Theoretical rates of convergence for Markov chain Monte Carlo

by

Jeffrey S. Rosenthal*

Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1

Phone: (416) 978-4594. Internet: jeff@utstat.toronto.edu

(Conference proceedings, Interface '94, June 1994.)

Abstract. We present a general method for proving rigorous, *a priori* bounds on the number of iterations required to achieve convergence of Markov chain Monte Carlo. We describe bounds for specific models of the Gibbs sampler, which have been obtained from the general method. We discuss possibilities for obtaining bounds more generally.

1. Introduction.

Markov chain Monte Carlo techniques, including the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), data augmentation (Tanner and Wong, 1986), and the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) have become very popular in recent years as a way of generating a sample from complicated probability distributions (such as posterior distributions in Bayesian inference problems). A fundamental issue regarding such techniques is their convergence properties, specifically whether or not the algorithm will converge to the correct distribution, and if so how quickly. Many general convergence results (e.g. Tierney, 1994), qualitative convergence-rate results (Schervish and Carlin, 1992; Liu, Wong, and Kong, 1991a, 1991b; Baxter and Rosenthal, 1994), and convergence diagnostics (e.g. Roberts, 1992; Gelman and Rubin, 1992; Mykland, Tierney, and Yu, 1992) have been developed. However, none of these approaches are entirely satisfactory (Matthews, 1991; Cowles and Carlin, 1994).

In a different direction, a number of papers have attempted to prove rigorous, quantitative bounds on convergence rates for these algorithms (Jerrum and Sinclair, 1989; Frieze, Kannan, and Polson, 1993; Meyn and Tweedie, 1993; Lund and Tweedie, 1993; Mengersen and Tweedie, 1993; Rosenthal, 1991, 1993a, 1993b, 1994). Such results often provide bounds which are very weak, and/or for very specific models, but the area appears to be worthy of further work.

In this paper we describe a general method (Section 2) for proving such quantitative bounds. The method requires only that we verify a drift condition and a minorization condition, for the Markov chain of interest. We describe (Section 3) the application of this (and related) methods to various specific examples of the Gibbs sampler, including variance components models, hierarchical Poisson models, and a model related to James-Stein estimators. In some cases, the bounds appear to be small enough to be of practical use. In other cases, they provide additional theoretical information about the Gibbs sampler for the model being studied.

We close (Section 4) with a brief discussion of possibilities for further bounds of this type.

2. The general method.

The simplest form of our general method is the following, taken from Rosenthal (1993b, Theorem 12).

Proposition. *Let $P(x, \cdot)$ be the transition probabilities for a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose there exist $\epsilon > 0$, $0 < \lambda < 1$, $0 < \Lambda < \infty$, $d > \frac{2\Lambda}{1-\lambda}$, a non-negative function $f : \mathcal{X} \to \mathbf{R}$, and a probability measure $Q(\cdot)$ on $\mathcal{X}$, such that*

$$\mathbf{E}\left(f(X_1) \,|\, X_0 = x\right) \;\leq\; \lambda f(x) \,+\, \Lambda\,, \qquad x \in \mathcal{X} \quad (1)$$

and

$$P(x, \cdot) \;\geq\; \epsilon\, Q(\cdot)\,, \qquad x \in f_d \quad (2)$$

where $f_d = \{x \in \mathcal{X} \,|\, f(x) \leq d\}$, and where $P(x, \cdot) \geq \epsilon\, Q(\cdot)$ means $P(x, S) \;\geq\; \epsilon\, Q(S)$ for every measurable $S \subseteq \mathcal{X}$. Then for any $0 < r < 1$, the total variation distance to the stationary distribution after k iterations is bounded above by

$$(1-\epsilon)^{rk} + \left(\alpha^{-(1-r)}\gamma^{r}\right)^{k} \left(1 + \frac{\Lambda}{1-\lambda} + \mathbf{E}\left(f(X_0)\right)\right),$$

* Supported in part by NSERC of Canada.

where

$$\alpha^{-1} \;=\; \frac{1+2\Lambda+\lambda d}{1+d} \;<\; 1\,; \quad \gamma \;=\; 1+2(\lambda d+\Lambda)\,.$$

Inequality (1) above is called a *drift condition*, while inequality (2) above is called a *minorization condition*. The proposition thus allows for precise, quantitative, exponentially-decreasing upper bounds on the distance to stationarity, as a function of the number of iterations k, using just these two inequalities.

The proof of this proposition involves the *coupling inequality*, which states that the total variation distance between the laws of two random variables is bounded by the probability that they are unequal. Proving the proposition thus amounts to (theoretically) constructing auxiliary random variables Y_k, so that $\mathcal{L}(Y_k) = \pi$ but $P(X_k = Y_k)$ is as large as possible. Inequality (2) allows us to construct X_k and Y_k jointly so that, whenever $(X_k, Y_k) \in f_d \times f_d$, they have probability ϵ of becoming equal on the next generation. Furthermore, inequality (1) implies that the number of iterations k for which $(X_k, Y_k) \in f_d \times f_d$ will be large with high probability. Combining these two facts, we can construct X_k and Y_k so that $P(X_k \neq Y_k)$ is small, and thus use the coupling inequality to prove the proposition. The reader is referred to Rosenthal (1993b) for details.

3. Applications to specific models.

The general method of Section 2 (and related methods) have been applied to a number of specific examples of the Gibbs sampler, to derive information about their rates of convergence to the appropriate posterior distributions.

In Rosenthal (1993), a version of the data augmentation algorithm (a special case of the Gibbs sampler) was applied to *finite* sample spaces. It was shown that, with n parameters and n observed data, the algorithm would converge in $O(\log n)$ iterations. Thus, the running time of the algorithm does not grow too quickly with the number of parameters.

In Rosenthal (1991), the Gibbs sampler applied to variance components models (as discussed in Gelfand and Smith, 1990; Gelfand et al., 1990) was analyzed. It was shown that, with K location parameters each having J observed data, the $(K+3)$-dimensional Gibbs sampler would approximately converge in $O\left(1 + \frac{\log K}{\log J}\right)$ iterations. So again, the running time of the algorithm does not grow too quickly with the number of parameters.

In Rosenthal (1993b), the Gibbs sampler applied to a hierarchical Poisson model was analyzed, using the same data as analyzed in Gelfand and Smith (1990). For this data, the (11-dimensional) Gibbs sampler was shown to have total variation distance to stationarity after k iterations bounded above by

$$(0.976)^k \;+\; (0.951)^k (6.2 + E\left((S^{(0)} - 6.5)^2\right))\,,$$

where $S^{(0)} = \sum_i \theta_i^{(0)}$ is a sum of initial values. The bound is thus explicit and quantitative, and depends explicitly on the initial distribution. The bound is also not absurdly large: for example, if $E\left((S^{(0)} - 6.5)^2\right) = 2$ and $k = 150$, this bound is equal to 0.03, implying that 150 iterations are sufficient to achieve randomness.

In Rosenthal (1994), the Gibbs sampler applied to a model related to James-Stein estimators (James and Stein, 1961) was analyzed. The model (suggested by Jun Liu) was designed to avoid the use of guesses and empirical estimates in the usual (empirical Bayes) formulation of James-Stein estimators. The Gibbs sampler was intended to facilitate computations related to the associated posterior distribution. A formula was provided which gave a bound on convergence of the Gibbs sampler explicitly, in terms of the number of iterations, the initial distributions, the prior distributions of the model, and the observed data. When applied to the baseball data analyzed in Efron and Morris (1975) and Morris (1983), it proved that the Gibbs sampler would converge in less than 200 iterations.

For certain other prior distributions, it was shown (Rosenthal, 1994) that this Gibbs sampler would in fact not converge at all. This information was used, together with standard convergence theory, to prove that for these (improper) priors, the model itself was improper, i.e. the posterior distribution was non-normalizable. Analysis of the Gibbs sampler was thus used to provide additional information about the model under consideration.

Our method has thus been applied to a variety of realistic examples of the Gibbs sampler. It has provided useful quantitative bounds, convergence information relating the running time to the number of parameters, and additional theoretical information about the underlying model itself.

4. Discussion.

It is now widely recognized that convergence issues are crucial for the successful implementation of Markov chain Monte Carlo algorithms. However, no method is entirely satisfactory for demonstrating such convergence or providing a convergence rate.

We have provided a general method (Section 2) for rigorously and explicitly bounding the convergence of these Markov chain algorithms. The method requires only that we verify a drift condition and a minorization condition for the Markov chain under consideration. In principle the method can be applied to virtually any Markov chain algorithm, and does not require special structure such as spectral information or reversibility. However, it is to be admitted that, in complicated high-dimensional problems, even the verification of the two required conditions can be quite difficult.

We have described the application of this method to several models of the Gibbs sampler. These models are realistic and non-trivial, and our method provides useful information about their convergence properties. The theoretical results appear to be at the point where they can begin to have practical implications.

However, each of these applications has required additional, extensive computation. Furthermore, similar computation may be extremely difficult for more complicated models. Hence, further work is required before these methods are easily usable in very general applied settings. It is possible that the theoretical approach described here can be combined with a more practical analysis, for example by attempting to verify drift and minorization conditions through additional simulation (Cowles and Rosenthal, 1994), which might allow for wider use.

In any case, while there is much work to be done, the methods described here appear to hold promise for providing rigorous rates of convergence for many additional examples of Markov chain Monte Carlo.

Acknowledgements. In the course of this research I have benefited from very helpful conversations with many people, including Persi Diaconis, Jun Liu, Peter Ney, Richard Tweedie, John Baxter, Kate Cowles, and Neal Madras.

REFERENCES

J.R. Baxter and J.S. Rosenthal (1994), Rates of convergence for everywhere-positive Markov chains. Tech. Rep. **9406**, Dept. of Statistics, University of Toronto.

M.K. Cowles and B.P. Carlin (1994), Evaluation and comparison of Markov chain Monte Carlo convergence diagnostics. Tech. Rep., Dept. of Biostatistics, University of Minnesota.

M.K. Cowles and J.S. Rosenthal (1994), work in progress.

B. Efron and C. Morris (1975), Data analysis using Stein's estimator and its generalizations. J. Amer. Stat. Assoc., Vol. **70**, No. **350**, 311-319.

A. Frieze, R. Kannan, and N.G. Polson (1993), Sampling from log-concave distributions. Tech. Rep., School of Computer Science, Carnegie-Mellon University.

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

A.E. Gelfand, S.E. Hills, A. Racine-Poon, and A.F.M. Smith (1990), Illustration of Bayesian inference in normal data models using Gibbs sampling. J. Amer. Stat. Soc. **85**, 972-985.

A. Gelman and D.B. Rubin (1992), Inference from iterative simulation using multiple sequences. Stat. Sci., Vol. **7**, No. **4**, 457-472.

S. Geman and D. Geman (1984), Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on pattern analysis and machine intelligence **6**, 721-741.

W. James and C. Stein (1961), Estimation with Quadratic Loss. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Vol. **1**, University of California Press, Berkeley, 361-379.

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

M. Jerrum and A. Sinclair (1989), Approximating the permanent. SIAM J. Comput. **18**, 1149-1178.

J. Liu, W. Wong, and A. Kong (1991a), Correlation structure and the convergence of the Gibbs sampler, *I*. Tech Rep. **299**, Dept. of Statistics, University of Chicago. Biometrika, to appear.

J. Liu, W. Wong, and A. Kong (1991b), Correlation structure and the convergence of the Gibbs sampler, *II*: Applications to various scans. Tech Rep. **304**, Dept. of Statistics, University of Chicago. J. Royal Stat. Sci. **(B)**, to appear.

R.B. Lund and R.L. Tweedie (1993), Geometric convergence rates for stochastically ordered Markov chains. Tech. Rep., Dept. of Statistics, Colorado State University.

P. Matthews (1993), A slowly mixing Markov chain with implications for Gibbs sampling. Stat. Prob. Lett. **17**, 231-236.

K.L. Mengersen and R.L. Tweedie (1993), Rates of convergence of the Hastings and Metropolis algorithms. Tech. Rep. **93/30**, Dept. of Statistics, Colorado State University.

N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1091.

S.P. Meyn and R.L. Tweedie (1993), Computable bounds for convergence rates of Markov chains. Tech. Rep., Dept. of Statistics, Colorado State University.

C. Morris (1983), Parametric empirical Bayes confidence intervals. Scientific Inference, Data Analysis, and Robustness, 25-50.

P. Mykland, L. Tierney, and B. Yu (1992), Regeneration in Markov chain samplers. Tech. Rep. **585**, School of Statistics, University of Minnesota.

G.O. Roberts (1992), Convergence diagnostics of the Gibbs sampler. In Bayesian Statistics **4** (J.M. Bernardo et al., eds.), 777-784. Oxford University Press.

J.S. Rosenthal (1991), Rates of convergence for Gibbs sampler for variance components models. Tech. Rep. **9322**, Dept. of Statistics, University of Toronto. (Tentatively accepted in *Annals of Statistics*.)

J.S. Rosenthal (1993a), Rates of convergence for Data Augmentation on finite sample spaces. Ann. Appl. Prob., Vol. **3**, No. **3**, 319-339.

J.S. Rosenthal (1993b), Minorization conditions and convergence rates for Markov chain Monte Carlo. Tech. Rep. **9321**, Dept. of Statistics, University of Toronto.

J.S. Rosenthal (1994), Analysis of the Gibbs sampler for a model related to James-Stein estimators. Tech. Rep. **9413**, Dept. of Statistics, University of Toronto.

M.J. Schervish and B.P. Carlin (1992), On the convergence of successive substitution sampling, J. Comp. Graph. Stat. **1**, 111–127.

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

L. Tierney (1994), Markov chains for exploring posterior distributions. Tech. Rep. **560**, School of Statistics, University of Minnesota. Ann. Stat., to appear.