Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo

(September, 1993; revised July, 1994.)

(Appeared in *Journal of the American Statistical Association* **90** (1995), 558–566.)

by

Jeffrey S. Rosenthal*

Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1
Phone: (416) 978-4594. Internet: `jeff@utstat.toronto.edu`

Summary. We provide general methods 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. We apply our results to two models of the Gibbs sampler, the first a bivariate normal model, the second a hierarchical Poisson model (with gamma conditionals). Our methods use the notion of *minorization conditions* for Markov chains.

Keywords. Gibbs sampler; Metropolis-Hastings algorithm; Coupling; Harris recurrence; Drift condition; Bivariate normal model; Hierarchical Poisson model; Regeneration time.

Acknowledgements. I am very grateful to John Baxter for his suggestions regarding Lemma 4 herein. I thank Persi Diaconis, Jun Liu, Peter Ney, Richard Tweedie, and Gareth Roberts for very helpful conversations. Finally, I thank the referees for many excellent comments and suggestions.

* Partially supported through NSF grant DMS-90-02899, and through a research grant from NSERC of Canada.

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 (Metropolis et al., 1953; Hastings, 1970), and include the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) 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), describe general results about exponential convergence (but without giving quantitative bounds), using the theory of compact operators. Similar approaches are used by Liu et al. (1991a, 1991b) and Baxter and Rosenthal (1994). A "discretized" Markov chain is analyzed by Applegate, Kannan, and Polson (1990) and Frieze, Kannan, and Polson (1993), who prove polynomial bounds on certain running times. A method for estimating the variance of the chain is suggested in Geyer (1992). There have also been various papers giving quantitative bounds on convergence rates for specific models, including Amit and Grenander (1991), Amit (1991, 1993), Rosenthal (1993, 1991), Liu (1992), Frigessi et al. (1992), Diaconis and Hanlon (1992), and Belsley (1993, Chapter 6). In addition, various "convergence diagnostics" have been suggested by a number of authors, including Roberts (1992) and Gelman and Rubin (1992).

In this paper, we provide 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 (splits) for Markov chains (see section 2) to establish results about convergence of MCMC. This amounts to showing that the Markov chain satisfies a condition of the form $P^{k_0}(x, \cdot) \geq \epsilon Q(\cdot)$ 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 our 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 our 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). Our results thus relate to work of Mykland et al. (1992) who discuss 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), which is also discussed in 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. While our bounds are not sharp numerically, they are not too wildly off, and they could be of use in guiding a simulation.

In section 6, we present (Theorem 12) a simplified version of our main result, which involves verifying a simpler drift condition than does Theorem 5. Finally, the theorem proofs are contained in the Appendix.

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 obtain computable bounds on the distance to stationarity under certain conditions. Their methods require slightly less information than do 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 $\epsilon > 0$, such that

$$(*) \qquad P^{k_0}(x, A) \geq \epsilon Q(A), \qquad \text{for all } x \in R,$$

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

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

Most of the present paper is based on the following theorem. Special cases of the theorem were used in Rosenthal (1993, 1991) for similar purposes.

Theorem 1. *Suppose 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* $j > 0$,

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

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 $(1-\epsilon)^{[j/k_0]}$ in the bound will then decrease quickly as the number of iterations k gets large (assuming 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 below.

At the other extreme is when the condition $(*)$ is satisfied with $R = \mathcal{X}$, i.e. on the entire state space. (This is called the Doeblin condition, and is equivalent (Nummelin, 1984, Theorem 6.15; Tierney, 1991, Proposition 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)

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

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

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

This proposition is discussed in Nummelin (1984), Roberts and Polson (1990), and elsewhere. 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 only work for models with bounded state spaces. However, our Example

#2 below 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 be able 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 $m \le k$ for which $\left(X^{(m)}, Y^{(m)}\right) \in R \times R$).

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

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

However, to make this bound go to 0 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 paper 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, it 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, Chapter XI, Section 3). However, it is not true, for example, that the coupling bound in Theorem 1 can be taken as $(1-\epsilon)^{E(N_k)}$. (To see this, consider a case where N_k is equal to either zero or one million, each with probability 1/2.) Thus, the mean of N_k is insufficient to establish exponential convergence of the chain; more information is needed.

The approach 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 i'th gap between such times (i.e., the "k_0-delayed i'th return time to $R \times R$"). Then for any $\alpha > 1$,*

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

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, Proposition 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 i'th return time to $R \times R$" as above. Suppose there is $\alpha > 1$ and a function $h : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ such that $h \geq 1$ and*

$$E\left(h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y\right) \leq \alpha^{-1} h(x, y), \qquad \text{for all } (x, y) \notin R \times R.$$

Then

$$(i) \qquad E(\alpha^{r_1}) \leq E\left(h(X^{(0)}, Y^{(0)})\right),$$

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

$$(ii) \qquad E(\alpha^{r_i} \mid r_1, \ldots, r_{i-1}) \leq \alpha^{k_0} \sup_{(x,y) \in R \times R} E\left(h(X^{(1)}, Y^{(1)}) \,|\, X^{(0)} = x, Y^{(0)} = y\right),$$

Putting all of the above together, we obtain the following.

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

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

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

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1 - \epsilon)^{[j/k_0]} + \alpha^{-k + jk_0 - 1} A^{j-1} E_{\nu \times \pi}\left(h(X^{(0)}, Y^{(0)})\right).$$

(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 below 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 its i'th 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 requires that we bound the expected value $E_{\nu\times\pi}\left(h(X^{(0)}, Y^{(0)})\right)$ which unfortunately depends on the unknown distribution π. However, if we have verified a drift condition of the form $E(V(X^{(1)}) \mid X^{(0)} = x) \leq \lambda V(x) + b$, then it is easily seen (cf. Meyn and Tweedie, 1993b, Proposition 4.3 (i)), by taking expectations of both sides with respect to π, that $E_\pi V \leq \frac{b}{1-\lambda}$. We make use of this fact in Example #2 below, 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 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 below.

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

Lemma 6.

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

$$P^{k_1}(x, R_2) \geq \epsilon_1 \qquad \textit{for all } x \in R_1$$

and

$$P^{k_2}(x, \cdot) \geq \epsilon_2 \, Q(\cdot) \qquad \textit{for all } x \in R_2 \, ,$$

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* $\epsilon = \epsilon_1 \epsilon_2$.

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

$$P^{k_0}(x, \cdot) \geq \epsilon \, Q(\cdot) \qquad \textit{for all } x \in R \, ,$$

where

$$\epsilon \; = \; \int\limits_{\mathcal{X}} \left(\inf_{x \in R} \, P^{k_0}(x, dy) \right) \, .$$

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

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

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 above. 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 of 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* $R \subseteq \mathcal{X}$, $\epsilon' > 0$ *and a probability measure* $Q'(\cdot)$ *on* $\mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ *such that*

$$\mathcal{L}(X_1^{(k_0)}, \ldots, X_d^{(k_0)} \mid (X_1^{(0)}, \ldots, X_n^{(0)}) = x) \; \geq \epsilon' \, Q'(\cdot) \, , \qquad \textit{for all } x \in R \, .$$

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

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

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 above 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. It is thus reasonable to ask if useful information can be obtained by just considering a single Markov chain, rather than attempting to couple two different chains.

An interesting possibility is suggested in Mykland et al. (1992), following Athreya and Ney (1978) and Nummelin (1984), who use 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 $\epsilon\, Q(dy)/P^{k_0}(x, dy)$. Let T_i be the i'th 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 *i.i.d.* 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 should therefore 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 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 $h : \mathcal{X} \to \mathbf{R}$ with $h \geq 1$, and some $\alpha > 1$,*

$$E\left(h(X^{(1)}) \mid X^{(0)} = x\right) \;\leq\; \alpha^{-1}\, h(x), \qquad \textit{for all } x \notin R\,.$$

Let T_1 be the time of the first regeneration as described above, 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 $(1-\epsilon)\alpha^{k_0} S_R < 1$, then

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

and for $i > 1$,

$$E\left(\alpha^{T_i - T_{i-1}} \mid T_1, \ldots, T_{i-1}\right) \;\leq\; \frac{\epsilon\, \alpha^{k_0}\, S_R}{1 - (1-\epsilon)\alpha^{k_0} S_R}\,.$$

where

$$S_R \;=\; \sup_{x \in R} E\left(h(X^{(1)}) \mid X^{(0)} = x\right)\,.$$

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 we have

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

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

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 *i.i.d.* quantities $\sum_{k=T_{i-1}+1}^{T_i} g(X^{(k)})$ that we are averaging. It can thus be used to get quantitative bounds on the error of our estimate after completing j tours, either through exponential bounds such as Cramér's Theorem (see Dembo and Zeitouni, 1993, Section 2.2.1), or through standard use of Chebychev's inequality (since 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 above. (It is then necessary to consider periodicity issues, which complicates the analysis.) A similar issue is considered in Lindvall (1992, Theorem 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.

We here apply Theorem 5 to two examples involving the Gibbs sampler, one a bivariate normal model and the other a hierarchical Poisson model.

Example #1. Bivariate Normal Model.

Schervish and Carlin (1992) analyze 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 shall write this as $N\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\right)$.) The conditional distributions are thus given by

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

and

$$\mathcal{L}(X_2 \mid X_1 = x) \;=\; N(\frac{x+\mu}{2},\, 1/2)\,.$$

They suggest 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 $N(\frac{X_1^{(1)}+\mu}{2}, 1/2)$, then generate $X_1^{(2)}$ from $N(X_2^{(1)}, 1)$, then generate $X_2^{(2)}$ from $N(\frac{X_1^{(2)}+\mu}{2}, 1/2)$, and so on.

In analyzing this use of the Gibbs sampler, we can ask whether $\mathcal{L}(X_1^{(k)}, X_2^{(k)})$ (the distribution of the Gibbs sampler after k iterations) converges to $N\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\right)$ and at what rate. Now, this example is simple enough that it permits an exact analysis (Schervish and Carlin, 1992, Theorem 4). However, it is instructive to proceed using the general method outlined above. We prove the following quantitative exponential bound on total variation distance.

Theorem 10. *The total variation distance between the distribution of the Gibbs sampler after k iterations when started in the initial distribution ν, and the true joint distribution of (X_1, X_2), satisfies*

$$\|\mathcal{L}(X_1^{(k)}, X_2^{(k)}) \;-\; N\left(\begin{pmatrix}\mu\\ \mu\end{pmatrix}, \begin{pmatrix}2 & 1\\ 1 & 1\end{pmatrix}\right)\|_{\text{var}} \;\le\; (0.964)^k \;+\; (0.953)^k\left(2 + E_\nu(x_2-\mu)^2\right).$$

Remark. The exact analysis of Schervish and Carlin (1992, Theorem 4) indicates that this total variation distance is actually decreasing at the rate $(0.5)^k$. Our results are therefore not sharp, though they are within a factor of about 20.

Example #2. Hierarchical Poisson Model.

We analyze here 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, Section 5) by

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

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

where $G(a,b)$ denotes the gamma distribution with density $b^a x^{a-1} e^{-bx}/\Gamma(a)$, where $\alpha_0 = 1.802$, $\gamma = 0.01$, 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 $k = 0, 1, 2, \ldots$.

Since we shall make use of this property, we note explicitly that (as in the Remark following Lemma 7) once a given k'th 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*

$$\|\mathcal{L}(\beta^{(k)}, \theta_1^{(k)}, \ldots, \theta_{10}^{(k)}) \;-\; \pi\|_{\text{var}} \;\leq\; (0.976)^k \;+\; (0.951)^k (6.2 + E_\nu\left((S^{(0)} - 6.5)^2\right)),$$

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 a Markov chain $P(x, dy)$ on a state space $\mathcal{X}$ satisfies the drift condition*

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

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

$$P(x, \cdot) \geq \epsilon\, Q(\cdot)\,, \qquad \text{for all } x \in \mathcal{X} \text{ with } V(x) \leq d\,,$$

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

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

where

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

7. Conclusions.

In this paper 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 Markov chain Monte Carlo, 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 Pitman (1976), Diaconis (1988, Chapter 4E), Asmussen (1990), Lindvall (1992), 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, choose $X^{(n+1)}$ and $Y^{(n+1)}$ independently, according to the "complementary" measures $\left(P(X^{(n)}, \cdot) - \epsilon Q(\cdot)\right) / (1-\epsilon)$ and $\left(P(Y^{(n)}, \cdot) - \epsilon Q(\cdot)\right) / (1-\epsilon)$, 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 above. It now follows from the *coupling inequality* that

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

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, since

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

we have that

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

and hence that

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

completing the proof. ∎

Proof of Lemma 3. This follows immediately from the fact that

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

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

$$g_i(k) \;=\; \begin{cases} \alpha^k\, h(X^{(k)}, Y^{(k)}), & k \le t_i \\ 0, & k > t_i \end{cases}$$

has non-increasing expectation as a function of k, at least for $k \ge t_{i-1} + k_0$. Statement (i) now follows (recalling that $h \ge 1$) from the fact that $E\,\alpha^{r_1} \le E\,g_1(r_1) \le E\,g_1(0)$. Similarly, statement (ii) follows from the fact that

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

∎

Proof of Lemma 6. Part (i) is obvious. For part (ii), define the measure $Q'(\cdot)$ on $\mathcal{X}$ by

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

Then it is easily seen that $P^{k_2}(x, \cdot) \ge Q'(\cdot)$ for $x \in R$. Assuming $Q'(\mathcal{X}) > 0$ (otherwise the lemma is vacuously true), the result now follows by setting $Q(\cdot) = Q'(\cdot)/Q'(\mathcal{X})$, and setting $\epsilon_2 = Q'(\mathcal{X})$. ∎

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

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

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 i'th 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 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, we see that F is a geometrically distributed random variable with parameter ϵ. Setting $m_0 = E(h(X^{(0)}))$, we have that

$$E\left(\alpha^{T_1}\right) \;\leq\; m_0\, E\left(\alpha^{k_0} S_R\right)^F \;=\; m_0 \sum_{\ell=0}^{\infty} \epsilon(1-\epsilon)^{\ell} (\alpha^{k_0} S_R)^{\ell} \;=\; \frac{\epsilon\, m_0}{1-(1-\epsilon)\alpha^{k_0} S_R}\,,$$

as desired. The second statement follows similarly. ∎

Proof of Theorem 10. We begin by noting (using $E(X^2) = (EX)^2 + Var(X)$) that

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

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

We recall (see the Remark following Theorem 7) that, since at each iteration the old value $X_1^{(k)}$ is discarded, our subset R and function h should only refer 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

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

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

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

Hence, we can take $\alpha = 16/13$.

To continue, we note that

$$A \;=\; \sup_{(x,y)\in R\times R} E\left(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y\right) = (9/4) + (1/4)(7) = 4\,.$$

Furthermore, since the stationary distribution for Y_2 is $N(\mu, 1)$, we have that $E_\pi(Y_2 - \mu)^2 = 1$, so that $E_{\nu\times\pi}\left(h(X^{(0)}, Y^{(0)})\right) = 2 + E_\nu\left(x_2 - \mu\right)^2$. (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.)

We obtain a value for ϵ from Lemma 6 (ii). Indeed, we can take

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

(where $N(a,b;y) = \frac{1}{\sqrt{2\pi b}} e^{-(y-a)^2/2b}$ 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 ≥ 0.31.

We now apply Theorem 5 with $k_0 = 1$, $A = 4$, $\alpha = 16/13$, and $\epsilon = 0.31$. We choose $j = k/10$. Since $(0.69)^{1/10} < 0.964$, and $(16/13)^{-9/10} 4^{1/10} < 0.953$, 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

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

$$Var\left(\beta^{(k+1)} \mid S^{(k)}\right) \;=\; \frac{\gamma + 10\alpha_0}{(\delta + S^{(k)})^2}\,,$$

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

and $$Var\left(S^{(k+1)} \mid \beta^{(k)}\right) \;=\; \sum_i \frac{\alpha_0 + s_i}{(t_i + \beta^{(k)})^2}\,.$$

Note that although the random variables involved here are not themselves bounded, the conditional means and variances given above *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 paper.

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

A cursory numerical examination of the conditional means above (for the given data) suggests that the value of $S^{(k)}$ roughly approaches the value 6.5. Thus, writing our 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

$$E\left(h(X^{(1)}, Y^{(1)}) \mid X^{(0)}, Y^{(0)}\right)\,.$$

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 proceed accordingly. We define the function $e(w)$ by

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

where we have used the conditional mean and variance of the θ_i, and the conditional distribution of the β, as given above (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, we find 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

$$
\sup_{w \notin [4.0, 9.0]} \left(\frac{1 + e(w)}{1 + (w - 6.5)^2}\right) < 0.46\,,
$$

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

$$
\sup_w \left(\frac{0.46}{1 + (w - 6.5)^2/7.25} + \frac{e(w)}{7.25 + (w - 6.5)^2}\right) < 0.66\,,
$$

with the supremum obtained near $w = 6.6$ (though there is a competing upturn to 0.405 near $w = 0$). Hence,

$$
\begin{aligned}
&\sup_{(x,y) \notin R \times R} \left(\frac{E\left(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y\right)}{h(x, y)}\right) \\
&= \sup_{\substack{w_1, w_2 \\ w_1 \notin [4.0, 9.0]}} \left(\frac{1 + e(w_1) + e(w_2)}{1 + (w_1 - 6.5)^2 + (w_2 - 6.5)^2}\right) \\
&\le \sup_{w_2} \left[\left(\sup_{w_1 \notin [4.0, 9.0]} \left(\frac{1 + e(w_1)}{1 + (w_1 - 6.5)^2}\right) \Big/ \left(1 + (w_2 - 6.5)^2/7.25\right)\right) \right. \\
&\qquad\qquad \left. + \left(\frac{e(w_2)}{7.25 + (w_2 - 6.5)^2}\right)\right] \\
&< 0.66\,,
\end{aligned}
$$

where we have used the above numerical bounds, and have also used the fact that $1 + (w_1 - 6.5)^2 \geq 7.25$. Hence, we can choose $\alpha = 1/0.66 > 1.5$.

We compute a value for ϵ using Lemma 6 (ii) 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

$$
\begin{aligned}
\epsilon &= \int_0^\infty \left(\inf_{w \in [4.0, 9.0]} G(\gamma + 10\alpha_0,\ \delta + w; x) \right) dx \\
&= \int_0^\infty \min \left(G(\gamma + 10\alpha_0,\ \delta + 4.0; x),\ G(\gamma + 10\alpha_0,\ \delta + 9.0; x) \right) dx \\
&> 0.14\,,
\end{aligned}
$$

where again we have done the integration numerically.

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

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

$$
E_\pi \left((S'^{(0)} - 6.5)^2 \right) \ \leq\ \sup_x E \left((S - 6.5)^2 \mid \beta = x \right) \ =\ E \left((S^{(1)} - 6.5)^2 \mid \beta^{(1)} = 0 \right) \ <\ 43\,.
$$

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 $E(V(X^{(1)} \mid X^{(0)} = x) \leq \lambda V(x) + b$ with $\lambda = 0.46$ and $b = 3.3$. The Remark then gives $E_\pi (S - 6.5)^2 \leq b/(1 - \lambda) < 6.2$. It follows that

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

We now apply Theorem 5, with $k_0 = 1$, $\epsilon = 0.14$, $\alpha = 1.5$, $A = 5.6$, and $j = k/6$. Since $(0, 86)^{1/6} < 0.976$, and $(1.5)^{-5/6} (5.6)^{1.6} < 0.951$, the result follows. ■

Proof of Theorem 12. We set $h(x, y) = 1 + V(x) + V(y)$, and set $R = \{x \in \mathcal{X} \mid V(x) \leq d\}$. Then if $(x, y) \notin R \times R$, then $h(x, y) \geq 1 + d$. Thus, in terms of a coupled chain as in Lemma 4, we have

$$
E \left(h(X^{(1)}, Y^{(1)}) \mid X^{(0)} = x, Y^{(0)} = y \right) \ \leq\ 1 + \lambda V(x) + \lambda V(y) + 2b
$$

$$\leq \left(\lambda + \frac{1-\lambda+2b}{1+d}\right) h(x,y) \;=\; \left(\frac{1+2b+\lambda d}{1+d}\right) h(x,y)\,,$$

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

Furthermore, with A as in Theorem 5, we have

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

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

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

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

REFERENCES

Y. Amit (1991), On the rates of convergence of stochastic relaxation for Gaussian and Non-Gaussian distributions. J. Multivariate Analysis **38**, 89-99.

Y. Amit (1993), Convergence properties of the Gibbs sampler for perturbations of Gaussians. Tech. Rep. **352**, Department of Statistics, University of Chicago.

Y. Amit and U. Grenander (1991), Comparing sweep strategies for stochastic relaxation. J. Multivariate Analysis **37**, No. **2**, 197-222.

D. Applegate, R. Kannan, and N.G. Polson (1990), Random polynomial time algorithms for sampling from joint distributions. Tech. Rep. **500**, School of Computer Science, Carnegie-Mellon University.

S. Asmussen (1987), Applied Probability and Queues. John Wiley & Sons, New York.

K.B. Athreya, D. McDonald, and P. Ney (1978), Limit theorems for semi-Markov processes and renewal theory for Markov chains. Ann. Prob. **6**, 788-797.

K.B. Athreya and P. Ney (1978), A new approach to the limit theory of recurrent Markov chains. Trans. Amer. Math. Soc. **245**, 493-501.

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. Stat. Prob. Lett., to appear.

E.D. Belsley (1993), Rates of convergence of Markov chains related to association schemes. Ph.D. dissertation, Dept. of Mathematics, Harvard University.

A. Dembo and O. Zeitouni (1993), Large deviations techniques and applications. Jones and Bartlett Publishers, Boston, Mass.

P. Diaconis (1988), Group representations in Probability and Statistics. IMS, Hayward, Calif.

P. Diaconis and P. Hanlon (1992), Eigen analysis for some examples of the Metropolis algorithm. Tech. Rep., Dept. of Mathematics, Harvard University.

W. Feller (1971), An introduction to Probability Theory and its applications, Vol. *II*, $2^{\mathbf{nd}}$ ed. Wiley & Sons, New York.

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

A. Frigessi, C.-R. Hwang, L. Younes (1992), Optimal spectral structure of reversible stochastic matrices, Monte Carlo methods and the simulation of Markov random fields. Ann. Appl. Prob. **2**, 610-628.

D. Gaver and I. O'Muircheartaigh (1987), Robust empirical Bayes analysis of event rates. Technometrics **29**, 1-15.

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

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.

C. Geyer (1992), Practical Markov chain Monte Carlo. Stat. Sci., Vol. **7**, No. **4**, 473-483.

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

T. Lindvall (1992), Lectures on the Coupling Method. Wiley & Sons, New York.

J. Liu (1992), Eigen analysis for a Metropolis sampling scheme with comparisons to rejection sampling and importance resampling. Research Rep. **R-427**, Dept. of Statistics, Harvard University.

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.

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

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 (1993a), Markov chains and stochastic stability. Springer-Verlag, London.

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

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

E. Nummelin (1984), General irreducible Markov chains and non-negative operators. Cambridge University Press.

J.W. Pitman (1976), On coupling of Markov chains. Z. Wahrsch. verw. Gebiete **35**, 315-322.

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.

G.O. Roberts and N.G. Polson (1990), On the geometric convergence of the Gibbs sampler. J. Royal Stat. Soc. **B**, to appear.

J.S. Rosenthal (1991), Rates of convergence for Gibbs sampler for variance components models. Tech. Rep. **9322**, Department of Statistics, University of Toronto. Ann. Stat., to appear.

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

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 (1991), Markov chains for exploring posterior distributions. Tech. Rep. **560**, School of Statistics, University of Minnesota.