Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo

Jeffrey S. Rosenthal

To cite this article: Jeffrey S. Rosenthal (1995) Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo, Journal of the American Statistical Association, 90:430, 558-566, DOI: 10.1080/01621459.1995.10476548

To link to this article: https://doi.org/10.1080/01621459.1995.10476548

Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo

Jeffrey S. ROSENTHAL*

General methods are provided for analyzing the convergence of discrete-time, general state-space Markov chains, such as those used in stochastic simulation algorithms including the Gibbs sampler. The methods provide rigorous, a priori bounds on how long these simulations should be run to give satisfactory results. Results are applied to two models of the Gibbs sampler: a bivariate normal model, and a hierarchical Poisson model (with gamma conditionals). The methods use the notion of *minorization conditions* for Markov chains.

KEY WORDS: Bivariate normal model; Coupling; Drift condition; Gibbs sampler; Harris recurrence; Hierarchical Poisson model; Metropolis–Hastings algorithm; Regeneration time.

1. INTRODUCTION

Markov chain Monte Carlo (MCMC) techniques have become very popular in the statistics literature, as a way of sampling from complicated probability distributions (such as those arising in Bayesian inference). These techniques have their roots in the Metropolis–Hastings algorithm (Hastings 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller 1953), and include the Gibbs sampler (Gelfand and Smith 1990; Geman and Geman 1984) and data augmentation (Tanner and Wong 1987).

One fundamental question about MCMC is its convergence rate. Specifically, how long does MCMC need to be run before it gives satisfactory answers? In applied problems this question is often answered heuristically, by "eyeballing" the MCMC output. Often this appears to suffice in practice; however, it can sometimes be quite misleading (Matthews 1991), and it is desirable to have more systematic methods of establishing convergence of MCMC.

There have been various approaches to this problem. Geman and Geman (1984) and Schervish and Carlin (1992) described general results about exponential convergence (but without giving quantitative bounds), using the theory of compact operators. Similar approaches have been used by Liu, Wong, and Kong (1991a,b) and Baxter and Rosenthal (1994). A "discretized" Markov chain was analyzed by Applegate, Kannan, and Polson (1990) and by Frieze, Kannan, and Polson (1993), who proved polynomial bounds on certain running times. A method for estimating the variance of the chain was suggested by Geyer (1992). There have also been various papers giving quantitative bounds on convergence rates for specific models, including those by Amit and Grenander (1991), Amit (1991, 1993), Rosenthal (1993, 1991), Liu (1992), Frigessi, Hwang, and Younes (1992), Diaconis and Hanlon (1992), and Belsley (1993, chap. 6). In addition, various "convergence diagnostics" have been suggested by a number of authors, including Roberts (1992) and Gelman and Rubin (1992).

This article provides a method (Theorem 5) for proving rigorous, a priori bounds on the number of iterations required until satisfactory convergence has taken place. We feel that such bounds provide increased confidence in the results of MCMC, and allow for improved analysis of the efficiency of various algorithms. It is our hope that the methods presented here can be applied quite generally to many different Markov chain samplers.

Our method involves establishing minorization conditions (i.e., splits) for Markov chains (see Sec. 2) to establish results about convergence of MCMC. This amounts to showing that the Markov chain satisfies a condition of the form $P^{k_0}(x, \cdot) \geq \varepsilon Q(\cdot)$ for all points x in a subset R of the state space. We have attempted to make the method simpler, and easier to apply, than previous related work (Rosenthal 1993, 1991). It is our hope that it can be applied with greater ease and to a wider variety of models.

In Section 2 we present the essence of the method (Theorem 1). In Section 3 we present a method for bounding (exponential moments of) the return time to R, in terms of a drift condition involving an auxiliary function h with decreasing expectation. We use this method to establish exponential convergence (in total variation distance), with explicit rate, in quite general situations (Theorem 5). We further provide some lemmas to facilitate the application of this theorem. We hope that the method presented here can be applied to a wide variety of MCMC's and can provide useful bounds on their time to convergence.

In Section 4 we use a simplified version of the method of Section 3 to analyze *regeneration points* of a Markov chain, without necessarily establishing convergence in total variation distance. In particular, we provide explicit exponential bounds on the time required to complete a fixed number of regeneration tours (Corollary 9). The results thus relate to work of Mykland, Tierney, and Yu (1992), who discussed how to identify regeneration times when running MCMC.

In Section 5 we apply our ideas to two examples of the Gibbs sampler. The first is a simple bivariate normal example, taken from Schervish and Carlin (1992). The second is a hierarchical Poisson model (with gamma conditionals) using actual data, taken from Gelfand and Smith (1990),

* Jeffrey S. Rosenthal is Assistant Professor, Department of Statistics, University of Toronto, Ontario, Canada M5S 1A1. This research was partially supported through National Science Foundation Grant DMS-90-02899, and through a research grant from Natural Science and Engineering Research Council of Canada. The author is very grateful to John Baxter for his suggestions regarding Lemma 4 herein, and thanks Persi Diaconis, Jun Liu, Peter Ney, Richard Tweedie, and Gareth Roberts for very helpful conversations and the referees for many excellent comments and suggestions.

which was also discussed by Tierney (1991) and Mykland, et al. (1992). For each of these two models, we provide explicit, numerical, exponentially decreasing bounds on total variation distance to stationarity. Although the bounds are not sharp numerically, they are not too wildly off, and could be of use in guiding a simulation.

Finally, in Section 6 we present (Theorem 12) a simplified version of the main result, which involves verifying a simpler drift condition than that in Theorem 5.

Remark. Since originally completing this manuscript, We have learned of recent similar work by Meyn and Tweedie (1993b). Using minorization conditions and a simple drift condition on the chain, they obtained computable bounds on the distance to stationarity under certain conditions. Their methods require slightly less information than ours; however, their bounds appear to be weaker in specific examples, I am very grateful to Richard Tweedie for discussing these issues with me in detail.

2. MINORIZATION CONDITIONS FOR MARKOV CHAINS

A Markov chain with transition kernel $P(x, dy)$ on a state space $\mathcal{X}$ is said to satisfy a *minorization condition* or *split* on a subset $R \subseteq \mathcal{X}$ if there is a probability measure $Q(\cdot)$ on $\mathcal{X}$, a positive integer k_0, and $\varepsilon > 0$, such that

$$P^{k_0}(x, A) \geq \varepsilon Q(A), \quad \forall\, x \in R, \qquad (*)$$

for all measurable subsets $A \subseteq \mathcal{X}$.

Minorization conditions are closely related to the notion of Harris recurrence. They were introduced by Athreya and Ney (1978); see also the work of Athreya, McDonald, and Ney (1978), Nummelin (1984), Asmussen (1989), Lindvall (1992), and Meyn and Tweedie (1993a). They have been used to analyze MCMC by Roberts and Polson (1990), Tierney (1991), Rosenthal (1993, 1991), and Mykland et al. (1992).

Most of this article is based on the following theorem. Special cases of the theorem were used in earlier work (Rosenthal 1991, 1993) for similar purposes.

Theorem 1. Suppose that a Markov chain $P(x, dy)$ on a state space $\mathcal{X}$ satisfies $(*)$ for some R, k_0, ε, and $Q(\cdot)$. Let $X^{(k)}$, $Y^{(k)}$ be two realizations of the Markov chain (started in any initial distribution), defined jointly as described in the proof. Let

$$t_1 = \inf\{m : (X^{(m)}, Y^{(m)}) \in R \times R\},$$

and for $i > 1$ let

$$t_i = \inf\{m : m \geq t_{i-1} + k_0, (X^{(m)}, Y^{(m)}) \in R \times R\}.$$

Set $N_k = \max\{i : t_i < k\}$. Then for any $j > 0$,

$$\|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\text{var}} \leq (1 - \varepsilon)^{[j/k_0]} + P(N_k < j),$$

where $[r]$ is the greatest integer not exceeding r.

This theorem would usually be applied with $\mathcal{L}(Y^{(0)}) = \pi$ (so that $\mathcal{L}(Y^{(k)}) = \pi$ for all times k). It thus gives a rigorous bound on the total variation distance between the distribution $\mathcal{L}(X^{(k)})$ of a Markov chain after k iterations and the target stationary distribution π.

If we take the subset R to be relatively small, then a good minorization can usually be found so that ε is reasonably large. The term $(1 - \varepsilon)^{[j/k_0]}$ in the bound will then decrease quickly as the number of iterations k gets large (assuming that j is chosen correspondingly large). The term $P(N_k < j)$ is more complicated and involves controlling the returns of the Markov chain to the subset R. This issue is explored in Section 3.

At the other extreme is when the condition $(*)$ is satisfied with $R = \mathcal{X}$; that is, on the entire state space. (This is called the Doeblin condition and is equivalent [Nummelin 1984, thm. 6.15; Tierney 1991, prop. 2] to the Markov chain being uniformly ergodic.) Clearly, if $R = \mathcal{X}$, then $N_k = [k/k_0]$ with probability 1, so we can take $j = [k/k_0]$ in Theorem 1 to conclude (as is well-known) the following.

Proposition 2. If a transition kernel P on a state space $\mathcal{X}$ satisfies $P^{k_0}(x, \cdot) \geq \varepsilon Q(\cdot)$ for all $x \in \mathcal{X}$, with $Q(\cdot)$ a probability distribution and $\varepsilon > 0$, then its variation distance to a stationary distribution π satisfies

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1 - \varepsilon)^{[k/k_0]}$$

for any starting distribution $\mathcal{L}(X^{(0)})$.

This proposition was discussed by Nummelin (1984), Roberts and Polson (1990), and others. It was used in Rosenthal (1993) to obtain convergence rates for the Gibbs sampler for a hierarchical Bernoulli model. Now one might suppose that this "uniform ergodicity" approach would work only for models with bounded state spaces. But Example 2 (in Sec. 5) has unbounded random variables, and yet it is easily seen that Proposition 2 still applies (though it gives a very small value of ε; hence we use a different approach).

3. BOUNDING THE TAIL OF N_k

In applying Theorem 1, it will usually not be possible to establish condition $(*)$ on the entire state space. Indeed, to keep ε reasonably large and k_0 reasonably small, it is often necessary to keep the subset R relatively small. To apply Theorem 1, it is then necessary to get bounds on the tail probabilities $P(N_k < j)$ of the random variable N_k (i.e., the number of times $m \leq k$ for which $(X^{(m)}, Y^{(m)}) \in R \times R$).

In earlier work (Rosenthal 1991), a special case of Theorem 1 was used in which $j = [k/k_0]$, using the obvious bound

$$P(N_k < [k/k_0]) \leq P(X^{(0)} \notin R) + P(Y^{(0)} \notin R) + 2[k/k_0] \sup_{x \in R} P^{k_0}(x, R^C).$$

But to make this bound go to zero as a function of k, it was necessary to let the subset R (and the value of k_0) grow larger and larger as a function of k. This made the calculations considerably more complicated.

One of the main goals of this article is to simplify such analyses. We propose to bound $P(N_k < j)$ more carefully, in a way that is more easily applicable and leads to useful, exponential bounds on variation distance. In particular, this allows for use of smaller values of k_0; in both of our examples below we use $k_0 = 1$.

It is well known that the expected number of returns to the set R by time k, $E(N_k)$, is bounded below by k/μ, where

μ is the mean return time to R (see Feller 1971, chap. XI, sec. 3). But it is not true, for example, that the coupling bound in Theorem 1 can be taken as $(1-\varepsilon)^{E(N_k)}$. (To see this, consider a case where N_k is equal to either zero or one million, each with probability $\frac{1}{2}$.) Thus the mean of N_k is insufficient to establish exponential convergence of the chain; more information is needed.

The approach that we take begins with the following.

Lemma 3. Let t_i be the "k_0-delayed hitting times of $R \times R$" as in Theorem 1, and let $r_i = t_i - t_{i-1}$ (with $r_1 = t_1$) represent the ith gap between such times (i.e., the "k_0-delayed ith return time to $R \times R$"). Then for any $\alpha > 1$,

$$P(N_k < j) \le \alpha^{-k} E\left(\prod_{i=1}^{j} \alpha^{r_i}\right).$$

Lemma 3 suggests that we attempt to bound the exponential moments $E(\alpha^{r_i})$ of the return times of $(X^{(k)}, Y^{(k)})$ to $R \times R$. An approach is suggested by the following lemma. It introduces an auxiliary function h whose expectation is decreasing rapidly when $(X^{(k)}, Y^{(k)}) \notin R \times R$, thus facilitating bounds on the return time to $R \times R$. It is somewhat related to the "drift condition" of Nummelin (1984, prop. 5.21).

Lemma 4. Let $X^{(k)}$ and $Y^{(k)}$ be two Markov chains on a state space $\mathcal{X}$, defined jointly as in Theorem 1, with $R \subseteq \mathcal{X}$, and with r_i the "k_0-delayed ith return time to $R \times R$" as earlier. Suppose that there is $\alpha > 1$ and a function $h : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ such that $h \ge 1$ and

$$E(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y) \le \alpha^{-1} h(x, y), \qquad \forall\, (x, y) \notin R \times R.$$

Then

$$E(\alpha^{r_1}) \le E(h(X^{(0)}, Y^{(0)})), \tag{a}$$

and for $i > 1$ and any choice of $r_1, \ldots, r_{i-1}$,

$$E(\alpha^{r_i} \mid r_1, \ldots, r_{i-1}) \le \alpha^{k_0} \sup_{(x,y) \in R \times R} E(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y). \tag{b}$$

Putting all of this together, we obtain the following.

Theorem 5. Suppose that a Markov chain $P(x, dy)$ satisfies condition $(*)$ for some R, k_0, and $\varepsilon > 0$ and satisfies the hypotheses of Lemma 4 for some h and α. Set

$$A = \sup_{(x,y) \in R \times R} E(h(X^{(k_0)}, Y^{(k_0)}) \mid X^{(0)} = x, Y^{(0)} = y).$$

Then if $\nu = \mathcal{L}(X^{(0)})$ is the initial distribution and π is a stationary distribution, then for any $j > 0$, the total variation distance to π after k steps satisfies

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \le (1-\varepsilon)^{\lfloor j/k_0 \rfloor} + \alpha^{-k+(j-1)k_0} A^{j-1} E_{\nu \times \pi}(h(X^{(0)}, Y^{(0)})).$$

(Here $E_{\nu \times \pi}$ means expectation with $X^{(0)}$ distributed according to ν and with $Y^{(0)}$ distributed independently according to π.)

This theorem provides a method for proving useful rates of convergence for a variety of Markov chains. Typically, one would choose j to be a small multiple of k (in Example 1 we use $j = k/10$). Also, a good choice for the function h appears to be of the form $h(x, y) = 1 + (x_i - a)^2 + (y_i - a)^2$, where x is a vector and x_i is its ith coordinate, which works well assuming that x_i tends to drift exponentially quickly towards the value a (at least while it's far away).

We illustrate this approach with two examples in Section 5.

Remarks.

1. Theorem 5 still necessitates bounding the expected value $E_{\nu \times \pi}(h(X^{(0)}, Y^{(0)}))$, which unfortunately depends on the unknown distribution π. However, if we have verified a drift condition of the form $E(V(X^{(1)}) \mid X^{(0)} = x) \le \lambda V(x) + b$, then it is easily seen (cf. Meyn and Tweedie 1993b, prop. 4.3(i)), by taking expectations of both sides with respect to π, that $E_\pi V \le b/(1-\lambda)$. We make use of this fact in Example 2 in Section 5, simplifying our original analysis. Furthermore, in Section 6 we state (Theorem 12) a modified version of our theorem based on this approach.

2. The inequality in Lemma 4 is stated in terms of Markov chains defined jointly, as described in Theorem 1. However, it clearly suffices to verify the inequality for Markov chains $(X'^{(k)}, Y'^{(k)})$ with a different joint definition, provided that the corresponding quantity N'_k is stochastically dominated by N_k. Furthermore, if the function h is of the additive form $h(x, y) = h_1(x) + h_2(y)$, then the joint structure of the two Markov chains does not matter. This is the case for both of our examples and also for Theorem 12.

We close this section with two lemmas that may help to establish a minorization condition $(*)$ in certain examples. Part (a) of the next lemma is not used in the examples presented in Section 5 herein, but it was used in earlier work (Rosenthal 1991, 1993). The other parts of the lemmas are used in Section 5.

Lemma 6.

(a) Suppose that a Markov transition kernel P on a state space $\mathcal{X}$ satisfies

$$P^{k_1}(x, R_2) \ge \varepsilon_1 \quad \forall\, x \in R_1$$

and

$$P^{k_2}(x, \cdot) \ge \varepsilon_2 Q(\cdot) \quad \forall\, x \in R_2,$$

for some probability measure $Q(\cdot)$ on $\mathcal{X}$. Then condition $(*)$ is satisfied with $k_0 = k_1 + k_2$, with $R = R_1$, and with $\varepsilon = \varepsilon_1 \varepsilon_2$.

(b) Given a positive integer k_0 and a subset $R \subseteq \mathcal{X}$, there exists a probability measure $Q(\cdot)$ so that

$$P^{k_0}(x, \cdot) \ge \varepsilon Q(\cdot) \quad \forall\, x \in R,$$

where

$$\varepsilon = \int_{\mathcal{X}} \left(\inf_{x \in R} P^{k_0}(x, dy)\right).$$

Finally, we intend to apply our method to the Gibbs sampler, where there are typically n random variables $X_1, \ldots, X_n$, which are updated repeatedly by

$$X_i^{(k)} \sim \mathcal{L}(X_i | X_j = X_j^{(k-1)} \text{ for } j < i, \text{ and } X_j = X_j^{(k)} \text{ for } j > i),$$

with (say) X_i taking values in $\mathcal{X}_i$. If the updating is done sequentially (i.e., each step of the Markov chain corresponds to updating first $X_1^{(1)}$, then $X_2^{(1)}$, and so on up to $X_n^{(1)}$), then the following lemma may help to establish condition (*). It says essentially that under a certain independence assumption, if we establish condition (*) for $X_1, \ldots, X_d$, then we can conclude condition (*) for all the variables $X_1, \ldots, X_n$, with the same value of ε.

Lemma 7. Consider a sequentially updated Gibbs sampler, as earlier. Suppose that for some d, conditional on values for $X_1^{(k)}, \ldots, X_d^{(k)}$, the random variables $X_{d+1}^{(k)}, \ldots, X_n^{(k)}$ are independent for all $X_i^{(k')}$ for all $k' < k$. (For example, for the Gibbs sampler this always holds with $d = n - 1$.) Suppose further that there is $R \subseteq \mathcal{X}$, $\varepsilon' > 0$ and a probability measure $Q'(\cdot)$ on $\mathcal{X}_1 \times \cdots \times \mathcal{X}_d$ such that

$$\mathcal{L}(X_1^{(k_0)}, \ldots, X_d^{(k_0)} | (X_1^{(0)}, \ldots, X_n^{(0)}) = x) \geq \varepsilon' Q'(\cdot), \quad \forall\, x \in R.$$

Then there is a probability measure $Q(\cdot)$ on $\mathcal{X}$ such that

$$P^{k_0}(x, \cdot) \geq \varepsilon' Q(\cdot).$$

Remark. This lemma exploits the specific structure of a sequentially updated Gibbs sampler. We shall also take advantage of another aspect of this structure. At each iteration, the Gibbs sampler begins by replacing the value of $X_1^{(k-1)}$ with the new value X_1^k. Thus, once a given iteration is completed, the value of $X_1^{(k-1)}$ is no longer used and has no effect on the future behavior of the chain. This suggests that it is unnecessary for such quantities as the subset R and the function h to make any reference to the value of $X_1^{(k-1)}$. This idea is used in both of the examples in Section 5.

4. REGENERATION TIMES

There may be cases in which it is too difficult to apply the foregoing methods and thus obtain bounds on convergence in total variation distance. Part of the difficulty might come from having to control a function $h(X^{(k)}, Y^{(k)})$ of *two* Markov chains instead of just one chain. It is thus reasonable to ask whether useful information can be obtained by just considering a single Markov chain, rather than attempting to couple two different chains.

An interesting possibility was suggested by Mykland et al. (1992), following Athreya and Ney (1978) and Nummelin (1984), who used minorization conditions to introduce *regeneration times* into a run of an MCMC. In the present context, this corresponds to the following. Given $X^{(k)} = x$ and $X^{(k+k_0)} = y$, if $X^{(k)} \in R$, then introduce a regeneration at time k with probability $\varepsilon Q(dy)/P^{k_0}(x, dy)$. Let T_i be the ith such regeneration, subject to $T_i \geq T_{i-1} + k_0$. Then, as is well known, the distribution of $X^{(T_i)}$ will be precisely $Q(\cdot)$. Thus the tours *between* regeneration times are actually independent. Furthermore, the stationary distribution π will satisfy

$$E_\pi(g) = (1/\mu) E\left(\sum_{k=T_{i-1}+1}^{T_i} g(X^{(k)}) \right),$$

where $\mu = E(T_i - T_{i-1})$ is the expected time between regenerations.

This suggests (Mykland et al. 1992) that if we run the Markov chain for precisely j complete tours, then we may estimate $E_\pi(g)$ as an average of j different iid quantities, thereby simplifying the analysis considerably.

One implication of this approach is that if the Markov chain is not *started* at a regeneration point (i.e., with initial distribution $Q(\cdot)$), then the initial values of $g(X^{(k)})$, before the first regeneration point, must be discarded. We believe that this provides an interesting resolution of the problem of *burn-in period* (i.e., the fact that the initial values in any MCMC run are too closely correlated with the starting distribution and thus should not be used for drawing inferences about the stationary distribution). Here a *random* number of initial iterations must be discarded. This corresponds to down-weighting the initial iterations in an interesting way.

A potential limitation of this approach is that it is unclear (at the beginning) how many iterations will be required to complete j tours. In this section we shall show that techniques similar to those of the previous section, but simpler to apply, can be used to bound exponential moments of the intervals $T_i - T_{i-1}$ between regeneration times. They thus provide exponential bounds on the waiting time until j tours are completed. This has the advantage that the target number of tours can be specified in advance, which avoids biases (related to the waiting-time paradox) associated with discarding an incomplete final tour.

We prove the following.

Theorem 8. Suppose that a Markov chain $P(x, dy)$ on a state space $\mathcal{X}$ satisfies condition (*) for some R, ε, and $Q(\cdot)$, and in addition has the property that for some function $h: \mathcal{X} \to R$ with $h \geq 1$, and some $\alpha > 1$,

$$E(h(X^{(1)}) | X^{(0)} = x) \leq \alpha^{-1} h(x), \quad \forall\, x \notin R.$$

Let T_1 be the time of the first regeneration as described earlier, and for $i > 1$ let T_i be the time of the first regeneration with $T_i \geq T_{i-1} + k_0$. Then if $(1 - \varepsilon)\alpha^{k_0} S_R < 1$, then

$$E(\alpha^{T_1}) \leq \frac{\varepsilon E(h(X^{(0)}))}{1 - (1 - \varepsilon)\alpha^{k_0} S_R},$$

and for $i > 1$,

$$E(\alpha^{T_i - T_{i-1}} | T_1, \ldots, T_{i-1}) \leq \frac{\varepsilon \alpha^{k_0} S_R}{1 - (1 - \varepsilon)\alpha^{k_0} S_R},$$

where

$$S_R = \sup_{x \in R} E(h(X^{(1)}) | X^{(0)} = x).$$

Note that in this lemma it is not necessary to consider a second chain $Y^{(k)}$ in stationary distribution. This simplifies the analysis in several places. It does of course come at the expense of no longer giving information directly about convergence in total variation distance.

This lemma immediately implies information about the time required to complete a particular number j of tours. Indeed, similar to Lemma 3 is the following corollary.

Corollary 9. Let U_k be the number of regenerations of our Markov chain up to time k. Then

$$P(U_k < j) \leq \alpha^{-k}\left(\frac{\varepsilon E(h(X^{(0)}))}{1-(1-\varepsilon)\alpha^{k_0}S_R}\right)\left(\frac{\varepsilon\alpha^{k_0}S_R}{1-(1-\varepsilon)\alpha^{k_0}S_R}\right)^{j-1}.$$

This corollary thus provides an exponential upper bound on the number of iterations required to complete j tours.

If we are estimating the mean $E_\pi(g)$ of a function g that is *bounded,* then Theorem 8 provides bounds on exponential moments of the iid quantities $\sum_{k=T_{i-1}+1}^{T_i} g(X^{(k)})$ that are being averaged. It can thus be used to get quantitative bounds on the error of the estimate after completing j tours, either through exponential bounds such as Cramér's theorem (see Dembo and Zeitouni 1993, sec. 2.2.1) or through standard use of Chebychev's inequality (because exponential moments imply second moments). This appears to be an interesting area for further research.

Finally, one can ask whether it is possible to obtain quantitative bounds on the convergence of $\mathcal{L}(X^{(k)})$ (as opposed to ergodic averages) to its stationary distribution, solely from information about the regeneration times as earlier. (It is then necessary to consider periodicity issues, which complicates the analysis.) A similar issue was considered by Lindvall (1992, thm. II.4.2), where finiteness of the moments of the coupling time are shown to follow from finiteness of corresponding moments (of order one less) of the return time. However, that work does not appear to extend easily to quantitative bounds and is thus difficult to apply in the present context. We leave this as an open question.

5. EXAMPLES

In this section Theorem 5 is applied to two examples involving the Gibbs sampler: a bivariate normal model and a hierarchical Poisson model.

Example 1: Bivariate Normal Model

Schervish and Carlin (1992) analyzed a model in which (X_1, X_2) are bivariate normally distributed, with common mean μ, with variances 2 and 1, respectively, and with covariance 1. (We write this as $N(\binom{\mu}{\mu}, \left(\begin{smallmatrix}2 & 1\\ 1 & 1\end{smallmatrix}\right))$.) The conditional distributions are thus given by

$$\mathcal{L}(X_1 | X_2 = x) = N(x, 1)$$

and

$$\mathcal{L}(X_2 | X_1 = x) = N\left(\frac{x+\mu}{2}, \frac{1}{2}\right).$$

Schervish and Carlin suggested running a Gibbs sampler on these two random variables, as follows. Given a value for $X_2^{(0)}$ (perhaps chosen from some initial distribution), generate $X_1^{(1)}$ from $N(X_2^{(0)}, 1)$, then generate $X_2^{(1)}$ from $N[(X_1^{(1)} + \mu)/2, \frac{1}{2}]$, then generate $X_1^{(2)}$ from $N(X_2^{(1)}, 1)$, then generate $X_2^{(2)}$ from $N[(X_1^{(2)} + \mu)/2, \frac{1}{2}]$, and so on.

In analyzing this use of the Gibbs sampler, one can ask whether $\mathcal{L}(X_1^{(k)}, X_2^{(k)})$ (the distribution of the Gibbs sampler after k iterations) converges to $N(\binom{\mu}{\mu}, \left(\begin{smallmatrix}2 & 1\\ 1 & 1\end{smallmatrix}\right))$ and at what rate. Now, this example is simple enough to permit an exact analysis (Schervish and Carlin 1992, thm. 4). However, it is instructive to proceed using the general method outlined earlier. We prove the following quantitative exponential bound on total variation distance. Because the law of $X_1^{(k+1)}$ is completely determined by the law of $X_2^{(k)}$, we concentrate here on the convergence of $\mathcal{L}(X_2^{(k)})$.

Theorem 10. The total variation distance between the distribution of $X_2^{(k)}$, when started in the initial distribution ν, and the true distribution of X_2 satisfies

$$\|\mathcal{L}(X_2^{(k)}) - N(\mu, 1)\|_{\text{var}} \leq (.964)^k + (.953)^k(2 + E_\nu(x_2 - \mu)^2).$$

Remark. The exact analysis of Schervish and Carlin (1992, thm. 4) indicates that this total variation distance is actually decreasing at the rate $(.5)^k$. Our results thus are not sharp, although they are within a factor of about 20.

Example 2: Hierarchical Poisson Model

Here we analyze the Gibbs sampler applied to a hierarchical Poisson model corresponding to failures in pumps at nuclear power plants. We use the same data studied by Gaver and O'Muircheartaigh (1987) using an empirical Bayes approach, and by Gelfand and Smith (1990), Tierney (1991), and Mykland, et al. (1992) using the Gibbs sampler.

The Gibbs sampler for this model is a Markov chain, $(\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)})_{k\geq 0}$ on $\mathcal{X} = (\mathbf{R}^{\geq 0})^{11}$, with updating scheme given (following Tierney 1991, sec. 5) by

$$\mathcal{L}(\beta^{(k+1)} | \{\theta_j^{(k)}\}) = G\left(\gamma + 10\alpha_0, \delta + \sum_{j=1}^{10}\theta_j^{(k)}\right),$$

$$\mathcal{L}(\theta_i^{(k+1)} | \beta^{(k+1)}, \{\theta_j^{(k+1)}\}_{j<i}, \{\theta_j^{(k)}\}_{j>i}) = G(\alpha_0 + s_i, t_i + \beta^{(k+1)}), \qquad (1 \leq i \leq 10),$$

where $G(a, b)$ denotes the gamma distribution with density $b^a x^{a-1} e^{-bx}/\Gamma(a)$, where $\alpha_0 = 1.802$, $\gamma = .01$, and $\delta = 1$, and with the data s_i and t_i as in Gelfand and Smith (1990, table 3). (Note that we write "α_0" rather than the usual "α" to avoid confusion with the α of Theorem 5.) Starting with initial values $\beta^{(0)}, \theta_1^{(0)}, \ldots, \theta_{10}^{(0)}$ (chosen from some initial distribution), the Markov chain proceeds by updating each of these random variables in turn, from these conditional distributions, for $k = 0, 1, 2, \ldots$.

Because we shall make use of this property, we note explicitly that (as in the Remark following Lemma 7) once a given kth iteration is completed, the value of $\beta^{(k)}$ is not used further and has no effect on the future behavior of the chain.

For this Markov chain, we prove the following.

Theorem 11. The total variation distance between the distribution of this Gibbs sampler after k iterations when started in the initial distribution ν and the true stationary distribution π satisfies

$$\|\mathcal{L}(\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)}) - \pi\|_{\text{var}} \leq (.976)^k + (.951)^k(6.2 + E_\nu((S^{(0)} - 6.5)^2)),$$

where $S^{(0)} = \sum_i \theta_i^{(0)}$.

6. A SIMPLIFICATION OF THE MAIN RESULT

Since originally completing this manuscript, we have realized that our main result (Theorem 5) can be stated in another form, using a drift condition on the original chain rather than on the coupled chain. This new form is inspired by the work of Meyn and Tweedie (1993b). (I am very grateful to Richard Tweedie for discussing these matters with me.)

A self-contained version of this new form of our result is the following.

Theorem 12. Suppose that a Markov chain $P(x, dy)$ on a state space $\mathcal{X}$ satisfies the drift condition

$$E(V(X^{(1)}) | X^{(0)} = x) \leq \lambda V(x) + b, \qquad x \in \mathcal{X}$$

for some $V : \mathcal{X} \to \mathbf{R}^{\geq 0}$ and some $\lambda < 1$ and $b < \infty$, and further satisfies a minorization condition

$$P(x, \cdot) \geq \varepsilon Q(\cdot), \quad \forall\, x \in \mathcal{X} \quad \text{with} \quad V(x) \leq d,$$

for some $\varepsilon > 0$, some probability measure $Q(\cdot)$ on $\mathcal{X}$, and some $d > 2b/(1 - \lambda)$. Then for any $0 < r < 1$, beginning in the initial distribution ν, we have

$$\|\mathcal{L}(X^{(k)} - \pi\|_{\text{var}} \leq (1 - \varepsilon)^{rk} + (\alpha^{-(1-r)}A^r)^k\left(1 + \frac{b}{1 - \lambda} + E_\nu(V(X_0))\right),$$

where

$$\alpha^{-1} = \frac{1 + 2b + \lambda d}{1 + d} < 1; \qquad A = 1 + 2(\lambda d + b).$$

7. CONCLUSIONS

In this article we have presented a general result (Theorem 5) giving upper bounds on the distance to stationarity of a Markov chain. We have provided two examples illustrating how this result can be applied to MCMC to provide rigorous, a priori upper bounds on the required running times. It is our hope that this method can be applied in the future to other, more complicated examples of MCMC.

APPENDIX: PROOFS

Proof of Theorem 1

We take $k_0 = 1$; the extension to general k_0 is straightforward.

The proof uses a coupling approach. (For background on coupling, see Asmussen 1990, Diaconis 1988, chap. 4E, Lindvall 1992, Pitman 1976, or Rosenthal 1991, appendix.) We begin by constructing $X^{(n)}$ and $Y^{(n)}$ simultaneously as follows. Let $X^{(0)}$ and $Y^{(0)}$ be chosen from the given initial distribution. For each time $n \geq 0$, given $X^{(n)}$ and $Y^{(n)}$, flip a coin with probability of heads equal to ε. If $X^{(n)}$ and $Y^{(n)}$ are both in R, and if the coin comes up heads, then choose a point $x \in \mathcal{X}$ according to the probability measure $Q(\cdot)$, set $X^{(n+1)} = Y^{(n+1)} = x$, and let the processes update so that they remain equal for all future times. If $X^{(n)}$ and $Y^{(n)}$ are both in R but the coin comes up tails, then choose $X^{(n+1)}$ and $Y^{(n+1)}$ independently, according to the "complementary" measures $(P(X^{(n)}, \cdot) - \varepsilon Q(\cdot))/(1 - \varepsilon)$ and $(P(Y^{(n)}, \cdot) - \varepsilon Q(\cdot))/(1 - \varepsilon)$, respectively. (Such a definition makes sense because $(*)$ holds.) Finally, if $X^{(n)}$ and $Y^{(n)}$ are *not* both in R, then simply update them independently, according to $P(X^{(n)}, \cdot)$ and $P(Y^{(n)}, \cdot)$ respectively, ignoring the coin flip.

It is easily checked that $X^{(n)}$ and $Y^{(n)}$ are each marginally updated according to the transition kernel P. Furthermore, $X^{(n)}$ and $Y^{(n)}$ are *coupled* the first time (call it T) that we choose them both from $Q(\cdot)$ as earlier. It now follows from the *coupling inequality* that

$$\|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\text{var}} \leq P(X^{(k)} \neq Y^{(k)}) \leq P(T > k).$$

Now it follows by construction that each time $X^{(n)}$ and $Y^{(n)}$ are both inside R, there is probability ε that they will couple on the next update. Thus, because

$$N_k = \#\{m < k : (X^{(m)}, Y^{(m)}) \in R \times R\},$$

we have that

$$P(T > k \text{ and } N_k \geq j) \leq (1 - \varepsilon)^j,$$

and hence that

$$P(T > k) \leq (1 - \varepsilon)^j + P(N_k < j),$$

completing the proof.

Proof of Lemma 3

This follows immediately from the fact that

$$P(N_k < j) = P(r_1 + \cdots + r_j > k) = P(\alpha^{r_1 + \cdots + r_j} > \alpha^k)$$

and from Markov's inequality.

Proof of Lemma 4

The hypotheses of the lemma imply that (with t_i as in Lemma 3) the function

$$g_i(k) = \alpha^k h(X^{(k)}, Y^{(k)}), \qquad k \leq t_i$$
$$= 0, \qquad k > t_i$$

has nonincreasing expectation as a function of k, at least for $k \geq t_{i-1} + k_0$. Statement (a) now follows (recalling that $h \geq 1$) from the fact that $E\alpha^{r_1} \leq Eg_1(r_1) \leq Eg_1(0)$. Similarly, statement (b) follows from the fact that

$$\begin{aligned} &E(\alpha^{r_i} | X^{(t_{i-1})}, Y^{(t_{i-1})}) \\ &= E(\alpha^{t_i - t_{i-1}} | X^{(t_{i-1})}, Y^{(t_{i-1})}) \\ &\leq E(\alpha^{-t_{i-1}} g_i(t_i) | X^{(t_{i-1})}, Y^{(t_{i-1})}) \\ &\leq E(\alpha^{-t_{i-1}} g_i(t_{i-1} + k_0) | X^{(t_{i-1})}, Y^{(t_{i-1})}) \\ &= \alpha^{k_0} E(h(X^{(t_{i-1}+k_0)}, Y^{(t_{i-1}+k_0)}) | X^{(t_{i-1})}, Y^{(t_{i-1})}) \\ &\leq \alpha^{k_0} \sup_{(x,y) \in R \times R} E(h(X^{(1)}, Y^{(1)}) | X^{(0)} = x, Y^{(0)} = y). \end{aligned}$$

Proof of Lemma 6

Part (a) is obvious. For part (b), define the measure $Q'(\cdot)$ on $\mathcal{X}$ by

$$Q'(A) = \int_A (\inf_{x \in R_2} P^{k_2}(x, dy)).$$

Then it is easily seen that $P^{k_2}(x, \cdot) \geq Q'(\cdot)$ for $x \in R$. Assuming that $Q'(\mathcal{X}) > 0$ (otherwise, the lemma is vacuously true), the result now follows by setting $Q(\cdot) = Q'(\cdot)/Q'(\mathcal{X})$ and setting $\varepsilon_2 = Q'(\mathcal{X})$.

Proof of Lemma 7

We define the measure $Q(\cdot)$ as follows. Marginally on the first d coordinates, $Q(\cdot)$ agrees with $Q'(\cdot)$. Conditional on the first d coordinates, $Q(\cdot)$ is defined by

$$Q(X_{x+1}, \ldots, X_n | X_1, \ldots, X_d) = \mathcal{L}(X_{d+1}, \ldots, X_n | X_1, \ldots, X_d).$$

By the independence hypothesis, the minorization condition for $Q'(\cdot)$ implies the minorization condition for $Q(\cdot)$.

Proof of Theorem 8

By reasoning similar to Lemma 4, letting r_i be the ith waiting time (subject to $r_i \geq k_0$) to return to the set R, we have that $E(\alpha^{r_1}) \leq E(h(X^{(0)}))$ and $E(\alpha^{r_i}) \leq \alpha^{k_0} S_R$. Now each time that the chain is inside R, it has probability ε of regenerating. Thus, letting F be the number of times the Markov chain is inside R (after waiting at least time k_0) before the next regeneration, F is a geometrically distributed random variable with parameter ε. Setting $m_0 = E(h(X^{(0)}))$, we have that

$$E(\alpha^{T_1}) \leq m_0 E(\alpha^{k_0} S_R)^F = m_0 \sum_{l=0}^{\infty} \varepsilon(1-\varepsilon)^l (\alpha^{k_0} S_R)^l$$

$$= \frac{\varepsilon m_0}{1 - (1-\varepsilon)\alpha^{k_0} S_R},$$

as desired. The second statement follows similarly.

Proof of Theorem 10

We begin by noting (using $E(X^2) = (EX)^2 + \operatorname{var}(X)$) that

$$\begin{aligned} &E((X_2^{(1)} - \mu)^2 | X_2^{(0)} = x_2) \\ &\quad = E(E((X_2^{(1)} - \mu)^2 | X_1^{(1)}) | X_2^{(0)} = x_2) \\ &\quad = E\left(\left(\frac{X_1^{(1)} - \mu}{2}\right)^2 + \left(\frac{1}{2}\right) \middle| X_2^{(0)} = x_2\right) \\ &\quad = \frac{1}{4}(x_2 - \mu)^2 + \frac{3}{4}. \end{aligned}$$

(Of course, here the simple nature of the problem makes this computation easy. In a more complicated situation, such as Example 2, this quantity may have to be estimated, numerically or otherwise. A good *upper bound* on the quantity is all that is required.)

Recall (see the Remark following Theorem 7) that because at each iteration the old value $X_1^{(k)}$ is discarded, our subset R and function h should refer only to the second components x_2 and y_2.

Thus, setting $h(x, y) = 1 + (x_2 - \mu)^2 + (y_2 - \mu)^2$ and considering two independent versions $X^{(k)} = (X_1^{(k)}, X_2^{(k)})$ and $Y^{(k)} = (Y_1^{(k)}, Y_2^{(k)})$ of the chain, we have that

$$E(h(X^{(1)}, Y^{(1)}) | X_2^{(0)} = x_2, Y_2^{(0)} = y_2) = \frac{9}{4} + \left(\frac{1}{4}\right) h(x, y).$$

Hence if we set $R = \{x \in \mathcal{X} | (x_2 - \mu)^2 \leq 3\}$, then if $(x, y) \notin R \times R$, $h(x, y) \geq 4$ and hence

$$E(h(X^{(1)}, Y^{(1)}) | X^{(0)} = x, Y^{(0)} = y) \leq \left(\frac{13}{16}\right) h(x, y).$$

Hence we can take $\alpha = \frac{16}{13}$.

To continue, note that

$$\begin{aligned} A &= \sup_{(x,y) \in R \times R} E(h(X^{(1)}, Y^{(1)}) | X^{(0)} = x, Y^{(0)} = y) \\ &= \left(\frac{9}{4}\right) + \left(\frac{1}{4}\right)(7) = 4. \end{aligned}$$

Furthermore, because the stationary distribution for Y_2 is $N(\mu, 1)$, we have that $E_\pi(Y_2 - \mu)^2 = 1$, so that $E_{\nu \times \pi}(h(X^{(0)}, Y^{(0)})) = 2 + E_\nu(x_2 - \mu)^2$. (Again, in a more complicated example these quantities may have to be estimated, perhaps using the first Remark after Theorem 5, but *bounds* on them are all that is required.)

A value for ε can be obtained from Lemma 6(b). Indeed, we can take

$$\begin{aligned} \varepsilon &= \int \left(\inf_{x \in R} N\left(\frac{x_2 + \mu}{2}, \frac{3}{4}; y\right) \right) dy \\ &= \int_{-\infty}^{0} N\left(\frac{\sqrt{3}}{2}, \frac{3}{4}; y\right) dy + \int_{0}^{\infty} N\left(-\frac{\sqrt{3}}{2}, \frac{3}{4}; y\right) dy, \end{aligned}$$

where $N(a, b; y) = (1/\sqrt{2\pi b}) e^{-(y-a)^2/2b}$ is the density function of $N(a, b)$. This last expression is just the probability that a normal random variable will be more than one standard deviation away from its mean, and is thus well known to be $\geq .31$.

We now apply Theorem 5 with $k_0 = 1$, $A = 4$, $\alpha = \frac{16}{13}$, and $\varepsilon = .31$. We choose $j = k/10$. Because $(.69)^{1/10} < .964$ and $(16/13)^{-9/10} 4^{1/10} < .953$, the result now follows from Theorem 5.

Proof of Theorem 11

To begin the analysis, note that the $\theta_i^{(k)}$ are conditionally independent given the value of $\beta^{(k-1)}$. Using this and recalling that $G(a, b)$ has mean a/b and variance a/b^2, it is easily verified (writing $S^{(k)}$ for $\sum_i \theta_i^{(k)}$) that

$$E(\beta^{(k+1)} | S^{(k)}) = \frac{\gamma + 10\alpha_0}{\delta + S^{(k)}},$$

$$\operatorname{var}(\beta^{(k+1)} | S^{(k)}) = \frac{\gamma + 10\alpha_0}{(\delta + S^{(k)})^2},$$

$$E(S^{(k+1)} | \beta^{(k)}) = \sum_i \frac{\alpha_0 + s_i}{t_i + \beta^{(k)}},$$

and

$$\operatorname{var}(S^{(k+1)} | \beta^{(k)}) = \sum_i \frac{\alpha_0 + s_i}{(t_i + \beta^{(k)})^2}.$$

Note that although the random variables involved here are not themselves bounded, the conditional means and variances *are* bounded. This suggests that it should be possible to apply Proposition 2 directly. Indeed, using Chebychev's inequality it is straightforward to establish a condition $(*)$ on the entire state space. Unfortunately, it appears to be very difficult to obtain a value of ε that is not extremely small. Thus we consider the other methods developed in this article.

Recall (see the Remark following Theorem 7) that because at each iteration the old value $\beta^{(k)}$ is discarded, the subset R and function h should refer only to the remaining components $\theta_1^{(k)}, \ldots, \theta_n^{(k)}$. Indeed, it is sufficient to refer only to their sum $S^{(k)}$.

A cursory numerical examination of the conditional means (for the given data) suggests that the value of $S^{(k)}$ roughly approaches the value 6.5. Thus, writing the two Markov chains as $X^{(k)} = (\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)})$ and $Y^{(k)} = (\beta'^{(k)}, \theta_1'^{(k)}, \ldots, \theta_{10}'^{(k)})$, with $S^{(k)} = \sum_i \theta_i^{(k)}$ and $S'^{(k)} = \sum_i \theta_i'^{(k)}$, we set

$$h(X^{(k)}, Y^{(k)}) = 1 + (S^{(k)} - 6.5)^2 + (S'^{(k)} - 6.5)^2.$$

To proceed, it is necessary to control quantities of the form

$$E(h(X^{(1)}, Y^{(1)}) | X^{(0)}, Y^{(0)}).$$

Because the Markov chain proceeds by first replacing the value $\beta^{(0)}$ by a new value $\beta^{(1)}$, it is easily seen that this quantity will depend

only on the values of $S^{(0)}$ and $S'^{(0)}$, so we must proceed accordingly. We define the function $e(w)$ by

$$\begin{aligned} e(w) &= E((S^{(1)} - 6.5)^2 | S^{(0)} = w) \\ &= \int_0^\infty E((S^{(1)} - 6.5)^2 | \beta^{(1)} = x) P(\beta^{(1)} = dx | S^{(0)} = w) \\ &= \int_0^\infty \left[\left(\sum_i \left(\frac{\alpha_0 + s_i}{t_i + x} \right) - 6.5 \right)^2 + \sum_i \left(\frac{\alpha_0 + s_i}{(t_i + x)^2} \right) \right] \\ &\quad \times G(\gamma + 10\alpha_0, \delta + w; x)\, dx, \end{aligned}$$

where we have used the conditional mean and variance of the θ_i and the conditional distribution of the β as given earlier (and where $G(a, b; x) = b^a x^{a-1} e^{-bx}/\Gamma(a)$ is the density of the gamma distribution). Now the function $e(w)$ is difficult to handle analytically, but it is easily evaluated numerically. Integrating $e(w)$ numerically over a fine grid of values of w gives the following. The function $e(w)$ changes slowly as a function of w, with a unique minimum of about 1.40 near $w = 5.8$. We compute numerically that

$$e(4.0) < 1.90; \qquad e(9.0) < 2.29.$$

This suggests that we choose $R = \{X^{(k)} : 4.0 \le S^{(k)} \le 9.0\}$.

To proceed, we verify numerically (as will be important shortly) that

$$\sup_{w \notin [4.0, 9.0]} \left(\frac{1 + e(w)}{1 + (w - 6.5)^2} \right) < .46,$$

with the supremum obtained at $w = 9.0$. Also,

$$\sup_w \left(\frac{.46}{1 + (w - 6.5)^2/7.25} + \frac{e(w)}{7.25 + (w - 6.5)^2} \right) < .66,$$

with the supremum obtained near $w = 6.6$ (though there is a competing upturn to .405 near $w = 0$). Hence

$$\begin{aligned} &\sup_{(x,y) \notin R \times R} \left(\frac{E(h(X^{(1)}, Y^{(1)}) | X^{(0)} = x, Y^{(0)} = y)}{h(x, y)} \right) \\ &= \sup_{\substack{w_1, w_2 \\ w_1 \notin [4.0, 9.0]}} \left(\frac{1 + e(w_1) + e(w_2)}{1 + (w_1 - 6.5)^2 + (w_2 - 6.5)^2} \right) \\ &\le \sup_{w_2} \left[\left(\sup_{w_1 \notin [4.0, 9.0]} \left(\frac{1 + e(w_1)}{1 + (w_1 + 6.5)^2} \right) \Big/ (1 + (w_2 - 6.5)^2/7.25) \right) \right. \\ &\qquad \left. + \left(\frac{e(w_2)}{7.25 + (w_2 - 6.5)^2} \right) \right] \\ &< .66, \end{aligned}$$

where we have used the foregoing numerical bounds and have also used the fact that $1 + (w_1 - 6.5)^2 \ge 7.25$. Hence we can choose $\alpha = 1/.66 > 1.5$.

We compute a value for ε using Lemma 6 (b) and Lemma 7 (with $d = 1$). We have (using that for fixed a and x, $G(a, b; x)$ is unimodal as a function of b) that

$$\begin{aligned} \varepsilon &= \int_0^\infty \left(\inf_{w \in [4.0, 9.0]} G(\gamma + 10\alpha_0, \delta + w; x) \right) dx \\ &= \int_0^\infty \min(G(\gamma + 10\alpha_0, \delta + 4.0; x), \\ &\qquad G(\gamma + 10\alpha_0, \delta + 9.0; x))\, dx \\ &> .14, \end{aligned}$$

where again we have done the integration numerically.

In the context of Theorem 5, because $\sup_{w \in R} e(w) < 2.3$, we have $A < 1 + 2.3 + 2.3 = 5.6$.

Finally, we need to bound $E_\pi((S'^{(0)} - 6.5)^2)$. Using the stationarity of π, we have the crude bound

$$\begin{aligned} E_\pi((S'^{(0)} - 6.5)^2) &\le \sup_x E((S - 6.5)^2 | \beta = x) \\ &= E((S^{(1)} - 6.5)^2 | \beta^{(1)} = 0) < 43. \end{aligned}$$

We can do better using the Remark following Theorem 5. Setting $V(X) = 1 + (S - 6.5)^2$, our previous calculations indicate that we will have $E(V(X^{(1)}) | X^{(0)} = x) \le \lambda V(x) + b$ with $\lambda = .46$ and $b = 3.3$. The Remark then gives $E_\pi(S - 6.5)^2 \le b/(1 - \lambda) < 6.2$. It follows that

$$E_{\nu \times \pi}(h(X^{(0)}, Y^{(0)}) < 6.2 + E((S^{(0)} - 6.5)^2).$$

We now apply Theorem 5, with $k_0 = 1$, $\varepsilon = .14$, $\alpha = 1.5$, $A = 5.6$, and $j = k/6$. Because $(0, 86)^{1/6} < .976$ and $(1.5)^{-5/6}(5.6)^{1.6} < .951$, the result follows.

Proof of Theorem 12

We set $h(x, y) = 1 + V(x) + V(y)$ and set $R = \{x \in \mathcal{X} \mid V(x) \le d\}$. Then if $(x, y) \notin R \times R$, then $h(x, y) \ge 1 + d$. Thus, in terms of a coupled chain as in Lemma 4, we have

$$\begin{aligned} &E(h(X^{(1)}, Y^{(1)}) | X^{(0)} = x, Y^{(0)} = y) \\ &\quad \le 1 + \lambda V(x) + \lambda V(y) + 2b \\ &\quad \le \left(\lambda + \frac{1 - \lambda + 2b}{1 + d} \right) h(x, y) = \left(\frac{1 + 2b + \lambda d}{1 + d} \right) h(x, y), \end{aligned}$$

so the hypothesis of Lemma 4 is satisfied with h, α, and R as given.

Furthermore, with A as in Theorem 5, we have

$$A = 1 + 2 \sup_{x \in R} E(V(X^{(1)}) | X^{(0)} = x) \le 1 + 2(\lambda d + b).$$

Finally, using the Remark following Theorem 5, we have that

$$E_{\nu \times \pi}(h(X^{(0)}, Y^{(0)})) \le 1 + E_\nu(V(X^{(0)})) + \frac{b}{1 - \lambda}.$$

Setting $j = rk + 1$, Theorem 12 now follows directly from Theorem 5.

[Received October 1993. Revised April 1994.]

REFERENCES

Amit, Y. (1991), "On the Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions," *Journal of Multivariate Analysis*, 38, 89–99.

——— (1993), "Convergence Properties of the Gibbs Sampler for Perturbations of Gaussians," Technical Report 352, University of Chicago, Dept. of Statistics.

Amit, Y., and Grenander, U. (1991), "Comparing Sweep Strategies for Stochastic Relaxation," *Journal of Multivariate Analysis*, 37, 197–222.

Applegate, D., Kannan, R., and Polson, N. G. (1990), "Random Polynomial Time Algorithms for Sampling From Joint Distributions," Technical Report 500, Carnegie-Mellon University, School of Computer Science.

Asmussen, S. (1987), *Applied Probability and Queues*, New York: John Wiley.

Athreya, K. B., McDonald, D., and Ney, P. (1978), "Limit Theorems for Semi-Markov Processes and Renewal Theory for Markov Chains," *The Annals of Probability*, 6, 788–797.

Athreya, K. B., and Ney, P. (1978), "A New Approach to the Limit Theory of Recurrent Markov Chains," *Transactions of the American Mathematical Society*, 245, 493–501.

Baxter, J. R., and Rosenthal, J. S. (1994), "Rates of Convergence for Everywhere-Positive Markov Chains," Technical Report 9406, University of Toronto, Dept. of Statistics.

Belsley, E. D. (1993), "Rates of Convergence of Markov Chains Related to Association Schemes," Ph.D. dissertation, Harvard University, Dept. of Mathematics.

Dembo, A., and Zeitouni, O. (1993), *Large Deviations Techniques and Applications,* Boston: Jones and Bartlett.

Diaconis, P. (1988), *Group Representations in Probability and Statistics,* Hayward, CA: IMS.

Diaconis, P., and Hanlon, P. (1992), "Eigen Analysis for Some Examples of the Metropolis Algorithm," technical report, Harvard University, Dept. of Mathematics.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* Vol. II (2nd ed.), New York: John Wiley.

Frieze, A., Kannan, R., and Polson, N. G. (1993), "Sampling From Log-Concave Distributions," Technical report, Carnegie-Mellon University, School of Computer Science.

Frigessi, A., Hwang, C.-R., and Younes, L. (1992), "Optimal Spectral Structure of Reversible Stochastic Matrices, Monte Carlo Methods, and the Simulation of Markov Random Fields," *Annals of Applied Probability,* 2, 610–628.

Gaver, D., and O'Muircheartaigh, I. (1987), "Robust Empirical Bayes Analysis of Event Rates," *Technometrics,* 29, 1–15.

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. (1992), "Inference From Iterative Simulation Using Multiple Sequences, *Statistical Science,* 7, 457–472.

Geman, S., and Geman, D. (1984), "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* PAMI-6, 721–741.

Geyer, C. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science,* 7, 473–483.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika,* 57, 97–109.

Lindvall, T. (1992), *Lectures on the Coupling Method,* New York: John Wiley.

Liu, J. (1992), "Eigen Analysis for a Metropolis Sampling Scheme With Comparisons to Rejection Sampling and Importance Resampling," Research Report R-427, Harvard University, Dept. of Statistics.

Liu, J., Wong, W., and Kong, A. (1991a), "Correlation Structure and the Convergence of the Gibbs Sampler, I." Technical Report 299, University of Chicago, Dept. of Statistics.

——— (1991b), "Correlation Structure and the Convergence of the Gibbs Sampler, II: Applications to Various Scans," Technical Report 304, University of Chicago, Dept. of Statistics.

Matthews, P. (1993), "A Slowly Mixing Markov Chain With Implications for Gibbs Sampling," *Statistics and Probability Letters,* 17, 231–236.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953), "Equations of State Calculations by Fast Computing Machines," *Journal of Chemical Physics,* 21, 1087–1091.

Meyn, S. P., and Tweedie, R. L. (1993a), *Markov Chains and Stochastic Stability,* London: Springer-Verlag.

——— (1993b), "Computable Bounds for Convergence Rates of Markov Chains," Technical report, Colorado State University, Dept. of Statistics.

Mykland, P., Tierney, L., and Yu, B. (1992), "Regeneration in Markov Chain Samplers," Technical Report 585, University of Minnesota, School of Statistics.

Nummelin, E. (1984), *General Irreducible Markov Chains and Nonnegative Operators,* Cambridge, U.K.: Cambridge University Press.

Pitman, J. W. (1976), "On Coupling of Markov Chains," *Z. Wahrsch. Verw. Gebiete,* 35, 315–322.

Roberts, G. O. (1992), "Convergence Diagnostics of the Gibbs Sampler," in *Bayesian Statistics 4,* eds. J. M. Bernardo et al., Oxford, U.K.: Oxford University Press, pp. 777–784.

Roberts, G. O., and Polson, N. G. (1990), "A Note on the Geometric Convergence of the Gibbs Sampler," *Journal of the Royal Statistical Society,* Ser. B, 56, 377–384.

Rosenthal, J. S. (1991), "Rates of Convergence for Gibbs Sampler for Variance Components Models," Technical Report 9322, University of Toronto, Dept. of Statistics. Ann. Stat., to appear.

——— (1993), "Rates of Convergence for Data Augmentation on Finite Sample Spaces," *Annals of Applied Probability,* 3, 319–339.

Schervish, M. J., and Carlin, B. P. (1992), "On the Convergence of Successive Substitution Sampling," *Journal of Computational and Graphical Statistics,* 1, 111–127.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.

Tierney, L. (1991), "Markov Chains for Exploring Posterior Distributions," Technical Report 560, University of Minnesota, School of Statistics.