Geometric Convergence Rates for Time-Sampled Markov Chains

by

Jeffrey S. Rosenthal*

(April 5, 2002; last revised April 17, 2003)

Abstract. We consider time-sampled Markov chain kernels, of the form $P^\mu = \sum_n \mu\{n\}P^n$. We prove bounds on the total variation distance to stationarity of such chains. We are motivated by the analysis of near-periodic MCMC algorithms.

1. Introduction.

Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$ and stationary distribution $\pi(\cdot)$. Such schemes are often used to estimate $\pi(h) \equiv \int_{\mathcal{X}} h \, d\pi$ for various functionals $h : \mathcal{X} \to \mathbf{R}$, by e.g. $\widehat{\pi}(h) = \frac{1}{n}\sum_{i=1}^n h(X_i)$. Specific examples of such "MCMC algorithms" include the Gibbs sampler and the Metropolis-Hastings algorithm; for background see e.g. Smith and Roberts (1993), Tierney (1994), and Gilks, Richardson, and Spiegelhalter (1996).

Certain Markov chains (e.g. nearly-periodic chains) may have good asymptotic variance properties (i.e. the variance of $\widehat{\pi}(h)$ above is relatively small as $n \to \infty$, when started with $X_0 \sim \pi(\cdot)$), but still have poor distributional convergence properties (so that $\mathcal{L}(X_i)$ is far from $\pi(\cdot)$ unless i is very large). This is particularly relevant for *antithetic* chains, which introduce negative correlations to reduce asymptotic variance, but at the expense of possibly introducing near-periodic behaviour which may slow the distributional convergence (see e.g. Green and Han, 1992; Craiu and Meng, 2001).

It was argued in Rosenthal (2001) that, in such cases, it was worthwhile to instead consider a *time-sampled chain* of the form $P^\mu = \sum_n \mu\{n\}P^n$, where μ is a probability measure on the non-negative integers (e.g. perhaps $\mu\{0\} = \mu\{1\} = 1/2$). Then $(P^\mu)^m = P^{\mu^{*m}}$, where μ^{*m} is the m-fold convolution of μ with itself (cf. Meyn and Tweedie, 1993, Lemma 5.5.2(i)). Equivalently, $(P^\mu)^m$ is generated by choosing $T_m \sim \mu^{*m}$ independently

* Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 3G3. Internet: jeff@math.toronto.edu. Supported in part by NSERC of Canada.

of $\{X_n\}$, and considering X_{T_m}. For a very simple example, if $\mathcal{X} = \{1, 2\}$, and $P(1, 2) = P(2, 1) = 1$, then P is periodic, with stationary distribution $\pi = \text{Unif}(\mathcal{X})$. But if $\mu\{0\} = \mu\{1\} = 1/2$, then $P^\mu(i, j) = 1/2$ for all $i, j \in \mathcal{X}$, so that P^μ coverges to stationarity in one step, even though P is periodic and never converges. We consider P^μ herein.

2. Distributional convergence rates.

We now discuss distributional convergence. For a signed measure ν on $\mathcal{X}$, we write $\|\nu\|_{TV} = \sup_{A \subseteq \mathcal{X}} |\nu(A)|$ for total variation distance. We are interested in bounding distance to stationarity of the form $\|P^m(x, \cdot) - \pi(\cdot)\|_{TV}$, as a function of m.

Now, Markov chain convergence rates can sometimes be proved by establishing *minorisation conditions* of the form

$$P(x, A) \;\geq\; \epsilon\, \nu(A)\,, \qquad x \in C, \quad A \subseteq \mathcal{X}\,; \tag{1}$$

here $C \subseteq \mathcal{X}$ is called a *small set*, and $\epsilon > 0$, and $\nu(\cdot)$ is some probability measure on $\mathcal{X}$. (We shall abbreviate this as $P(x, \cdot) \geq \epsilon\, \nu(\cdot)$, $x \in C$.) Indeed, for *uniformly ergodic* Markov chains, where we can take $C = \mathcal{X}$, (1) is all that is required. Indeed, in that case $\|P^m(x, \cdot) - \pi(\cdot)\|_{TV} \leq (1 - \epsilon)^m$ for all $m \in \mathbf{N}$, as is well known. (This can be proved by coupling; see e.g. Doeblin, 1938; Doob, 1953; Griffeath, 1975; Pitman, 1976; Nummelin, 1984; Lindvall, 1992; Meyn and Tweedie, 1993; Rosenthal, 1995a, 1995b.)

Now, if P is (say) a nearly periodic chain, then it is unlikely we will have $P(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$ for any non-negligible ϵ. On the other hand, it is more likely that we will have $P^\mu(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$, where P^μ represents (as before) the same Markov chain but run for a random number of iterations. In that case, applying the above result to P^μ instead of P, we see (cf. Rosenthal, 2001) that if $P^\mu(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$, then

$$\|(P^\mu)^m(x, \cdot) - \pi(\cdot)\|_{TV} \;\equiv\; \|\mathcal{L}(X_{T_m} \,|\, X_0 = x) - \pi(\cdot)\|_{TV} \;\leq\; (1 - \epsilon)^m$$

for all $x \in \mathcal{X}$, where $T_m \sim \mu^{*m}$ is chosen independently of $\{X_n\}$.

Often, especially on infinite state spaces, the minorisation condition (1) will only be satisfied on a subset $C \subseteq \mathcal{X}$. In that case, suppose we also know that a *drift condition*

$$(P \times P)h(x, y) \;\leq\; h(x, y) \,/\, \alpha\,, \qquad (x, y) \notin C \times C \tag{2}$$

is satisfied, for some function $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$ and constant $\alpha > 1$, where

$$(P \times P)h(x, y) \;\equiv\; \int_{\mathcal{X}} \int_{\mathcal{X}} h(z, w)\, P(x, dz)\, P(y, dw)\,.$$

A slight modification of the argument and bound in Rosenthal (1995b), which follows as a special case of Douc et al. (2002), and which also takes into account the ϵ-improvement [i.e., replacing A by $A - \epsilon$ in (4)] of Roberts and Tweedie (1999), then yields the following.

Proposition 1. *Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$ and stationary distribution $\pi(\cdot)$. Suppose there is $C \subseteq \mathcal{X}$, $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$, a probability distribution $\nu(\cdot)$ on $\mathcal{X}$, $\alpha > 1$, and $\epsilon > 0$, such that (1) and (2) hold. Suppose also that*

$$\sup_{(x,y) \in C \times C} (P \times P)h(x, y) \;\leq\; A\,. \tag{3}$$

Then for any initial distribution $\mathcal{L}(X_0)$, and any integer $j \leq k$,

$$\|\mathcal{L}(X_k) - \pi(\cdot)\|_{TV} \;\leq\; (1 - \epsilon)^j + \alpha^{-k} \max[1,\, (\alpha(A - \epsilon))^{j-1}]\, E[h(X_0, Y_0)]\,, \tag{4}$$

with the expectation taken with respect to any joint law of X_0 and Y_0 provided their marginal distributions are $\mathcal{L}(X_0)$ and $\pi(\cdot)$ respectively.

Versions of Proposition 1 have been applied to a number MCMC examples in a variety of ways (see e.g. Meyn and Tweedie, 1994; Rosenthal, 1995a, 1995b, 1996; Roberts and Tweedie, 1999; Jones and Hobert, 2001, 2002; Cowles and Rosenthal, 1998; Cowles, 2001). The resulting bounds are useful, although they are usually conservative.

Remark. If $P(x, dy) \geq h(x, y)\, dy$, then we can achieve (1) by setting

$$\epsilon \;=\; \int_{\mathcal{X}} \inf_{x \in C} h(x, y)\, dy \tag{5}$$

and $\nu(dy) = \epsilon^{-1} \inf_{x \in C} h(x, y)\, dy$.

Remark. As observed in Roberts and Rosenthal (2000), small-set conditions of the form $P(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in C$, can be replaced by *pseudo-small* conditions of the form $P(x,\cdot) \geq \epsilon\nu_{xy}(\cdot)$ and $P(y,\cdot) \geq \epsilon\nu_{xy}(\cdot)$ for all $x, y \in C$, without affecting any bounds which use coupling (which includes all the bounds considered here).

Finally, we note that in verifying (2), it is often simpler to verify a univariate drift condition which bounds PV, where $V : \mathcal{X} \to \mathbf{R}$. One can then construct a bivariate function h from V. The following result summarises various possibilities, following Rosenthal (1995b,c), Cowles and Rosenthal (1998), and Roberts and Tweedie (1999). Parts (i) to (iv) follow by direct computation, since if $(x,y) \notin C \times C$ and $d_* = \inf_{x \notin C} V(x)$, then either $V(x) \geq d_*$ or $V(y) \geq d_*$ (or both). Part (v) is easily seen by taking expectations with respect to π of both sides of $PV \leq \lambda V + b$ (cf. Meyn and Tweedie, 1993, Proposition 4.3(i)).

Proposition 2. *Consider a Markov chain on a state space $\mathcal{X}$, with transition probabilities $P(x,\cdot)$. Let $V : \mathcal{X} \to \mathbf{R}$, let $C \subseteq \mathcal{X}$, let $\mathbf{1}_C$ be the indicator function of C, let $d_* = \inf_{x \notin C} V(x)$, let $d^* = \sup_{x \in C} V(x)$, and let $M > 0$. (Typically $M = 1$, $C = \{x \in \mathcal{X};\ V(x) \leq d\}$, and $d_* = d^* = d$.)*

(i) If $PV(x) \leq \lambda V(x) + b$ for all $x \in \mathcal{X}$, where $V \geq 0$, then (2) and (3) are satisfied with $h(x,y) = 1 + MV(x) + MV(y)$, $\alpha^{-1} = \lambda + \frac{1+2Mb-\lambda}{1+Md_*}$, and $A = 1 + 2M(\lambda d^* + b)$.

(ii) If $PV(x) \leq \lambda V(x) + b$ for all $x \in \mathcal{X}$, where $V \geq 1$, then (2) and (3) are satisfied with $h(x,y) = (M/2)(V(x) + V(y)) + (1 - M)$, $\alpha^{-1} = \lambda + \frac{Mb+(1-\lambda)(1-M)}{(M/2)(d_*+1)+(1-M)}$, and $A = M(\lambda d^* + b) + (1 - M)$.

(iii) If $PV(x) \leq \lambda V(x) + b\mathbf{1}_C(x)$ for all $x \in \mathcal{X}$, where $V \geq 0$, then (2) and (3) are satisfied with $h(x,y) = 1 + MV(x) + MV(y)$, $\alpha^{-1} = \lambda + \frac{1+Mb-\lambda}{1+Md_*}$, and $A = 1 + 2M(\lambda d^* + b)$.

(iv) If $PV(x) \leq \lambda V(x) + b\mathbf{1}_C(x)$ for all $x \in \mathcal{X}$, where $V \geq 1$, then (2) and (3) are satisfied with $h(x,y) = (M/2)(V(x) + V(y)) + (1 - M)$, $\alpha^{-1} = \lambda + \frac{(M/2)b+(1-\lambda)(1-M)}{(M/2)(d_*+1)+(1-M)}$, and $A = M(\lambda d^* + b) + (1 - M)$.

(v) Furthermore, under any of (i) to (iv), we have $E_\pi[V(Y_0)] \leq \frac{b}{1-\lambda}$, where the expectation is taken with respect to $Y_0 \sim \pi(\cdot)$. Hence, $E_\pi[h(x,Y_0)] \leq 1 + MV(x) + \frac{Mb}{1-\lambda}$ under (i) or (iii), and $E_\pi[h(x,Y_0)] \leq (M/2)(V(x) + \frac{b}{1-\lambda}) + (1 - M)$ under (ii) or (iv).

3. Application to time-sampled chains.

Suppose now that we only have $P^\mu(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in C$, where $C \subseteq \mathcal{X}$, for some probability distribution μ on the non-negative integers. (This means that C is *petite* for P in the language of Meyn and Tweedie, 1993; if P is aperiodic then this implies that C is also small for P, but without any control over the corresponding values of k_0 and ϵ.) Suppose also that (2) holds for P. That is, suppose we have a drift condition for P, but a minorisation condition for P^μ. How can we obtain convergence bounds in that case?

One method is to convert the drift condition for P to one for P^μ, as follows.

Proposition 3. *Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x,\cdot)$.*
(i) Suppose $PV(x) \leq \phi(V(x))$ for all $x \in \mathcal{X}$, where $\phi : [1,\infty) \to [1,\infty)$ is non-decreasing and (weakly) concave. Then

$$P^n V(x) \leq \phi(\phi(\ldots\phi(V(x))\ldots)) \equiv \phi_n(V(x))\,,$$

and

$$P^\mu V(x) \leq \sum_n \mu\{n\}\phi_n(V(x))\,.$$

(ii) In the special case $\phi(t) = \lambda t + b$ with $\lambda < 1$, then $P^\mu V \leq \lambda_\mu V + b_\mu$, where

$$\lambda_\mu = R_\mu(\lambda)\,; \qquad b_\mu = b\left(\frac{1-R_\mu(\lambda)}{1-\lambda}\right) < \frac{b}{1-\lambda}\,;$$

here $R_\mu(s) = E_\mu[s^Z] = \sum_n \mu\{n\}s^n$ is the probability generating function of μ.

Proof. Part (i) follows by induction and Jensen's inequality, since for $n = 1$ it is trivial, and assuming it true for n, then $P^{n+1}V = P\big(P^n V\big) \leq P\big(\phi_n \circ V\big) \leq \phi_n \circ (PV) \leq \phi_n \circ (\phi \circ V) = \phi_{n+1} \circ V$. For part (ii), we compute that if $\phi(t) = \lambda t + b$, then

$$\phi_n(t) = \lambda^n t + (\sum_{i=0}^{n-1} \lambda^i)b = \lambda^n t + b\left(\frac{1-\lambda^n}{1-\lambda}\right)\,.$$

Hence,

$$P^\mu V(x) \leq \sum_n \mu\{n\}\phi_n(V(x)) = \sum_n \mu\{n\}\left(\lambda^n V(x) + b\left(\frac{1-\lambda^n}{1-\lambda}\right)\right)$$

$$= R_\mu(\lambda)V(x) + b\left(\frac{1 - R_\mu(\lambda)}{1-\lambda}\right) . \qquad \blacksquare$$

That is, to replace P by P^μ, we must replace λ by $\lambda_\mu = R_\mu(\lambda)$, and must replace b by $b_\mu = b\left(\frac{1-R_\mu(\lambda)}{1-\lambda}\right)$. Some special cases are worth noting:

(a) If $\mu\{1\} = 1$, then $\lambda_\mu = \lambda$ and $b_\mu = b$, as they must.

(b) If $\mu\{k_0\} = 1$, then $\lambda_\mu = \lambda^{k_0}$ and $b_\mu = b\left(\frac{1-\lambda^{k_0}}{1-\lambda}\right)$.

(c) If $\mu\{0, 1, 2, \ldots, k_0 - 1\} = 0$, then $\lambda_\mu \le \lambda^{k_0}$.

Combining Proposition 3 part (ii) with Proposition 1 applied to P^μ, and with Proposition 2 parts (i) and (ii) and (v) (with $M = 1$, for simplicity), we obtain the following.

Theorem 4. *Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$ and stationary distribution $\pi(\cdot)$. Suppose $PV(x) \le \lambda V(x) + b$ where $\lambda < 1$ and $V : \mathcal{X} \to [0, \infty)$. Suppose also that $P^\mu(x, \cdot) \ge \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$ such that $V(x) \le d$. Then for any integer $j \le k$,*

$$\|\mathcal{L}(X_{T_k}) - \pi(\cdot)\|_{TV} \;\le\; (1-\epsilon)^j + \alpha_\mu^{-k} \max[1, (\alpha_\mu(A_\mu - \epsilon))^{j-1}]\,(1 + \frac{b}{1-\lambda} + E[V(X_0)])\,,$$

where $T_k \sim \mu^{*k}$ is chosen independently of $\{X_n\}$, and where

$$\alpha_\mu^{-1} \;=\; \lambda_\mu + \frac{1 - \lambda_\mu + 2b_\mu}{d+1} \;=\; R_\mu(\lambda) + \frac{1 - R_\mu(\lambda) + 2b\left(\frac{1-R_\mu(\lambda)}{1-\lambda}\right)}{d+1}\,,$$

and

$$A_\mu \;=\; \sup_{x,y \in C} (P^\mu \times P^\mu)(1 + V(y) + V(x)) \;\le\; 1 + 2(\lambda_\mu d + b_\mu)\,.$$

If $V \ge 1$, the value of α_μ^{-1} can be decreased slightly to $\alpha_\mu^{-1} = \lambda_\mu + \frac{2b_\mu}{d+1}$.

Another approach is to try to modify the proofs in Rosenthal (1995b) and Douc et al. (2002), to take into account jumping a random number $\sim \mu$ of iterations at each attempted regeneration, instead of just 1 iteration (or just k_0 iterations). The following theorem is proved in Section 5.

Theorem 5. *Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$ and stationary distribution $\pi(\cdot)$. Suppose $(P \times P)h(x, y) \leq h(x, y) / \alpha$ for $(x, y) \notin C \times C$, where $\alpha > 1$ and $h : \mathcal{X} \times \mathcal{X} \to [1, \infty)$. Suppose also that $P^\mu(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in C$. Then for any non-negative integers j and m,*

$$\|\mathcal{L}(X_{m+T_j}) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^j + \alpha^{-m-1} A_\mu^{j-1}\, E_\pi[h(X_0, Y_0)]\,,$$

where $T_j \sim \mu^{*j}$ is chosen independently of $\{X_n\}$, where the expectation is taken with respect to $\mathcal{L}(X_0)$ and $Y_0 \sim \pi(\cdot)$, and where $A_\mu = \sup_{x,y \in C}(P^\mu \times P^\mu)h(x, y)$.

We note that in Theorem 5, unlike Theorem 4, we can verify the drift condition $(P \times P)h(x, y) \leq h(x, y) / \alpha$ by any of the methods of Proposition 2.

If $\mu\{1\} = 1$, then $m + T_j = m + j$, so the bound of Theorem 5 becomes

$$\|\mathcal{L}(X_{m+j}) - \pi(\cdot)\|_{TV} \leq (1-\epsilon)^j + \alpha^{-m-1} A^{j-1} E[h(X_0, X_0'')]\,.$$

which is similar to (in fact a slight improvement of) Theorem 12 of Rosenthal (1995b). If $\mu\{k_0\} = 1$, where $k_0 \in \mathbf{N}$, then $k = m + T_j = m + k_0 j$, and $A_\mu = A_{k_0} \equiv \sup_{X \in C} P^{k_0} V(x)$, so the bound of Theorem 5 becomes

$$\|\mathcal{L}(X_{m+k_0 j}) - \pi(\cdot)\|_{TV} \leq (1-\epsilon)^j + \alpha^{-m-1} A_{k_0}^{j-1} E[h(X_0, X_0'')]\,,$$

which is similar to (in fact a slight improvement of) Theorem 5 of Rosenthal (1995b).

Remark. Another way of obtaining convergence bounds for near-periodic chains is through the use of shift-coupling (Aldous and Thorisson, 1993; Roberts and Rosenthal, 1997), whereby the minorisation condition holds for the original chain P throughout C, but the two chain copies do not need to regenerate at the same time. However, this approach bound only convergence of full ergodic averages $\frac{1}{n}\sum_{i=1}^n P^i(x, \cdot)$ of distributions. Furthermore it leads to bounds which decrease just linearly (rather than exponentially) in the number of iterations, so that more iterations will always be required to get within ϵ of stationarity for sufficiently small $\epsilon > 0$. In addition, in the examples below we typically have a minorisation condition for P^μ but not for P, making shift-coupling difficult to apply.

Hence, we conclude that such shift-coupling bounds are both weaker and more difficult to apply than the bounds herein.

4. Examples.

We now present some examples to which we apply the above theory.

Example 1. *A small set in many pieces.*

Suppose now that the state space $\mathcal{X}$ contains disjoint subsets $C_1, C_2, \ldots, C_D$ such that $P(x, \cdot) \geq \epsilon_0 \nu(\cdot)$ for all $x \in C_D$, and $P(x, C_{i+1}) \geq \delta_i$ for all $x \in C_i$ for $1 \leq i \leq D-1$, for some $D \geq 2$.

Let μ be uniform on $\{1, 2, \ldots, D\}$. Then we see by inspection that the union $\bigcup_i C_i$ is small for P^μ, with

$$P^\mu(x, \cdot) \;\geq\; \frac{1}{D}\delta_1 \ldots \delta_{D-1}\epsilon_0 \, \nu(\cdot) \;\equiv\; \epsilon \, \nu(\cdot)\,,$$

where $\epsilon = \frac{1}{D}\delta_1 \ldots \delta_{D-1}\epsilon_0$. (Such considerations generalise the notion of *transfer condition* discussed in Roberts and Rosenthal, 1997, Theorem 6.)

Suppose also that $PV \leq \lambda V + b\mathbf{1}_C$, where $V : \mathcal{X} \to [1, \infty)$, and where $C = \bigcup_i C_i = \{x \in \mathcal{X};\ V(x) \leq d\}$. Then the bounds of Theorem 4 and Theorem 5 can be applied.

We compute numerically with $D = 20$, $\epsilon = 0.3$, $\delta_i = 0.8$ for all i, $\lambda = 0.9$, $b = 10$, and $d = 200$. For Theorem 4, we compute using Proposition 3 that $\lambda_\mu = 0.395291$ and $b_\mu = 60.4709$. Then from Proposition 2, $\alpha_\mu^{-1} = \lambda_\mu + \frac{2b_\mu}{d+1} = 0.797091$, and $A_\mu = \lambda_\mu d + b_\mu = 179.058$. The bound of Theorem 4 then becomes

$$\|\mathcal{L}(X_{T_k}) - \pi(\cdot)\|_{TV} \;\leq\; (0.996541)^j + 101(3.0383)^{j-1}(20.5523)^k\,,$$

which is equal to $0.00782318 < 0.01$ if $j = 1,400$ and $k = 34,000$. Since μ has mean 10.5, this requires about $(10.5)k = 357,000$ iterations.

For Theorem 5, we see from Proposition 2 that $\alpha^{-1} = \lambda + \frac{b}{d+1} = 0.933223$, with A_μ as above. The bound of Theorem 4 then becomes

$$\|\mathcal{L}(X_{m+T_j}) - \pi(\cdot)\|_{TV} \;\leq\; (0.996541)^j + 101(179.058)^{j-1}(0.933223)^{1+m}\,,$$

which is equal to $0.00782318 < 0.01$ if $j = 1,400$ and $m = 106,000$. This requires about $m + (10.5)j = 120,700$ iterations.

We thus see that each of Theorem 4 and Theorem 5 provide rigorous bounds on convergence after a randomised number of iterations. Each of the bounds requires quite a large number of iterations to converge. However, the bound of Theorem 4 requires over $350,000$ iterations while the bound of Theorem 5 requires about $120,000$ iterations. Hence, for this example, the bound of Theorem 5 is nearly three times stronger than that of Theorem 4.

We note that in this example, the information available does not preclude the possibility that the original chain is, say, periodic, in which case it would not converge in distribution at all. Hence, without further information, minorisation and drift conditions for the original chain P (as opposed to P^μ) cannot be used to establish convergence rates.

Remark. In this example, we have only limited information available about the chain (the minorisation, transfer, and drift conditions). In fact, it is even possible for this chain to be *reversible*. Nevertheless, given the information we have available, the choice of μ being uniform on $\{1, 2, \ldots, D\}$ was preferred. (Similar comments apply to Example 3 below.) It is an open question whether, given *complete* information about a reversible chain, it would ever be preferred to have μ concentrated on more than two points.

Example 2. *An antithetic Metropolis algorithm.*

Let $\mathcal{X} = \mathbf{R}$, let $\gamma > 1$, let $a > 0$, and let $\pi(dx) \propto f(x)\,dx$ where the density f is defined by

$$f(x) \;=\; e^{-a|x-\mathrm{sign}(x)\gamma|}\,, \qquad x \in \mathbf{R}\,,$$

where $\mathrm{sign}(x) = 1$ for $x \geq 0$ and $\mathrm{sign}(x) = -1$ for $x < 0$. The density of f is thus a bimodal distribution with modes at $\pm\gamma$, which represents the continuous merging of two double-exponential densities.

We shall consider running a Metropolis algorithm for $\pi(\cdot)$. One possible proposal distribution is $\mathrm{Unif}[x-1, x+1]$; however this would take a very long time to move between the two modes. Instead, we shall use the proposal distribution given by $Q(x, \cdot) = \mathrm{Unif}[-x-1, -x+1]$, to do faster mode-hopping. That is, $Q(x, dy) = q(x, y)dy$ where $q(x, y) = \frac{1}{2}\, I(|y + x| \leq 1)$. (This proposal is "antithetic" in the sense that if X is random and $Y \sim Q(X, \cdot)$, then X and Y will be negatively correlated.)

Clearly, this Metropolis algorithm will not be uniformly ergodic. Indeed, we always have $\big||X_{n+1}| - |X_n|\big| \le 1$, while $\mathcal{X}$ is unbounded, so clearly $\{X_n\}$ cannot converge from *everywhere* in $\mathcal{X}$ in a fixed number of iterations. We thus consider applying Theorems 4 and 5.

We let $V(x) = f(x)^{-1/2} = e^{a|x - \text{sign}(x)\gamma|/2}$ (so $V \ge 1$), and let $C = \{x \in \mathcal{X};\ |x - \text{sign}(x)\gamma| \le 1\} = \{x \in \mathcal{X};\ V(x) \le e^{a/2}\}$ (so $d = e^{a/2}$). We then see (using symmetry) that for $x \notin C$,

$$PV(x) \,/\, V(x) \ = \ \frac{1}{2}\int_{-1}^{1} \min\left[1,\ e^{-az}\right]\, e^{az/2}\, dz + r \tag{6}$$

where $r = \frac{1}{2}\int_0^1 (1 - e^{-az})\, dz$ is the rejection probability from x.

Also for $x \in C$, the quantity $PV(x) - \lambda V(x)$ is maximised at $x = \pm\gamma$. Hence, $PV \le \lambda V + b$ if

$$b \ = \ PV(0) - \lambda V(0) \ = \ \int_0^1 e^{az/2} e^{-az} dz + \int_0^1 (1 - e^{-az}) dz - \lambda\,.$$

We next turn to the minorisation condition. Now, since C consists of two intervals, one near γ and one near $-\gamma$, and $\gamma > 1$, there is clearly no overlap at all in $\{P(x,\cdot)\}_{x\in C}$. Even $\{P^{k_0}(x,\cdot)\}_{x\in C}$ will have very little overlap unless k_0 is extremely large. Furthermore, if $\mu\{0\} = \mu\{1\} = \frac{1}{2}$, then $\{P^\mu(x,\cdot)\}_{x\in C} = \{\overline{P}(x,\cdot)\}_{x\in C}$ will again have no overlap at all.

On the other hand, if $\mu\{2\} = \mu\{3\} = \frac{1}{2}$, then $\{P^\mu(x,\cdot)\}_{x\in C}$ will have substantial overlap. Indeed, let $C^+ = \{x \in \mathcal{X};\ x > 0\}$ and $C^- = \{x \in \mathcal{X};\ x < 0\}$. Then for $x \in C^+$, we will always have $P(x,\, [-\gamma - \frac{1}{2},\, -\gamma + \frac{1}{2}]) \ge 1/4$. Hence, $P^2(x,\cdot)$ will always have density at least $1/8$ throughout the interval $[\gamma - \frac{1}{2},\, \gamma + \frac{1}{2}]$. Furthermore the acceptance probability at the point $-\gamma + z$ will be at least $e^{-a|z|}$. Hence, $P^2(x, dw) \ge \frac{1}{4}\kappa\, dw$ for $x \in C^+$ and $w \in [\gamma - \frac{1}{2},\, \gamma + \frac{1}{2}]$, where $\kappa \equiv \frac{1}{2}\int_{-1/2}^{1/2} e^{-a|z|}\, dz = \int_0^{1/2} e^{-az}\, dz$. Iterating this argument, we see that $P^3(x, dw) \ge \frac{1}{4}\kappa^2 dw$ for $x \in C^+$ and $w \in [\gamma - \frac{1}{2},\, \gamma + \frac{1}{2}]$. We conclude by symmetry that with $\mu\{2\} = \mu\{3\} = \frac{1}{2}$, $P^\mu(x, dw) \ge \frac{1}{4}\kappa^2 dw$ for $x \in C$ and either $|w - \gamma| \le \frac{1}{2}$ or $|w + \gamma| \le \frac{1}{2}$. Hence, by (5), we have $P^\mu(x,\cdot) \ge \epsilon\, \nu(\cdot)$ for all $x \in C$, with

$$\epsilon = 2(1/4)\kappa^2 = \kappa^2/2 = \frac{1}{2}\left(\frac{1}{2}\int_{-1/2}^{1/2} e^{-a|z|}\, dz\right)^2 .$$

We compute the bounds of Theorem 4 and Theorem 5 numerically with $a = 10$. The above arguments give $d = 148.413$, $\lambda = 0.648655$, $b = 0.450002$, $\kappa = 0.0993262$, and

$\epsilon = 0.00493285$. In the context of Theorem 4, we then have $\lambda_\mu = 0.346838$, $b_\mu = 0.836568$, $\alpha_\mu^{-1} = 0.358036$, and $A_\mu = 52.3119$. Setting $j = 1,000$ and $k = 5,000$, the bound of Theorem 4 gives

$$\|\mathcal{L}(X_{T_k}) - \pi(\cdot)\|_{TV} \;\leq\; 0.00711853\,,$$

where $E[T_k] = 2.5\,k = 12,500$. On the other hand, in the context of Theorem 5 we have $\alpha^{-1} = 0.654678$. Setting $j = 1,000$ and $m = 10,000$, the bound of Theorem 5 gives

$$\|\mathcal{L}(X_{m+T_j}) - \pi(\cdot)\|_{TV} \;\leq\; 0.00711853\,,$$

where $E[m + T_j] = m + 2.5\,j = 12,500$.

Hence, Theorem 4 and Theorem 5 give very similar convergence bounds for this chain. Each of them provides a result which is conservative, but not totally infeasible (i.e. it is quite possible to simulate X_{T_k} or X_{m+T_j} on a computer). Furthermore, each of the bounds requires doing a *random* number of iterations of the original chain, to reasonably bound the convergence.

We note that in this example, the antithetic Metropolis chain P itself must converge to $\pi(\cdot)$ by irreducibility and aperiodicity. Hence, in principle minorisation and drift conditions for P (as opposed to P^μ) could be used to establish convergence rates directly. However, if (say) a and γ are fairly large, then the convergence of P will be infeasibly slow, and any corresponding convergence rate bounds would necessarily be huge.

Example 3. *Random Walk on a Weighted Tree.*

Consider an infinite tree, defined as follows. There is a single root node $x_{1,1}$. This node has $C(x_{1,1}) = N_2 \geq 1$ children, labeled $x_{2,1}, \ldots, x_{2,N_2}$. Then $x_{2,i}$ has $C(x_{2,i}) \geq 0$ children ($1 \leq i \leq N_2$), for a total of $N_3 = \sum_{i=1}^{N_2} C(x_{2,i})$ nodes $x_{3,1}, \ldots, x_{3,N_3}$. We continue in the same manner. Thus, the j^{th} row of the tree has $N_j = \sum_{i=1}^{N_{j-1}} C(x_{j-1,i})$ nodes (where $N_1 = 1$). The set of all nodes is thus $\mathcal{X} = \{x_{i,j} :\; i \geq 1,\; 1 \leq j \leq N_i\}$. We assume that $N_j \geq 1$ for $j = 1, 2, 3, \ldots$, so $|\mathcal{X}| = \infty$. To make $\mathcal{X}$ into a weighted undirected tree, we add an edge between each node $x \in \mathcal{X} \setminus \{x_{1,1}\}$ and its parent, of weight $W(x) > 0$.

For $x \in \mathcal{X}$, we let $D(x) > 0$ equal the sum of the weights of all edges touching x, and let $\text{row}(x)$ be the row of x (so that $\text{row}(x_{i,j}) = i$).

We then define a random walk $X_0, X_1, \ldots$ on this weighted tree in the usual fashion. That is, for $x, y \in \mathcal{X}$, $P(X_{n+1} = y \,|\, X_n = x)$ is equal to the weight of the edge (if any) between x and y, divided by $D(x)$. This walk changes its row by one each time, and hence is clearly periodic with period two, so it does not converge in distribution in any fixed number of steps.

We regard the tree (given by the parameters $\{C(x)\}_{x\in\mathcal{X}}$ and $\{W(x)\}_{x\in\mathcal{X}}$) as fixed. We further assume that the weights $W(x_{i,j})$ go to 0 as $i \to \infty$, quickly enough that $W^* \equiv \sum_{x\in\mathcal{X}} W(x) < \infty$. In this case, the random walk has a stationary distribution, given by $\pi(x) = D(x) \,/\, 2W^*$; in fact the walk is reversible with respect to π.

To study the convergence of this random walk to its stationary distribution, we also make assumptions about the probability of the walk moving up. We assume that

$$\frac{W(x)}{D(x)} \;\geq\; \delta > 0\,, \qquad x \in \mathcal{X} \setminus \{x_{1,1}\}\,,$$

and for some $K \geq 2$,

$$\frac{W(x_{i,j})}{D(x_{i,j})} \;\geq\; p > 1/2\,, \qquad i > K,\ 1 \leq j \leq N_i\,. \tag{7}$$

The meaning of (7) is that, beyond row K of the tree, the random walk always has probability $\geq p$ of moving up (towards the root node) rather than down.

To proceed, let $\beta = 2 - (2p)^{-1}$, so $\beta > 1$. Define $V : \mathcal{X} \to [1, \infty)$ by $V(x) = \beta^{\mathrm{row}(x)-1}$. Let $C = \{x \in \mathcal{X};\ \mathrm{row}(x) \leq K + J\}$ for some $J \geq 0$. Then for $i > K$, since $(1+z)^{-1} \leq 1 - z + z^2$ for $z \geq 1$, we have

$$PV(x_{i,j}) \leq p\beta^{i-1} + (1-p)\beta^{i+1} = \beta^i\big[p\beta^{-1} + (1-p)\beta\big]$$

$$\leq \beta^i\big[p(2 - \beta + (\beta-1)^2) + (1-p)\beta\big] = \beta^i\big[2 - p - (4p)^{-1}\big] \equiv \lambda\, V(x_{i,j})\,,$$

where $\lambda = 2 - p - (4p)^{-1} < 1$.

We now let $h(x, y) = 1 + V(x) + V(y)$. Then for $\mathrm{row}(x), \mathrm{row}(y) \geq K + 1$,

$$(P \times P)h(x,y) \leq 1 + \lambda V(x) + \lambda V(y) \leq h(x,y)\Big[\lambda + \frac{1-\lambda}{1 + 2\beta^K}\Big] \equiv h(x,y)\lambda'\,.$$

For $\mathrm{row}(x) \geq K + J + 1$ and $\mathrm{row}(y) \leq K$, $PV(y) \leq (1-\delta)\beta V(y) + \delta\beta^{-1}V(y)$, so

$$\frac{(P \times P)h(x,y)}{h(x,y)} \leq \frac{1 + \lambda V(x) + [(1-\delta)\beta + \delta\beta^{-1}]V(y)}{1 + V(x) + V(y)}\,.$$

Now, if $(1-\delta)\beta + \delta\beta^{-1} \leq \frac{1+\lambda V(x)}{1+V(x)}$, then this is

$$\leq \frac{1+\lambda V(x)}{1+V(x)} \leq \frac{1+\lambda\beta^{K+J}}{1+\beta^{K+J}} \equiv \lambda'' < 1\,.$$

If $(1-\delta)\beta + \delta\beta^{-1} \geq \frac{1+\lambda V(x)}{1+V(x)}$, then this is

$$\leq \frac{1+\lambda V(x) + [(1-\delta)\beta + \delta\beta^{-1}]\beta^{-J-1}V(x)}{1+V(x)+\beta^{-J-1}V(x)}$$

$$\leq \frac{1+\lambda\beta^{K+J} + [(1-\delta)\beta + \delta\beta^{-1}]\beta^{K-1}}{1+\beta^{K+J}+\beta^{K-1}} \equiv \lambda'''\,.$$

We will often (but not always) have $\lambda''' < 1$. In this case, by the above, (2) is satisfied with $h(x,y) = 1 + V(x) + V(y)$, $C = \{x_{i,j} \in \mathcal{X};\ 1 \leq i \leq K\}$, and $\alpha^{-1} = \max(\lambda', \lambda'', \lambda''')$.

As for a minorisation condition, we let $\nu(\{x_{1,1}\}) = 1$. Then for $x_{i,j} \in C$, there is probability $\geq \delta^{i-1}$ that it will move up on each of its first $i-1$ jumps. Hence, if $\mu\{0\} = \mu\{1\} = \ldots = \mu\{K+J-1\} = 1/(K+J-1)$, then we have

$$P^{\mu}(x,\cdot) \geq \frac{1}{K+J-1}\,\delta^{K+J-1}\nu(\cdot) \equiv \epsilon\,\nu(\cdot)\,, \qquad x \in C\,.$$

For a crude bound on A_μ, we note that for $x \in C$, clearly $P^i V(x) \leq \beta^{K+J+i-1}$, so $P^\mu V(x) = (K+J-1)^{-1}\sum_{i=0}^{K-1} P^i V(x) \leq (K+J-1)^{-1}\sum_{i=0}^{K-1}\beta^{K+J+i-1} = (K+J-1)^{-1}\frac{\beta^{2K+J-1}-\beta^{K+J-1}}{\beta-1}$, so $A_\mu \leq 1 + 2(K+J-1)^{-1}\frac{\beta^{2K+J-1}-\beta^{K+J-1}}{\beta-1}$.

Finally, we note from the above that clearly $PV \leq \lambda V + b\mathbf{1}_C$ with λ as above and $b = \beta^{K+J}$. Hence, from Proposition 2 part (v), we have $\mathbf{E}_\pi[V(Y_0)] \leq \beta^{K+J}/(1-\lambda)$, so that $\mathbf{E}_\pi[h(x,Y_0)] \leq 1 + \beta^{\mathrm{row}(x)-1} + \beta^{K+J}/(1-\lambda)$.

We are now able to apply Theorem 5, to conclude that if $T_j \sim \mu^{*j}$ is chosen independently of $\{X_n\}$, then with α^{-1}, ϵ, and β as above,

$$\|\mathcal{L}(X_{m+T_j}) - \pi(\cdot)\|_{TV} \ \leq\ (1-\epsilon)^j +$$

$$\alpha^{-m-1}\Big(1 + 2(K+J-1)^{-1}\frac{\beta^{2K+J-1}-\beta^{K+J-1}}{\beta-1}\Big)^{j-1}\Big(1+\beta^{\mathrm{row}(X_0)-1}+\beta^{K+J}/(1-\lambda)\Big)\,.$$

As a specific numerical illustration, if $p = 4/5$ and $\delta = 1/2$, with $K = 3$ and $J = 4$, then if $\mathrm{row}(X_0) = 1$ and $j = 1,800$ and $m = 43,000$, the bound is equal to $0.00915354 <$

0.01. Since the mean of μ is equal to $(K + J - 1)/2 = 3$, this requires approximately $43,000 + 1,800(3) = 48,400$ iterations. On the other hand, if $p = 3/4$ and $\delta = 2/5$, with $K = 5$ and $J = 2$, then if $\text{row}(X_0) = 1$ and $j = 7,000$ and $m = 275,000$ the bound is equal to $0.00839349 < 0.01$. Here the mean of μ is again $(K + J - 1)/2 = 3$, so this requires approximately $275,000 + 7,000(3) = 296,000$ iterations. These bounds are obviously very conservative, and can doubtless be improved by a more problem-specific analysis. On the other hand, the bounds still provide numbers of iterations which can be done efficiently on a computer, so they may still be useful in providing guaranteed accuracy when doing simulations.

We note that in this example, the original chain P is periodic by construction. Hence, minorisation and drift conditions for the original chain P (as opposed to P^μ) cannot be used to establish convergence rates.

5. Proof of Theorem 5.

Let $\{\beta_1, \beta_2, \ldots, I_1, I_2, \ldots\}$ be a collection of independent random variables, where $\beta_i \sim \mu(\cdot)$, and $\mathbf{P}(I_i = 1) = 1 - \mathbf{P}(I_i = 0) = \epsilon$. Let $U_0 = 0$, and $U_k = \beta_1 + \ldots + \beta_k$ for $1 \leq k \leq j$. Then $U_j \sim \mu^{*j}$, so $\mathcal{L}(U_j) = \mathcal{L}(T_j)$ and $\mathcal{L}(X_{m+U_j}) = \mathcal{L}(X_{m+T_j})$.

We shall define *three* processes $\{X_t\}_{t=0}^{m+U_j}$, $\{X'_t\}_{t=0}^{m+j}$, and $\{X''_t\}_{t=0}^{m+j}$, each on $\mathcal{X}$. The idea is that X will start at X_0 and follow P, while X' and X'' will start at $X_0 \sim \mathcal{L}(X_0)$ and $X''_0 \sim \pi(\cdot)$ respectively, but will each follow a "collapsed time scale" where jumps of time β_i for X will correspond to jumps of time 1 for X' and X''. In particular, X is just a time change of X'.

We shall also define auxiliary variables $\{d_t, S_t, N_t\}_{t=0}^{m+j}$, where: d_t is the indicator function of whether or not X' and X'' have coupled by time t; S_t represents the time index for X which corresponds to the time index t for X' and X''; N_t represents the number of times X' and X'' have *attempted* to couple by time t.

Formally, we begin by setting $X'_0 = X_0$ where $\mathcal{L}(X_0)$ is the given initial distribution, and choosing $X''_0 \sim \pi(\cdot)$, with the pair (X_0, X''_0) following any joint law (e.g. independent). We also set $d_0 = S_0 = N_0 = 0$. Then iteratively for $n \geq 0$, given $X'_n, X''_n, d_n, S_n, N_n, X_{S_n}$:

1. If $d_n = 1$, then we must have $X'_n = X''_n = X_{S_n} \equiv x$, in which case

a. If $(X'_n, X''_n) \notin C \times C$ or $N_n = j$, then set $d_{n+1} = 1$, and $S_{n+1} = S_n + 1$, and $N_{n+1} = N_n$. Then choose $X'_{n+1} = X''_{n+1} = X_{S_{n+1}} \sim P(x, \cdot)$,

b. If $(X'_n, X''_n) \in C \times C$, and $N_n < j$, then set $d_{n+1} = 1$, and $N_{n+1} = N_n + 1$, and $S_{n+1} = S_n + \beta_{N_{n+1}}$. Then choose $X'_{n+1} = X''_{n+1} = X_{S_{n+1}} \sim P^{\beta_{N_{n+1}}}(x, \cdot)$, Then fill in $X_{S_n+1}, X_{S_n+2}, \ldots, X_{S_{n+1}-1}$ according to the transition kernel P, conditional on the values of S_n, S_{n+1}, X_{S_n}, and $X_{S_{n+1}}$.

2. If $d_n = 0$, then

a. If $(X'_n, X''_n) \notin C \times C$ or $N_n = j$, then set $d_{n+1} = 0$, and $S_{n+1} = S_n + 1$, and $N_{n+1} = N_n$. Then independently choose $X'_{n+1} = X_{S_{n+1}} \sim P(X'_n, \cdot)$, and $X''_{n+1} \sim P(X''_n, \cdot)$.

b. If $(X'_n, X''_n) \in C \times C$ and $N_n < j$ then set $d_{n+1} = I_{n+1}$, and $N_{n+1} = N_n + 1$. and $S_{n+1} = S_n + \beta_{N_{n+1}}$. Then

i. If $I_{n+1} = 1$, choose $X'_{n+1} = X''_{n+1} = X_{S_{n+1}} \sim \nu(\cdot)$,

ii. If $I_{n+1} = 0$, then independently choose

$$X_{S_{n+1}} = X'_{n+1} \sim (1 - \epsilon^{-1})(P^{\beta_{N_{n+1}}}(X'_n, \cdot) - \epsilon\nu(\cdot)),$$

and

$$X''_{n+1} \sim (1 - \epsilon^{-1})(P^{\beta_{N_{n+1}}}(X''_n, \cdot) - \epsilon\nu(\cdot)).$$

Under either i or ii, then fill in $X_{S_n+1}, X_{S_n+2}, \ldots, X_{S_{n+1}-1}$ according to the transition kernel P, conditional on the values of S_n, S_{n+1}, X_{S_n}, and $X_{S_{n+1}}$.

[To better understand the above construction, we note that steps (a) involve updating each of the three processes according to P, while steps (b) involve updating X according to P repeated S_n times, while updating (X', X'') according to P^μ (and attempting to couple them if they are not already coupled). Furthermore, step 2.b.i. involves the actual coupling, while step 2.b.ii. involves updating the processes from their "residual" kernels so that overall they are updated according to their correct transition kernels. Steps 1. involve simply *maintaining* the coupling (i.e. $X'_n = X''_n$) once it has already occurred.]

This construction is designed so that $\{X_t\}$ marginally follows its correct transition

kernel P (and, in particular, is marginally independent of the $\{\beta_i\}$). Also $0 \leq N_k \leq j$ for all k. Furthermore, $X_{S_k} = X'_k$ for all k, and $S_k = (k - N_k) + U_{N_k}$ for all k.

Lemma 4. *On the event* $\{N_{m+j} = j\}$, *we have* $X_{m+U_j} = X'_{m+j}$.

Proof. It follows from the above observations that if $N_{j+m} = j$, then $S_{m+j} = (m + j - j) + U_j = m + U_j$, so that $X_{m+U_j} = X_{S_{m+j}} = X'_{m+j}$. ∎

Now, since $X''_0 \sim \pi(\cdot)$, we have by stationarity that $X''_k \sim \pi(\cdot)$, for all k. Hence, using the coupling inequality (e.g. Lindvall, 1992; Rosenthal, 1995a,b), we see that

$$\|\mathcal{L}(X_{m+T_j}) - \pi(\cdot)\|_{TV} \;=\; \|\mathcal{L}(X_{m+U_j}) - \mathcal{L}(X''_{m+j})\|_{TV} \;\leq\; P(X_{m+U_j} \neq X''_{m+j})$$

$$=\; P[X_{m+U_j} \neq X''_{m+j},\; N_k \geq j] \;+\; P[X_{m+U_j} \neq X''_{m+j},\; N_k \leq j-1]\,. \tag{8}$$

By Lemma 4,

$$P[X_{m+U_j} \neq X''_{m+j},\; N_k \geq j] \;=\; P[X'_{m+j} \neq X''_{m+j},\; N_k \geq j]$$

$$\leq\; \mathbf{P}[I_1 = I_2 = \ldots = I_j = 0] \;=\; (1-\epsilon)^j\,, \tag{9}$$

which bounds the first term in (8).

Also,

$$P[X_{m+U_j} \neq X''_{m+j},\; N_{m+j} \leq j-1] \;\leq\; P[N_{m+j} \leq j-1]\,. \tag{10}$$

We bound this as in Rosenthal (1995b) by setting

$$M_k \;=\; \alpha^k (\alpha A_\mu)^{-N_k} h(X'_k, X''_k)\,.$$

Then using (2), M_k is easily seen (cf. Rosenthal, 1995b; Douc et al., 2001) to be a *supermartingale*, with $\mathbf{E}[M_{k+1} \,|\, X'_k, X''_k, M_k = m] \leq m$ (consider separately the cases $(X'_k, X''_k) \in C \times C$ and $(X'_k, X''_k) \notin C \times C$). Hence, since $\alpha A_\mu > 1$,

$$P[N_{m+j} \leq j-1] \;=\; P[(\alpha A_\mu)^{-N_{m+j}} \geq (\alpha A_\mu)^{-(j-1)}]$$

$$\leq (\alpha A_\mu)^{j-1}\mathbf{E}[(\alpha A_\mu)^{-N_{m+j}}] \quad \text{(by Markov's inequality)}$$

$$\leq (\alpha A_\mu)^{j-1}\mathbf{E}[(\alpha A_\mu)^{-N_{m+j}}\, h(X_{m+j}, X'_{m+j})] \quad \text{(since } h \geq 1\text{)}$$

$$= (\alpha A_\mu)^{j-1}\mathbf{E}[\alpha^{-(m+j)} M_{m+j}]$$

$$\leq (\alpha A_\mu)^{j-1}\alpha^{-m-j}\mathbf{E}[M_0] \quad \text{(since } \{M_k\} \text{ is supermartingale)}$$

$$= \alpha^{-m-j}(\alpha A_\mu)^{j-1}\mathbf{E}[h(X'_0, X''_0)]\,. \tag{11}$$

The result now follows by plugging (9) and (11) into (8). ■

Remark. If we could replace (10) by

$$P[X_{m+U_j} \neq X''_{m+j},\ N_{m+j} < j] \ \leq\ P[d_{m+j} = 0,\ N_{m+j} < j]\,, \tag{12}$$

then we could replace αA_μ by $\max[1,\ \alpha(A_\mu - \epsilon)]$ in the conclusion of the theorem, thus very slightly improving the result. Indeed, if $\beta_i \equiv 1$ then we can do precisely this (Douc et al., 2002), leading to Proposition 4 above. However, (12) is not true for general β_i. Indeed, in general if $N_{m+j} < j$ then we will not have $X_{m+U_j} = X'_{m+j}$. Hence, we might have $X'_{m+j} = X''_{m+j}$ and $d_{m+j} = 1$, even though $X_{m+U_j} \neq X''_{m+j}$. One can attempt to modify the construction of X' and X'' so that they sometimes jump according to P^μ even when they are not in $C \times C$, in an effort to force $X'_{m+j} = X_{m+U_j}$ no matter what; however, this then invalidates e.g. the drift condition (2), (One can even let X' and X'' jump according to P^μ when not in $C \times C$ *only* if they have already coupled; but this still does not take into account cases where e.g. they couple just before time $m + j$ even though N_{j+m} is far less than j.) Hence, we are unable to achieve the ϵ-improvement (12) when dealing with random β_i values.

Acknowledgements. I thank the organiser Petros Dellaportas and all the participants in the TMR Workshop on MCMC Model Choice, in Spetses, Greece, in August 2001, for inspiration related to this paper. I thank the two anonymous referees for very helpful comments.

REFERENCES

D.J. Aldous and H. Thorisson (1993), Shift-coupling. Stoch. Proc. Appl. **44**, 1-14.

M.K. Cowles (2001). MCMC Sampler Convergence Rates for Hierarchical Normal Linear Models: A Simulation Approach. Statistics and Computing, to appear.

M.K. Cowles and J.S. Rosenthal (1998), A simulation approach to convergence rates for Markov chain Monte Carlo algorithms. Statistics and Computing **8**, 115–124.

R.V. Craiu and X.-L. Meng (2001). Antithetic Coupling for Perfect Sampling. In Proceedings of the 2000 ISBA conference.

W. Doeblin (1938), Exposé de la theorie des chaînes simples constantes de Markov à un nombre fini d'états. Rev. Math. Union Interbalkanique **2**, 77–105.

R. Douc, E. Moulines, and J.S. Rosenthal (2002), Quantitative convergence rates for inhomogeneous Markov chains. Submitted for publication.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds. (1996), *Markov chain Monte Carlo in practice*. Chapman and Hall, London.

P.J. Green and X.-L. Han (1992), Metropolis methods, Gaussian proposals, and antithetic variables. In Stochastic Models, Statistical Methods and Algorithms in Image Analysis (P. Barone et al., Eds.). Springer, Berlin.

D. Griffeath (1975), A maximal coupling for Markov chains. Z. Wahrsch. verw. Gebiete **31**, 95–106.

G.L. Jones and J.P. Hobert (2001), Honest exploration of intractable probability distributions via Markov chain Monte Carlo. Statistical Science **16**, 312–334.

G.L. Jones and J.P. Hobert (2002), Sufficient Burn-in for Gibbs Samplers for a Hierarchical Random Effects Model. Ann. Stat., to appear.

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

S.P. Meyn and R.L. Tweedie (1993), Markov chains and stochastic stability. Springer-Verlag, London.

S.P. Meyn and R.L. Tweedie (1994), Computable bounds for convergence rates of Markov chains. Ann. Appl. Prob. **4**, 981–1011.

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 and J.S. Rosenthal (1997), Shift-coupling and convergence rates of ergodic averages. Communications in Statistics – Stochastic Models, Vol. **13**, No. **1**, 147–165.

G.O. Roberts and J.S. Rosenthal (2000), Small and Pseudo-Small Sets for Markov Chains. Communications in Statistics – Stochastic Models, to appear.

G.O. Roberts and R.L. Tweedie (1999), Bounds on regeneration times and convergence rates for Markov chains. Stoch. Proc. Appl. **80**, 211–229.

J.S. Rosenthal (1995a), Rates of Convergence for Gibbs Sampling for Variance Components Models. Ann. Stat. **23**, 740–761.

J.S. Rosenthal (1995b), Minorization conditions and convergence rates for Markov chain Monte Carlo. J. Amer. Stat. Assoc. **90**, 558–566.

J.S. Rosenthal (1995c), One-page supplement to Rosenthal (1995b). Available from `http://probability.ca/jeff/research.html`

J.S. Rosenthal (1996), Analysis of the Gibbs sampler for a model related to James-Stein estimators. Stat. and Comput. **6**, 269–275.

J.S. Rosenthal (2001), Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms. J. Amer. Stat. Assoc., to appear.

A.F.M. Smith and G.O. Roberts (1993), Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (with discussion). J. Roy. Stat. Soc. Ser. B **55**, 3–24.

L. Tierney (1994), Markov chains for exploring posterior distributions (with discussion). Ann. Stat. **22**, 1701–1762.