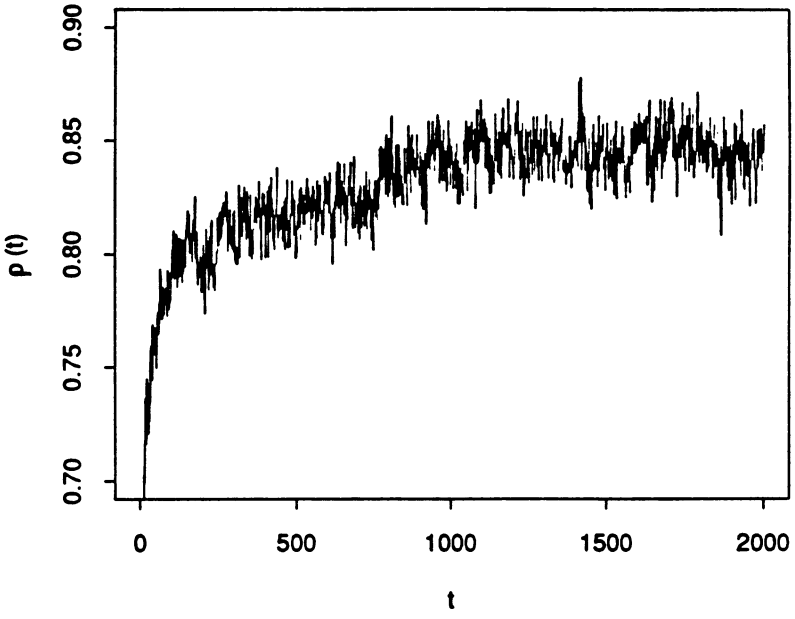


Figure 2: First 500 steps

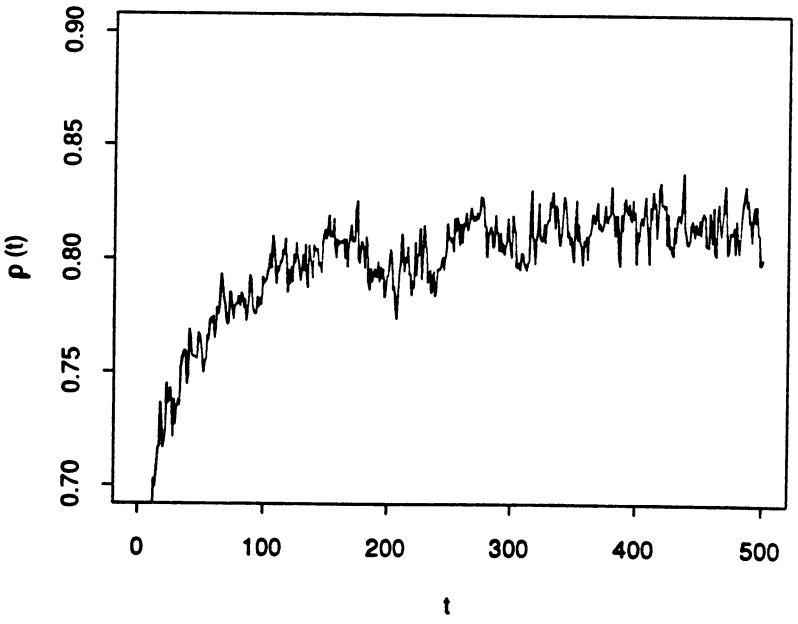
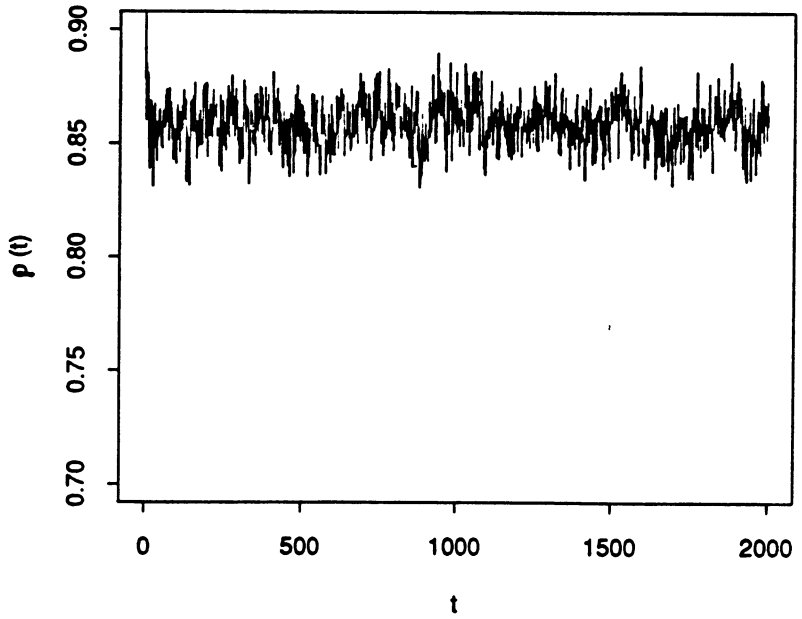


Figure 3: 2000 steps, starting at $\rho = 1$



which zooms in on the first 500 steps of the series. Figure 2 looks to have converged after about 300 steps, but a glance at the next 1500 iterations, as displayed in Figure 1, shows that the apparent convergence is illusory. For comparison, the Gibbs sampler was run again for 2000 steps, but this time starting at a point $Y^{(0)}$ for which $\rho(Y^{(0)}) = 1$; Figure 3 displays the series $\rho(Y^{(t)})$, which again seems to have converged nicely. To destroy all illusions about convergence in any of these figures, compare Figures 1 and 3: the two iteratively simulated sequences appear to have "converged," but to different distributions! The series in Figure 3 has "stabilized" to a higher value of $\rho(Y)$ than that of Figure 1. We are of course still observing transient behavior, and an estimate of the distribution of $\rho(Y)$ based on Figure 1 alone, or on Figure 3 alone, would be misleadingly—falsely—precise. Furthermore, neither series alone carries with it the information that it has not stabilized after 2000 steps.

5 Discussion

This example shows that the Gibbs sampler can stay in a small subset of its space for a long time, without any evidence of this problematic behavior being provided by one simulated series of finite length. The simplest way to run into trouble is with a two-chambered space, in which the probability of switching chambers is very low, but Figures 1–3 are especially disturbing because $\rho(Y)$ in the Ising model has a unimodal and approximately Gaussian marginal distribution, at least on the gross scale of interest. That is, the example is not pathological; the Gibbs sampler is just very slow. Rather than being a worst-case example, the Ising model is typical of the probability distributions for which iterative simulation methods were designed, and may be typical of many posterior distributions to which the Gibbs sampler is being applied.

A method designed for routine use must at the very least "work" for examples like the Ising model. By "work" we mean roughly that routine application should give valid conservative inferential summaries of the target distribution, that is, conservative relative to inferences that would be obtained from direct simulation of independent samples from the target distribution. Iterative simulation involves additional uncertainty due to the finite length of the simulated sequences, and so an appropriate inferential summary should reflect this in the same way that multiple imputation methods, which also involve the simulation of posterior distributions, reflect the uncertainty due to a finite number of imputations (Rubin, 1987).

In many cases, choosing a variety of dispersed starting points and running independent series may provide adequated diagnostic information, as in the example of Section 4. Nonetheless, for general practice, a more principled analysis of the between-series and within-series components of variability is far more convenient and useful. Gelman and Rubin (1991) offer such a solution based on simple statistics, which applies not only to the

Gibbs sampler but to any other method of iterative simulation, such as the Metropolis algorithm.

6 Using components of variance from multiple sequences

This section briefly presents, without derivations, our approach to inference from multiple iteratively simulated sequences; Gelman and Rubin (1991) present details in the context of a real example.

First, independently simulate $m \geq 2$ sequences, each of length $2n$, with starting points drawn from a distribution that is overdispersed relative to the target distribution (in the sense that overdispersed distributions are used in importance sampling). To limit the effect of the starting distribution, ignore the first n iterations of each sequence and focus attention on the last n.

Second, for each scalar quantity X of interest (e.g., $\rho(Y)$ in the above example), calculate the sample mean $\overline{x}_{i.} = \frac{1}{n}\sum_j x_{ij}$ and variance $s_i^2 = \frac{1}{n-1}\sum_j (x_{ij} - \overline{x}_{i.})^2$, for each sequence $i = 1, \ldots, m$. Then calculate the variance components,

$$\begin{aligned} W &= \text{the average of the } m \text{ within-sequence variances, } s_i^2, \\ &\quad \text{each based on } n-1 \text{ degrees of freedom, and} \\ B/n &= \text{the variance between the } m \text{ sequence means, } \overline{x}_{i.}, \\ &\quad \text{each based on } n \text{ values of } X. \end{aligned}$$

If the average within-sequence variance, W, is not substantially larger than the between-sequence variance, B/n, then the m sequences have not yet come close to converging to a common distribution. With only one sequence, between and within variabilities cannot be compared.

Third, estimate the target mean $\mu = \int X P(X) dX$ by $\hat{\mu} = \frac{1}{m}\sum_i \overline{x}_{i.}$, the sample mean of all mn simulated values of X.

Fourth, estimate the target variance, $\sigma^2 = \int (X-\mu)^2 P(X) dX$, by a weighted average of W and B, namely $\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{1}{n} B$, which overestimates σ^2 assuming $P_0(X)$ is overdispersed. The estimate $\hat{\sigma}^2$ is unbiased under stationarity (i.e., if $P_0(X) = P(X)$), or in the limit $n \to \infty$.

Fifth, create a conservative Student-t distribution for X with center $\hat{\mu}$, scale $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + B/mn}$, and degrees of freedom $\nu = 2\hat{V}^2 / \widehat{\text{var}}(\hat{V})$, where

$$\begin{aligned} \widehat{\text{var}}(\hat{V}) &= \left(\frac{n-1}{n}\right)^2 \frac{1}{m} \text{var}(s_i^2) + \left(\frac{m+1}{mn}\right)^2 \frac{2}{m-1} B^2 + \\ &\quad + 2\frac{(m-1)(n-1)}{mn^2} \cdot \frac{n}{m} \left[\text{cov}(s_i^2, \overline{x}_{i\cdot}^2) - 2\overline{x}_{\cdot\cdot} \text{cov}(s_i^2, \overline{x}_{i\cdot})\right], \end{aligned}$$

and the variances and covariances are estimated from the m sample values of s_i^2, $\overline{x}_{i\cdot}$,, and $\overline{x}_{i\cdot}^2$.

Sixth, monitor convergence of the iterative simulation by estimating the factor by which the scale of the Student-t distribution for X might be reduced if the simulations were continued in the limit $n \to \infty$. This potential scale reduction is estimated by $\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}}{W} \frac{\nu}{\nu-2}}$, which declines to 1 as $n \to \infty$.

A computer program implementing the above steps appears in the appendix to Gelman and Rubin (1991) and is available from the authors.

7 Previous multiple-sequence methods

Our approach is related to previous uses of multiple sequences to monitor convergence of iterative simulation procedures. Fosdick (1959) simulated multiple sequences, stopping when the difference between sequence means was less than a prechosen error bound, thus basically using W but without using B as a comparison. Similarly, Ripley (1987) suggested examining at least three sequences as a check on relatively complicated single-sequence methods involving graphics and time-series analysis, thereby essentially estimating W quantitatively and B qualitatively. Tanner and Wong (1987)

and Gelfand and Smith (1990) simulated multiple sequences, monitoring convergence by qualitatively comparing the set of m simulated values at time s to the corresponding set at a later time t; this general approach can be thought of as a qualitative comparison of values of B at two time points in the sequences, without using W as a comparison.

Our method differs from previous multiple-sequence methods by being fully quantitative, differs from single-sequence methods by relying on only a few assumptions, and differs from previous approaches of either kind by incorporating the uncertainty due to finite-length sequences into the distributional estimates. Current work focuses on the possibility of obtaining automatically overdispersed starting distributions and more efficient estimated target distributions for specific models.

References

Ehrman, J. R., Fosdick, L. D., and Handscomb, D. C. (1960). Computation of order parameters in an Ising lattice by the Monte Carlo method. *Journal of Mathematical Physics*, **1**, 547–558.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, section XV.7. New York: Wiley.

Fosdick, L. D. (1959). Calculation of order parameters in a binary alloy by the Monte Carlo method. *Physical Review*, **116**, 565–573.

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. *Journal of the American Statistical Association*, **85**, 398–409.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. and Rubin, D. B. (1991). Honest inferences from iterative simulation. Technical report #307, Department of Statistics, University of

California, Berkeley.

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.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In these proceedings.

Kinderman, R. and Snell, J. L. (1980). *Markov Random Fields and Their Applications.* Providence, R.I.: American Mathematical Society.

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

Pickard, D. K. (1987). Inference for discrete Markov fields: the simplest nontrivial case. *Journal of the American Statistical Association*, **82**, 90–96.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the Gibbs sampler? In these proceedings.

Ripley, B. D. (1987). *Stochastic Simulation*, chapter 6. New York: Wiley.

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

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.