Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms

by

Jeffrey S. Rosenthal*

(October 4, 2001; last revised April 5, 2002)

Abstract. We consider nearly-periodic Markov chains, which may have excellent functional-estimation properties but poor distributional convergence rate. We show how simple modifications of the chain (involving using a random number of iterations) can greatly improve the distributional convergence of the chain. We prove various theoretical results about convergence rates of the modified chains. We also consider a number of examples.

1. Introduction.

Consider a Markov chain Monte Carlo (MCMC) sampling algorithm $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with updating 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.

$$\widehat{\pi}(h) = \frac{1}{n} \sum_{i=1}^{n} h(X_i) \, . \tag{1}$$

Specific examples of 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).

There are two different notions of such a sampling algorithm being a "good" one:

1. *Distributional convergence.* The MCMC algorithm is "good" if the chain converges quickly in distribution, i.e. i does not have to be too large to make $\mathcal{L}(X_i)$

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

be close to $\pi(\cdot)$. (This implies in turn that the mean of $\widehat{\pi}(h)$ above is close to $\pi(h)$.)

2. *Asymptotic variance.* Alternatively, the algorithm is "good" if the variance of $\widehat{\pi}(h)$ above is relatively small as $n \to \infty$, when started in stationarity (i.e., with $X_0 \sim \pi(\cdot)$).

These two goals have been described as "conflicting", and it has even been proposed to begin with a rapidly-converging chain and then later *switch* to a small-variance chain (e.g. Besag and Green, 1993; Mira, 2001). Indeed, it is true that if the underlying Markov chain is (say) periodic or nearly periodic, then the convergence of $\mathcal{L}(X_i)$ to $\pi(\cdot)$ could be slow, even though $\widehat{\pi}(h)$ is a good approximation to $\pi(h)$. 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).

On the other hand, in the present paper we argue that the above two goals are not as conflicting as they might appear. In particular, we show that given a reversible sampler with good asymptotic variance properties, a very slight modification of the sampler (the *binomial modification*) will also have good distributional convergence properties. We then generalise this idea to consider *sampled chains* of the form $P^{\mu} = \sum_n \mu\{n\} P^n$ for probability distributions μ on the non-negative integers. We prove various results about the spectra and quantitative convergence rates of such chains.

2. A very simple example.

To motivate what follows, consider the simplest example of a periodic chain. Specifically, let $\mathcal{X} = \{1, 2\}$, with transition matrix P given by

$$P \;=\; \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

That is, this Markov chain always moves from 1 to 2 and from 2 to 1. The stationary distribution $\pi(\cdot)$ of this chain is given by the uniform distribution on $\mathcal{X}$.

This chain has excellent asymptotic variance properties. Indeed, if $h : \mathcal{X} \to \mathbf{R}$, and if $X_0 \sim \pi(\cdot)$, then we always have $\widehat{\pi}(h) = \pi(h)$ exactly (so the variance is zero).

On the other hand, the chain has very poor distributional convergence properties. Indeed, for any $x \in \mathcal{X}$ and any $n \in \mathbf{N}$, the distribution $P^n(x, \cdot)$ is always concentrated on just one point, so it never converges to $\pi(\cdot)$ (it is periodic).

Now, let $\overline{P}$ be the Markov chain which either does nothing (with probability 1/2), or does the same as P (with probability 1/2). Then $\overline{P} = \frac{1}{2}(I + P)$ where I is the identity matrix. (Similar such "mixtures" are considered e.g. in Proposition 3 of Tierney, 1994.) Hence, the matrix of $\overline{P}$ is given by

$$\overline{P} \;=\; \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} .$$

We thus see that the chain $\overline{P}$ converges to $\pi(\cdot)$ immediately, and therefore has excellent distributional convergence properties. Similarly, if we let $\widehat{P}^n$ equal either P^n or P^{n+1} with probability 1/2 each, then $\widehat{P}^n$ also converges immediately to π.

Furthermore, running $\widehat{P}^n$ is very similar to running P^n. Also, running $\overline{P}$ for $2n$ steps is equivalent (in terms of the distribution of the final value obtained) to running P for a random number of steps having distribution Binomial($2n$, 1/2) (hence, we call $\overline{P}$ the *binomial modification* of P).

We thus see that minor modifications to the original, periodic (but good for estimation) Markov chain results in new Markov chains which have excellent distributional convergence properties. This theme is explored further herein.

In addition, Markov chain convergence rates can sometimes be proved by establishing minorisation conditions such as

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

For the chain P given above, this is clearly impossible due to the periodicity problem. On the other hand, for the modified chain $\overline{P}$ this is easy; in fact

$$\overline{P}(x, A) \;\geq\; \pi(A)\,, \qquad x \in \mathcal{X}, \quad A \subseteq \mathcal{X}\,,$$

so we may take $\epsilon = 1$ in that case. Issues of proving convergence rates of the modified chain are explored in later sections of this paper.

Finally, we note that the general idea of considering a random number of iterations is not new. For example, if $T_n \sim \text{Unif}\{1, 2, \ldots, n\}$ (as opposed to $B_n \sim \text{Binomial}(2n, 1/2)$), then the distance of $\mathcal{L}(X_{T_n})$ to stationarity can be bounded using *shift-coupling* (Aldous and Thorisson, 1993; Roberts and Rosenthal, 1997a; Roberts and Tweedie, 1999). However, the resulting shift-coupling bounds are $O(1/n)$ rather than decreasing exponentially with n, and are thus weaker than the bounds considered here.

3. The Spectrum of P.

In this section we consider reversible Markov chain kernels P, and review two spectral quantities, $interval(P)$ and $gap(P)$, which are closely related to the asymptotic variance and convergence rates of P, respectively.

Let $\pi(\cdot)$ be stationary for a reversible Markov transition kernel P. Suppose the chain is in stationarity, i.e. that $\mathcal{L}(X_n) = \pi(\cdot)$ for every $n \in \mathbf{Z}$. Then it is known (e.g. Geyer, 1992) that

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{Var}_\pi \left(\sum_{i=1}^{n} g(X_i) \right) \;=\; \sum_{t=-\infty}^{\infty} \mathbf{Cov}\left(g(X_0), g(X_t)\right) \;=\; \mathbf{Var}_\pi(g) + 2\sum_{t=1}^{\infty} \mathbf{Cov}\left(g(X_0), g(X_t)\right) .$$

This asymptotic variance is also related to the spectrum of the operator P, as follows.

Define the inner product $\langle f,\, g\rangle = \int_{\mathcal{X}} f(x)g(x)\pi(dx)$ for $f, g \in L^2(\pi)$, where

$$L^2(\pi) \;=\; \{f : \mathcal{X} \to \mathbf{R}\,;\; \pi(f^2) < \infty\}\,.$$

Assume P is reversible, so that P defines a self-adjoint operator on $L^2(\pi)$. Let $P_0 = P\big|_{L_0^2(\pi)}$ be the restriction of P to $L_0^2(\pi)$, where

$$L_0^2(\pi) \;=\; \{f : \mathcal{X} \to \mathbf{R}\,;\; \pi(f^2) < \infty,\; \pi(f) = 0\}\,.$$

(This restriction is made to exclude the non-zero constant functions, which are eigenvectors corresponding to the eigenvalue 1 of stationarity.) Let $\sigma(P_0)$ be the spectrum of P_0 (see e.g. Conway, 1985; roughly the spectrum corresponds to the set of eigenvalues of the matrix P_0, but generalised to continuous state spaces). Assume P is ϕ-irreducible, so that $\sigma(P_0) \subseteq [-1, 1)$ (cf. Mira and Geyer, 1999). We shall see that the distance of the spectrum

to the value 1 (in two senses, one with absolute values and one without) is closely related to convergence and variance properties of the corresponding MCMC algorithm.

Let $E_{P_0}(\cdot)$ be the resolution of the identity associated with P_0, as in the spectral theorem (see e.g. Conway, 1985; Reed and Simon, 1972; Geyer, 1992; Chan and Geyer, 1994; Mira and Geyer, 1999), so that

$$g(P_0) = \int_{\sigma(P_0)} g(\lambda)\, E_{P_0}(d\lambda)\,,$$

for every bounded Borel-measurable function $g : \sigma(P_0) \to \mathbf{R}$. Given a bounded Borel-measurable function g, let E_{g,P_0} be the spectral measure associated with g and P_0, so that $E_{g,P_0}(A) = \langle g,\, E_{P_0}(A)\, g\rangle$ and

$$\langle g,\, h(P_0)g\rangle = \int_{\sigma(P_0)} h(\lambda) E_{g,P_0}(d\lambda)\,, \tag{2}$$

for every bounded Borel-measurable function $h : \mathbf{R} \to \mathbf{R}$. In particular, setting $h(P_0) \equiv 1$ in (2), we see that

$$\langle g,\, g\rangle = \pi(g^2) = \int_{\sigma(P_0)} E_{g,P_0}(d\lambda)\,. \tag{3}$$

Then the following is known (Kipnis and Varadhan, 1986; see also Geyer, 1992; Chan and Geyer, 1994; Mira and Geyer, 1999).

Proposition 1. *Let P be the kernel for a reversible, ϕ-irreducible Markov chain $\{X_n\}$, and let $E_{g,P_0}(\cdot)$ be as above. Then*

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{Var}\left(\sum_{i=1}^{n} g(X_i)\right) \;=\; \int_{\sigma(P_0)} \frac{1+\lambda}{1-\lambda}\, E_{g,P_0}(d\lambda)\,.$$

From Proposition 1, we easily see the following.

Corollary 2. *Let P be the kernel for a reversible, ϕ-irreducible Markov chain $\{X_n\}$, and let $\Lambda = \Lambda(P_0) = \sup_{\lambda\in\sigma(P_0)} \lambda$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{Var}\left(\sum_{i=1}^{n} g(X_i)\right) \;\le\; \frac{1+\Lambda}{1-\Lambda}\, \pi(g^2) \;<\; \frac{2}{1-\Lambda}\, \pi(g^2)\,.$$

Proof. Since $\lambda \to \frac{1+\lambda}{1-\lambda}$ is an increasing function for $\lambda \in \sigma(P_0) \subseteq [-1, 1)$, we have from Proposition 1 that

$$\lim_{n\to\infty} \frac{1}{n} \mathbf{Var}\left(\sum_{i=1}^{n} g(X_i)\right) \;=\; \int_{\sigma(P_0)} \frac{1+\lambda}{1-\lambda}\, E_{g,P_0}(d\lambda)$$

$$\leq \int_{\sigma(P_0)} \frac{1+\Lambda}{1-\Lambda}\, E_{g,P_0}(d\lambda) \;=\; \frac{1+\Lambda}{1-\Lambda} \int_{\sigma(P_0)} E_{g,P_0}(d\lambda) \;=\; \frac{1+\Lambda}{1-\Lambda}\, \pi(g^2)$$

by (3). Also $\Lambda < 1$, so $1 + \Lambda < 2$. ∎

We conclude that the quantity

$$interval(P) \equiv 1 - \Lambda(P_0) \equiv 1 - \sup_{\lambda\in\sigma(P_0)} \lambda$$

is very closely related to the asymptotic variance of empirical estimators of functionals as in (1).

We next turn to distributional convergence. The following is essentially standard spectral theory, though we include a proof for completeness. For a signed measure ν on $\mathcal{X}$, we write $\|\nu\|_{TV} = \sup_{A\subseteq\mathcal{X}} |\nu(A)|$ for total variation distance, and write $\|\nu\|_{L^2(\pi)} = \int_{\mathcal{X}} (\frac{d\nu}{d\pi})^2\, d\pi$ (with $\|\nu\|_{L^2(\pi)} = \infty$ if ν is not absolutely continuous with respect to π) for $L^2(\pi)$ distance.

Proposition 3. *Let P be the kernel for a reversible Markov chain. Let $r(P_0) = \sup_{\lambda\in\sigma(P_0)} |\lambda|$ be the spectral radius of P_0. Then*

$$\sup_{\|\mu\|_{L^2(\pi)}<\infty} \lim_{n\to\infty} \frac{1}{n} \log \|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} \;=\; \log r(P_0)\,,$$

where the supremum is taken over all probability distributions μ on $\mathcal{X}$ having finite $L^2(\pi)$-norm.

Proof. It follows from Roberts and Rosenthal (1997b) (cf. Roberts and Tweedie, 2000, Theorem 3) that

$$\sup_{\|\mu\|_{L^2(\pi)}<\infty} \lim_{n\to\infty} \frac{1}{n} \log \|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} = \sup_{\|\mu\|_{L^2(\pi)}<\infty} \lim_{n\to\infty} \frac{1}{n} \log \|\mu P^n(\cdot) - \pi(\cdot)\|_{L^2(\pi)}\,,$$

i.e. that we can replace TV distance by $L^2(\pi)$ distance in the statement of the Proposition.

We have

$$\|\mu P^n(\cdot) - \pi(\cdot)\|_{L^2(\pi)} \leq \|\mu(\cdot) - \pi(\cdot)\|_{L^2(\pi)} \|P_0^n\|_{L^2(\pi)}$$

$$\leq \|\mu(\cdot) - \pi(\cdot)\|_{L^2(\pi)} r(P_0)^n \,.$$

Hence, taking logs, dividing by n, and letting $n \to \infty$, we see that

$$\sup_{\mu \in L^2(\pi)} \lim_{n \to \infty} \frac{1}{n} \log \|\mu P^n(\cdot) - \pi(\cdot)\|_{L^2(\pi)} \;\leq\; \log r(P_0) \,.$$

Conversely, by the spectral radius formula (e.g. Conway, 1985), we have

$$r(P_0)^n \;=\; \|P_0^n\|^n \;=\; \sup\{ \left(\frac{\|P^n f\|_{L_2(\pi)}}{\|f\|_{L_2(\pi)}} \right)^{1/n} \;;\; f \in L_0^2(\pi) \}$$

$$\leq \; \sup\{ \left(\frac{\|P^n (g-1)\|_{L_2(\pi)}}{\|g-1\|_{L_2(\pi)}} \right)^{1/n} \;;\; g \in L^2(\pi),\; g \geq 0,\; \pi(g) = 1 \}$$

$$= \; \sup\{ \left(\frac{\|P^n (\frac{d(\mu-\pi)}{d\pi})\|_{L_2(\pi)}}{\|\frac{d(\mu-\pi)}{d\pi}\|_{L_2(\pi)}} \right)^{1/n} \;;\; \mu \text{ prob dist},\; \|\mu\|_{\mathcal{L}^2(\pi)} < \infty \}$$

$$= \; \sup\{ \left(\frac{\|(\mu-\pi) P^n\|_{L_2(\pi)}}{\|\mu-\pi\|_{L_2(\pi)}} \right)^{1/n} \;;\; \mu \text{ prob dist},\; \|\mu\|_{\mathcal{L}^2(\pi)} < \infty \} \,.$$

Hence, taking logs, dividing by n, and letting $n \to \infty$, we see that

$$\log r(P_0) \;\leq\; \sup_{\mu \in L^2(\pi)} \lim_{n \to \infty} \frac{1}{n} \log \|\mu P^n(\cdot) - \pi(\cdot)\|_{L^2(\pi)} \,.$$

The result follows. ∎

Proposition 3 says that for large n, we roughly have

$$\|\mu P^n(\cdot) - \pi(\cdot)\|_{TV} \;\approx\; C\, r(P_0)^n \;=\; C\,(1 - (1 - r(P_0)))^n$$

$$\approx\; C\,(e^{-(1-r(P_0))})^n \;=\; C\,(e^{-n(1-r(P_0))}) \,,$$

at least if $r(P_0) \approx 1$ as it usually would be. Hence, the quantity

$$gap(P) \equiv 1 - r(P_0) \equiv 1 - \sup_{\lambda \in \sigma(P_0)} |\lambda|$$

is a good measure of the distributional convergence rate of P. (Similar considerations are also discussed elsewhere, see e.g. Schervish and Carlin, 1992.)

4. Modifications for near-periodic chains.

The previous section showed that $interval(P)$ is a good measure of a chain's asymptotic variance properties, while $gap(P)$ is a good measure of a chain's distributional convergence properties.

Now, clearly $interval(P) \geq gap(P)$. Also, these two quantities will often be similar or identical. However, they could be very different if e.g. all $\lambda \in \sigma(P)$ are far from 1, but one of them is close to -1, so $interval(P)$ is large but $gap(P)$ is small. On the other hand, we now argue that simple modifications of the Markov chain itself allow us to deal with this situation quite easily.

Let

$$B_n \sim \text{Binomial}(2n,\ 1/2)\,,$$

with $\{B_n\}$ chosen independently of the Markov chain $\{X_n\}$. Then $B_n/n \to 1$ as $n \to \infty$, so $B_n \approx n$ for large n. Also $\sum_m \mathbf{P}(B_n = m)P^m = (\overline{P})^n$, where $\overline{P} = \frac{1}{2}I + \frac{1}{2}P$. That is, $(\overline{P})^n$ corresponds to running the original Markov chain P for B_n steps instead of n. Hence, $\overline{P}$ is just a slight modification of P.

The following result shows that in the reversible case at least, if P has good asymptotic variance properties, then $\overline{P}$ also has good convergence rate properties (and hence could be used to generate a random variable having distribution very close to stationary). To state it, let $\zeta(\epsilon) = \epsilon - \frac{1}{4}\epsilon^2$, so that $\zeta(\epsilon) \leq \epsilon$, and $\zeta(\epsilon) \approx \epsilon$ for small ϵ.

Theorem 4. *If P is reversible, then $gap(\overline{P}) = \zeta(interval(P))$.*

Proof. We have that

$$\overline{P} = \left(\frac{I+P}{2}\right)^2 .$$

Now, let $\eta(\lambda) = (\frac{1}{2}(1+\lambda))^2$. Then since P is self-adjoint, we have (see e.g. Conway, 1985) that

$$\sigma(\overline{P}_0) = \sigma\left(\left(\frac{I_0+P_0}{2}\right)^2\right) = \{(\frac{1}{2}(1+\lambda))^2\,;\ \lambda \in \sigma(P_0)\} = \{\eta(\lambda)\,;\ \lambda \in \sigma(P_0)\}\,.$$

Note that for $\lambda \in \sigma(P_0) \subseteq \mathbf{R}$, we have $\eta(\lambda) \geq 0$. Also, $\eta(\lambda)$ is an increasing function of λ for $\lambda \in \sigma(P_0) \subseteq [-1, 1]$. Hence,

$$r(\overline{P}_0) = \sup_{\lambda \in \sigma(\overline{P}_0)} |\lambda| = \sup_{\lambda \in \sigma(\overline{P}_0)} |\eta(\lambda)| = \sup_{\lambda \in \sigma(P_0)} \eta(\lambda) = \eta\left(\sup_{\lambda \in \sigma(P_0)} \lambda\right).$$

The statement now follows since $1 - \eta(x) = \zeta(1 - x)$. ■

It follows from Theorem 4 that the convergence rate properties of $\overline{P}$ are at least as good (and essentially the same) as the asymptotic variance properties of P. That is, the simple modification of using $\overline{P}$ instead of P gives us distributional convergence which is as fast as would be indicated by the asymptotic variance properties. (In particular, if $\mathit{interval}(P) \approx 0$, then $\mathit{gap}(\overline{P}) \approx \mathit{interval}(P)$. On the other hand, if $\mathit{interval}(P) \approx 2$ as for an extremely antithetic chain, then $\mathit{gap}(\overline{P}) \approx 1$, indicating extremely fast convergence.)

We shall refer to $\overline{P}$ as the *binomial modification* of P. More generally, we shall later consider $P^\mu \equiv \sum_n \mu\{n\} P^n$ for various probability measures μ on the non-negative integers; we then have $\overline{P} = P^\mu$ for the special case $\mu\{0\} = \mu\{1\} = 1/2$. On the other hand, if (say) P were nearly periodic with period 3, then one might instead choose $\mu\{0\} = \mu\{1\} = \mu\{2\} = 1/3$.

Next define $\widehat{P}^n$ by $\widehat{P}^n = \frac{1}{2}(P^n + P^{n+1})$. That is, $\widehat{P}^n$ corresponds to running P for L_n iterations, where $P(L_n = n) = P(L_n = n + 1) = 1/2$, with $\{L_n\}$ chosen independently of the Markov chain itself. Set $\theta_n(\lambda) = \frac{1}{2}\lambda^n + \frac{1}{2}\lambda^{n+1} = \lambda^n(1 - \frac{1}{2}(1 - \lambda))$. Then we have the following.

Theorem 5. *If P is reversible, then*

$$r(\widehat{P}_0^n) \;=\; \sup_{\lambda \in \sigma(P_0)} \theta_n(\lambda)\,.$$

(In particular, if $\sup \sigma(P_0) \approx 1$, then $r(\widehat{P}_0^n) \approx \sup \sigma(P_0^n)$, while if $\sup \sigma(P_0) \approx -1$, then $r(\widehat{P}_0^n) \approx 0$.)

Proof. We have using self-adjointedness of P that

$$\sigma(\widehat{P}_0^n) \;=\; \{\frac{1}{2}\lambda^n + \frac{1}{2}\lambda^{n+1}\,;\; \lambda \in \sigma(P_0)\} \;=\; \{\theta_n(\lambda)\,;\; \lambda \in \sigma(P_0)\}\,.$$

The result follows by taking supremums. ∎

More generally, we could consider $P^\mu P^n$ in place of $\widehat{P}^n$, for various probability measures μ on the non-negative integers. We then have $\widehat{P}^n = P^\mu P^n$ for the special case $\mu\{0\} = \mu\{1\} = 1/2$. In fact, running $P^\mu P^n$ on an initial distribution ρ is precisely equivalent to running P^n on the initial distribution ρP^μ. That is, modifications such as $\widehat{P}$ (as opposed to $\overline{P}$) correspond merely to choosing a more intelligent initial distribution.

Since intelligent initial distributions generally provide only slight improvement in convergence properties, in this paper we mostly concentrate on generalisations of $\overline{P}^n$ (i.e., $(P^\mu)^n$ for various μ) as opposed to generalisations of $\widehat{P}^n$ (i.e., $P^\mu P^n$ for various μ).

5. Uniform convergence rates.

We now turn our attention to methods of proving convergence rates for Markov chains with kernels of the form P^μ as above. We first recall a well-known fact about Markov chains and minorisation conditions, which 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).

Proposition 6. *Let P be the transitions for a Markov chain on a state space $\mathcal{X}$, having stationary distribution $\pi(\cdot)$. Suppose P satisfies the minorisation condition $P(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$, where $\epsilon > 0$ and where $\nu(\cdot)$ is any probability measure on $\mathcal{X}$. Then*

$$\|P^m(x,\cdot) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^m\,.$$

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.

To proceed, let μ be any probability measure on the non-negative integers. Let $P^\mu = \sum_n P^n \mu\{n\}$ (where P^0 is the identity operator, i.e. $P^0(x,\cdot) = \delta_x(\cdot)$). (In the language of Meyn and Tweedie (1993), P^μ is a *sampled chain*.) 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 of $\{X_n\}$, and considering X_{T_m}.

In terms of P^μ, we have the following.

Theorem 7. *Suppose $P^\mu(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$, where $\epsilon > 0$ and where $\nu(\cdot)$ is any probability measure on $\mathcal{X}$. Then for all $x \in \mathcal{X}$,*

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

where $T_m \sim \mu^{*m}$ is chosen independently of $\{X_n\}$.

Proof. Simply apply Proposition 6 to P^μ. ∎

Theorem 7 thus says that, if we have found a distribution μ such that P^μ satisfies a minorisation condition, and we run our original Markov chain for an appropriate random number of steps, then the resulting value will be very close to stationary. This provides a simple mechanism for obtaining a sample from a given distribution $\pi(\cdot)$, even if the corresponding MCMC algorithm is periodic (or nearly so). Some examples applying Theorem 7 are presented in Section 7.

Of course, in some MCMC applications, one wishes to average the results obtained from a single long run of the chain. In this case, the asymptotic variance is a more relevant quantity. What the above results say is that, if the original chain has good asymptotic variance properties (and hence is good for averaging), then the modified chain has in addition good convergence properties (and hence is good for obtaining a sample).

On the other hand, Theorem 7 can only be applied if P^μ is *uniformly ergodic*. The next section considers modifications for non-uniform chains.

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). That is, rather than having a single minorising measure $\nu(\cdot)$ for all $x \in C$, it suffices to have a different minorising measure $\nu_{xy}(\cdot)$ for each pair $x, y \in C$. For ease of exposition we do not emphasise this fact here. However, it should be noted that all bounds presented here such as Theorems 7, 11, and 12 all go through without change if the minorising measure $\nu(\cdot)$ is allowed to vary depending on the pair $x, y \in C$.

6. Non-uniform convergence rates.

Suppose we know only that

$$P(x,\cdot) \geq \epsilon\nu(\cdot)\,, \qquad x \in C\,, \tag{4}$$

where $C \subseteq \mathcal{X}$ (as opposed to $C = \mathcal{X}$ as in Proposition 6). Suppose we also know that a *drift condition*

$$(P \times P)h(x,y) \;\leq\; h(x,y)\,/\,\alpha\,, \qquad (x,y) \notin C \times C \tag{5}$$

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

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

Under such conditions, non-uniform convergence rates are available. In particular, a slight modification of the argument and bound in Rosenthal (1995b), which follows as a special case of Douc et al. (2001), and which also takes into account the ϵ-improvement [i.e., replacing A by $A - \epsilon$ in (7)] of Roberts and Tweedie (1999), is the following.

Proposition 8. *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 (4) and (5) hold. Suppose also that*

$$\sup_{(x,y)\in C\times C} (P \times P)h(x,y) \;\leq\; A\,. \tag{6}$$

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

$$\|\mathcal{L}(X_k) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^j + \alpha^{-k}\max[1,\,(\alpha(A-\epsilon))^{j-1}]\,E[h(X_0,Y_0)]\,, \tag{7}$$

with the expectation taken with respect to $\mathcal{L}(X_0)$ and with respect to $Y_0 \sim \pi(\cdot)$.

Versions of Proposition 8 have been applied to a number of simplified examples in Meyn and Tweedie (1994), Rosenthal (1995a,b), and Roberts and Tweedie (1999). They have also been applied to more substantial examples of the Gibbs sampler, including a hierarchical Poisson model (Rosenthal, 1995b), a version of the variance components model (Rosenthal, 1996), and a number of newer examples (Jones and Hobert, 2001). Furthermore, with the aid of auxiliary simulation to approximately verify the drift and minorisation conditions, approximate versions of Proposition 8 have been applied to more complicated Gibbs sampler examples (Cowles and Rosenthal, 1998; Cowles, 2001).

Note that if $P(x, dy) \geq h(x, y)\, dy$, then we can achieve (4) by setting

$$\epsilon \;=\; \int_{\mathcal{X}} \inf_{x \in C} h(x, y)\, dy \tag{8}$$

and $\nu(dy) = \epsilon^{-1} \inf_{x \in C} h(x, y)\, dy$. Note also that the quantity $\mathbf{E}[h(X_0, Y_0)]$ in Proposition 8 may be computed 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, though typically one will take X_0 and Y_0 to be independent.

In verifying (5), 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, and conclude a drift condition of the form (5) for h. 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, simply noting that if $(x, y) \notin C \times C$, 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 9. *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 (5) and (6) 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 (5) and (6) 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 (5) and (6) 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 (5) and (6) 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).*

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 (5) 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 10. *(i) Suppose* $PV(x) \leq \phi(V(x))$ *for all* $x \in \mathcal{X}$, *where* $\phi : [1, \infty) \to [1, \infty)$ *is non-decreasing. Then*

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

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* $\lambda \leq 1$, *then* $P^\mu V \leq \lambda_\mu V + b_\mu$, *where*

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

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

Proof. (i) follows immediately by iterating the inequalities. For (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,

$$\begin{aligned} P^\mu V(x) &\le \sum_n \mu\{n\}\phi_n(V(x)) = \sum_n \mu\{n\}\left(\lambda^n V(x) + b\left(\frac{1-\lambda^n}{1-\lambda}\right)\right) \\ &= M_\mu(\lambda)V(x) + b\left(\frac{1-M_\mu(\lambda)}{1-\lambda}\right) , \end{aligned}$$

as claimed. ∎

That is, to replace P by P^μ, we must replace λ by $\lambda_\mu = M_\mu(\lambda)$, and must replace b by $b_\mu = b\left(\frac{1-M_\mu(\lambda)}{1-\lambda}\right)$. 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 10 with Proposition 8 applied to P^μ, and with Proposition 9 parts (i) and (ii) and (v) (with $M = 1$, for simplicity), we obtain the following.

Theorem 11. *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

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

and

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

If $V \geq 1$, 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. (2001), 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 the Appendix.

Theorem 12. *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 $x \in \mathcal{X}$ such that $V(x) \leq d$. 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 12, unlike Theorem 11, we can verify the drift condition $(P \times P)h(x,y) \leq h(x,y) \,/\, \alpha$ by any of the methods of Proposition 9.

If $\beta_i = 1$ for all i, then $m + T_j = m + j$, so the bound of Theorem 12 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 $\beta_i = k_0$ for all i, 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 12 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).

7. Examples.

We now present a number of examples, to which we apply the theory of the previous sections.

Example 1. *A periodic continuous chain.*

Let $\mathcal{X} = [0, 2]$, and define P as follows. For $x \in [0, 1]$, $P(x, \cdot) = \text{Unif}[1, 2]$, while for $x \in (1, 2]$, $P(x, \cdot) = \text{Unif}[0, 1]$. This chain is reversible with respect to $\pi(\cdot) = \text{Unif}[0, 2]$.

This example is simple enough that we can understand its spectrum exactly. Indeed, note that $Ph = h$ if h is constant; $Ph = -h$ if $h(x) = C$ for $x > 1$ and $h(x) = -C$ for $x \leq 1$ for some constant C; and $Ph = 0$ if $\int_0^1 h = \int_1^2 h = 0$. This shows that P has a one-dimensional eigenspace corresponding to the eigenvalue 1, a one-dimensional eigenspace corresponding to the eigenvalue -1, and an infinite-dimensional eigenspace corresponding to the eigenvalue 0. Furthermore, since every measurable function can be written as a linear combination from these three eigenspaces, we see that this completely specifies the spectrum of P. Thus, $\sigma(P) = \{-1, 0, 1\}$ and $\sigma(P_0) = \{-1, 0\}$.

Hence, $\mathit{interval}(P) = 1$ while $\mathit{gap}(P) = 0$. In words, we see that this example (like that of Section 2) has excellent asymptotic variance properties, but very poor distributional convergence properties.

On the other hand, by Theorem 4, we see that $\mathit{gap}(\overline{P}) = 1$, i.e. the binomial-modified chain $\overline{P}^m$ converges to $\pi(\cdot)$ extremely quickly, as does $\widehat{P}^m$. (On the other hand, unlike the simple example of Section 2, this chain will not converge *exactly* after one iteration, since for any m, $\overline{P}^m$ always includes probability 2^{-2m} of not moving at all.)

Furthermore, from the perspective of Theorem 7, we see that we cannot have $P^{k_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ with $\epsilon > 0$, for all $x \in \mathcal{X}$ for any k_0 and $\nu(\cdot)$. On the other hand, with $\mu\{1\} = \mu\{2\} = 1/2$, we have $P^\mu(x, \cdot) = \pi(\cdot)$ for all $x \in \mathcal{X}$, so we can take $\epsilon = 1$ in the uniform minorisation context of Theorem 7, to get that $\|(P^\mu)^m(x, \cdot) - \pi(\cdot)\|_{TV} = 0$ for any $m \geq 1$ and all $x \in \mathcal{X}$.

Example 2. *A nearly-periodic chain.*

Again let $\mathcal{X} = [0, 2]$, and suppose now that we only know there is some $\delta_1, \delta_2 > 0$ such that for $x \in [0, 1]$, $P(x, \cdot) \geq \delta_1 \text{Unif}[1, 2]$, while for $x \in (1, 2]$, $P(x, \cdot) \geq \delta_2 \text{Unif}[0, 1]$. (This means that e.g. for $x \in [0, 1]$ and $1 \leq a < b \leq 2$, $P(x, [a, b]) \geq \delta_1(b - a)$.) Suppose the

chain has some stationary (though perhaps non-uniform) distribution $\pi(\cdot)$. (The previous example corresponds to $\delta_1 = \delta_2 = 1$ and $\pi(\cdot) = \text{Unif}[0, 2]$.) Since we know less about this chain, it is more difficult to directly understand its spectral properties.

On the other hand, we can still use Theorem 7. Indeed, we have $P(x, \cdot) \geq \delta_2 \text{Unif}[0, 1]$ for $x \in (1, 2]$, and $P^2(x, \cdot) \geq \delta_1\delta_2 \text{Unif}[0, 1]$ for $x \in [0, 1]$. Similarly $P(x, \cdot) \geq \delta_1 \text{Unif}[1, 2]$ for $x \in [0, 1]$, and $P^2(x, \cdot) \geq \delta_1\delta_2 \text{Unif}[1, 2]$ for $x \in (1, 2]$. Hence, with $\mu\{1\} = \mu\{2\} = 1/2$, we have $P^\mu(x, \cdot) \geq \epsilon\, \nu(\cdot)$ for all $x \in \mathcal{X}$, where $\epsilon = \min[\delta_1, \delta_2, \delta_1\delta_2] = \delta_1\delta_2$, and $\nu(\cdot) = \text{Unif}[0, 2]$.

Hence, by Theorem 7, $\|(P^\mu)^m(x, \cdot) - \pi(\cdot)\|_{TV} \leq (1 - \delta_1\delta_2)^m$. This provides a bound on how many iterations of P^μ should be done (or equivalently, how many *random* iterations of P should be done), to get sufficiently close to (say, within 0.01 of) the stationary distribution $\pi(\cdot)$.

Example 3. *A chain of period $D \geq 3$.*

Suppose now that $\mathcal{X} = \{1, 2, \ldots, D\}$, where $D \geq 3$. Suppose further that $P(i, \{i + 1\}) = 1$ for $1 \leq i \leq D - 1$, and $P(D, \{1\}) = 1$. This chain has stationarity distribution $\pi(\cdot) = \text{Unif}(\mathcal{X})$. However, the chain is periodic of degree D. Hence, it does not converge in distribution at all.

We note that the modification $\widehat{P}$ from Section 4 does not help. Indeed, $\widehat{P}^n(i, \{j\}) = 0$ unless $j \equiv i + n \pmod{D}$ or $j \equiv i + n + 1 \pmod{D}$. Indeed, the distribution $\widehat{P}^n(i, \cdot)$ always satisfies $\|\widehat{P}^n(i, \cdot) - \pi(\cdot)\| = (D - 2)/D$, and does not go to zero as $n \to \infty$.

The modification $\overline{P}$ from Section 4 does indeed help. In that case, the distribution $\overline{P}^n(i, \cdot)$ is equal to the distribution of $Y_n = B_n + i \pmod{D}$ where $B_n \sim \text{Binomial}(2n, 1/2)$. Hence, $\|\overline{P}^n(i, \cdot) - \pi(\cdot)\| = \|\mathcal{L}(Y_n) - \pi(\cdot)\|$, which goes to zero, gradually, as $n \to \infty$.

Even better is to consider P^μ, where μ is uniform on $\{0, 1, 2, \ldots, D - 1\}$. In that case $P^\mu(i, \cdot) = \pi(\cdot)$ for any i, so $\|(\mathbf{P}^\mu)^m(i, \cdot) - \pi(\cdot)\| = 0$ for any $i \in \mathcal{X}$ and any $m \geq 1$. That is, P^μ converges to stationarity in just one step.

Example 4. *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 $1 \leq i \leq D - 1$.

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, 1997a, Theorem 6.)

Suppose also that $PV \leq \lambda V + b\mathbf{1}_C$, where $V : \mathcal{X} \to [1, \infty)$, and where $\bigcup_i C_i = \{x \in \mathcal{X};\ V(x) \leq d\}$. Then the bounds of Theorem 11 and Theorem 12 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 11, we compute using Proposition 10 that $\lambda_\mu = 0.395291$ and $b_\mu = 60.4709$. Then from Proposition 9, $\alpha_\mu^{-1} = \lambda_\mu + \frac{2b_\mu}{d+1} = 0.797091$, and $A_\mu = \lambda_\mu d + b_\mu = 179.058$. The bound of Theorem 11 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 if $j = 1,400$ and $k = 34,000$. Since μ has mean 10.5, this proves convergence (with a randomised number of iterations) after about $(10.5)k = 357,000$ iterations.

For Theorem 12, we see from Proposition 9 that $\alpha^{-1} = \lambda + \frac{b}{d+1} = 0.933223$, with A_μ as above. The bound of Theorem 11 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 if $j = 1,400$ and $m = 106,000$. This proves convergence (again with a randomised number of iterations) after about $m + (10.5)j = 120,700$ iterations.

We thus see that each of Theorem 11 and Theorem 12 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 11 requires over $350,000$ iterations while the bound of Theorem 12 requires about $120,000$ iterations. Hence, for this example, the bound of Theorem 12 is nearly three times stronger than that of Theorem 11.

Example 5. *A dimension-jumping Metropolis-Hastings algorithm.*

Consider the chain of Proposition 3.1 of Brooks, Guidichi, and Roberts (2001). This is a very simple example of a dimension-jumping Metropolis-Hastings algorithm, in the spirit of e.g. Norman and Filinov (1969), Preston (1977), and Green (1995).

The Markov chain is defined as follows. Let $\mathcal{X} = \{e\} \cup [0,1]$, and $\pi(\{e\}) = p$, and $\pi(dy) = (1-p)f(y)$ for $y \in [0,1]$, where $0 < p < 1$ and $\int_0^1 f(y)dy = 1$. We run a Metropolis-Hastings algorithm for $\pi(\cdot)$, with proposal kernel $\{Q(x,\cdot)\}_{x\in\mathcal{X}}$ defined by $Q(y,\{e\}) = 1$ for $y \in [0,1]$, and $Q(e,dy) = q(y)dy$ for $y \in [0,1]$, where $\int_0^1 q(y)dy = 1$.

It seems reasonable to try to get a minorisation condition with $\nu(\{e\}) = 1$, i.e. to show that $P(x,\{e\}) \geq \epsilon$ for all $x \in \mathcal{X}$, or perhaps that $P^\mu(x,\{e\}) \geq \epsilon$ for all $x \in \mathcal{X}$.

We compute that

$$S \equiv P(e,\{e\}) = 1 - P(e,\,[0,1]) = 1 - \int_0^1 \min\left[1,\ \frac{(1-p)f(y)}{p}\,\frac{1}{q(y)}\right] q(y)dy\,.$$

If $q \equiv f$, then $S = \max[0,\ \frac{2p-1}{p}]$. Also,

$$I \equiv \inf_{0\leq y\leq 1} P(y,\{e\}) = \inf_{0\leq y\leq 1} \min\left[1,\ \frac{p}{(1-p)f(y)}\,\frac{q(y)}{1}\right]$$

$$= \min\left[1,\ \frac{p}{(1-p)}\inf_{0\leq y\leq 1}\frac{q(y)}{f(y)}\right]\,.$$

If $q \equiv f$, then $I = \min[1,\ \frac{p}{1-p}]$.

We therefore see that $P(x,\{e\}) \geq \epsilon$ for all $x \in \mathcal{X}$, where $\epsilon = \min[S,\, I]$. However, if e.g. $q \equiv f$ (as suggested by Brooks et al., 2001) and $p \leq 1/2$, then $S = 0$ and so $\epsilon = 0$. In fact, if $q \equiv f$ and $p = 1/2$, then the chain is periodic, always accepting its moves and therefore always jumping back and forth between $\{e\}$ and $[0,1]$. Hence, in this case we will never have $P^{k_0}(x,\{e\}) \geq \epsilon$ for all $x \in \mathcal{X}$, for any $\epsilon > 0$.

On the other hand, obviously $P^0(e,\{e\}) = 1$ (by definition, in fact). Hence, if $\mu\{0\} = \mu\{1\} = 1/2$, then $P^\mu(x,\{e\}) \geq I$ for all $x \in \mathcal{X}$. Hence, by Theorem 7, we have

$$\|(P^\mu)^m(x,\cdot) - \pi(\cdot)\| \ \leq\ (1-I)^m\,, \qquad x \in \mathcal{X}\,,$$

so that $\|\mathcal{L}(X_{B_n}) - \pi(\cdot)\| \leq (1-I)^m$ regardless of the initial distribution $\mathcal{L}(X_0)$ (where $B_n \sim$ Binomial$(2n,\,1/2)$ is independent of $\{X_n\}$). Note that if $q \equiv f$, then $1 - I = \max[0,\ \frac{p}{1-p}]$,

so in that case we obtain

$$\|(P^\mu)^m(x,\cdot) - \pi(\cdot)\| \;\le\; \max[0,\; (\frac{p}{1-p})^m]\,, \qquad x \in \mathcal{X}\,.$$

Example 6. *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 $x \ge 0$ 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 antithetic 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 $q(x,y) = \frac{1}{2}\,I(|y+x| \le 1)$.

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. It is thus necessary to turn to the results of Section 6.

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

$$PV(x)\,/\,V(x) \;=\; \frac{1}{2}\int_{-1}^{1} \min\left[1,\; e^{-az}\right]\, e^{az/2}\,dz + r \tag{9}$$

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 $\gamma \geq 1$, 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 $P(x,\, [-\gamma - \frac{1}{2},\, -\gamma + \frac{1}{2}]) \geq 1/4$. 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, $P^2(x, dw) \geq \frac{1}{4}\kappa\, dw$ 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 $P^3(x, dw) \geq \frac{1}{4}\kappa^2 dw$ for $x \in C^+$ and $w \in [\gamma - \frac{1}{2},\, \gamma + \frac{1}{2}]$. We conclude by symmetry that with $\mu\{2\} = \mu\{3\} = \frac{1}{2}$, $P^\mu(x, dw) \geq \frac{1}{4}\kappa^2 dw$ for $x \in C$ and either $|w - \gamma| \leq \frac{1}{2}$ or $|w + \gamma| \leq \frac{1}{2}$. Hence, by (8), we have $P^\mu(x,\cdot) \geq \epsilon\, \nu(\cdot)$ 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 11 and Theorem 12 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 11, 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 11 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 12 we have $\alpha^{-1} = 0.654678$. Setting $j = 1,000$ and $m = 10,000$, the bound of Theorem 12 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 11 and Theorem 12 give very similar convergence bounds for this chain. Each of them provides a result which is overly conservative, but not totally unreasonable (i.e. it is quite feasible to simulate X_{T_k} or X_{m+T_j} here). Furthermore, each of the

bounds requires doing a *random* number of iterations of the original chain, to reasonably bound the convergence.

Example 7. *A multi-dimensional antithetic Metropolis simulation.*

Let $\mathcal{X} = \mathbf{R}^{50}$ be fifty-dimensional space. Let $\pi(d\mathbf{x}) = f(\mathbf{x})\, d\mathbf{x}$, where

$$f(\mathbf{x}) \;\propto\; e^{-\sum_{j=1}^{50} (x_j - \gamma \,\mathrm{sign}(\sum_i x_i)\mathbf{1})^2}\,, \qquad x \in \mathbf{R}^{50}\,,$$

where $\gamma > 0$ and $\mathbf{1} = (1, 1, \ldots, 1)$. The distribution $\pi(\cdot)$ is thus a "merging" of two normal distributions, with modes at $\pm\gamma\mathbf{1}$.

Consider running a Metropolis algorithm $\mathbf{X}_0, \mathbf{X}_1, \ldots$ for $\pi(\cdot)$, with one of two different proposal distributions: $Q_1(\mathbf{x}, \cdot) = N(\mathbf{x},\, \sigma^2\, I)$, and $Q_2(\mathbf{x}, \cdot) = N(-\mathbf{x},\, \sigma^2\, I)$. That is, the proposals are normally distributed, with variance σ^2 times the identity matrix, and with mean either $\mathbf{x}$ or $-\mathbf{x}$. Hence, Q_1 is a non-antithetic proposal, while Q_2 is an antithetic proposal.

We simulated this chain numerically with $\gamma = 10$ and $\sigma = 0.01$, starting at the mode $\gamma\mathbf{1}$. With proposal Q_1, the chain is essentially unable to reach the other mode, and indeed even after a million iterations there is not a single time n with $\sum_i X_{n,i} < 0$ (where $\mathbf{X}_n = (X_{n,1}, \ldots, X_{n,d})$). Hence, with proposal Q_1, the chain converges very, very slowly.

With proposal Q_2, the chain is antithetic, and jumps between the two modes very easily. In this case, the autocorrelations of x_1 (say) are essentially zero. (The autocorrelations of $(x_1)^2$ are not zero but are still very small, since they are equivalent to the autocorrelations within a single mode which are very small.)

On the other hand, even with proposal Q_2, the chain converges quite slowly in distribution. This is because there are so few rejections (since σ is so small, and f is symmetric) that the chain exhibits near-periodic behaviour. This is corrected by the use of the schemes $\overline{P}$ and $\widehat{P}$ from Section 4, each of which effectively causes convergence.

We simulated this model in dimension 50, with $\gamma = 10$ and $\sigma = 0.01$, for each of the proposals Q_1 and Q_2, and for each of the sampling schemes P, $\overline{P}$, and $\widehat{P}$. For each of the six combinations, we ran 100,000 separate runs, each for 20 iterations started at the mode $\gamma\,\mathbf{1}$, and computed the mean of the resulting distribution of $X_{20,1}$ (which should be zero in stationarity). We illustrate our results in the following table.

	P	$\overline{P}$	$\widehat{P}$
Q_1	9.999944	10.000033	9.999950
Q_2	8.127410	0.038961	0.048713

Means of the quantity $X_{20,1}$ (which should have mean zero in stationarity) for each of the proposals Q_1 and Q_2, and for each of the schemes P, $\overline{P}$, and $\widehat{P}$.

We thus see that, regardless of which scheme is used, the non-antithetic proposal Q_1 is unable to produce a simulatation of $X_{20,1}$ whose distribution is close to the stationary distribution (which would have a mean of zero). Rather, it always concentrates around the mean $\gamma = 10$ of the mode in which it starts. For the antithetic proposal Q_2, the original chain P is nearly periodic, so again the simulation of $X_{20,1}$ is far from stationarity and has an incorrect mean. However, the modified schemes $\overline{P}$ and $\widehat{P}$, used in combination with the antithetic proposal Q_2, each produce a simulation which is very close to stationarity (having mean close to zero).

This provides numerical support, in high dimensions, for the claim that the modifications $\overline{P}$ and $\widehat{P}$ may be useful to produce good distributional convergence from nearly-periodic chains.

8. Conclusion.

It is true that nearly-periodic Markov chains may have very low asymptotic variance when estimating functionals, even though they have very slow distributional convergence to stationarity. However, we have argued in this paper that simple modifications of such chains (involving using a random number of iterations) can produce chains which also have excellent converence properties. We have also provided a number of theoretical results concerning the distributional convergence rates of such chains.

It is possible that these ideas can best be used in conjunction with the creation of antithetic chains. Indeed, it may be possible (as in Example 7 above) to first modify the transitions of a given chain to create an antithetic chain, and then modify the number of iterations of the antithetic chain to create a chain with excellent convergence properties.

Appendix: Proof of Theorem 12.

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 $\mathbf{P}(I_i = 1) = 1 - \mathbf{P}(I_i = 0) = \epsilon$. Assume the β_i were chosen so that $\beta_1 + \ldots + \beta_j = T_j$. More generally, let $T_0 = 0$, and $T_k = \beta_1 + \ldots + \beta_k$ for $1 \leq k \leq j$.

We shall define *three* processes $\{X_t\}_{t=0}^{m+T_j}$, $\{X'_t\}_{t=0}^{m+j}$, 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''.

We shall also define auxiliary variables $\{d_t, A_t, N_t\}_{t=0}^{m+j}$, where: d_t is the indicator function of whether or not X' and X'' have coupled by time t; A_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 $d_0 = A_0 = N_0 = 0$. Then iteratively for $n \geq 0$, given $X'_n, X''_n, d_n, A_n, N_n, X_{A_n}$:

1. If $d_n = 1$, then we must have $X'_n = X''_n = X_{A_n} \equiv x$, in which case
   a. If $(X'_n, X''_n) \notin C \times C$ or $N_n = j$, then set $d_{n+1} = 1$, and $A_{n+1} = A_n + 1$, and $N_{n+1} = N_n$. Then choose $X'_{n+1} = X''_{n+1} = X_{A_{n+1}} \sim P(x, \cdot)$,
   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 $A_{n+1} = A_n + \beta_{N_{n+1}}$. Then choose $X'_{n+1} = X''_{n+1} = X_{A_{n+1}} \sim P^{\beta_{N_{n+1}}}(x, \cdot)$, Then fill in $X_{A_n+1}, X_{A_n+2}, \ldots, X_{A_{n+1}-1}$ according to the transition kernel P, conditional on the values of A_n, A_{n+1}, X_{A_n}, and $X_{A_{n+1}}$.
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 $A_{n+1} = A_n + 1$, and $N_{n+1} = N_n$. Then independently choose $X'_{n+1} = X_{A_{n+1}} \sim P(X'_n, \cdot)$, 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 $A_{n+1} = A_n + \beta_{N_{n+1}}$. Then
      i. If $I_{n+1} = 1$, choose $X'_{n+1} = X''_{n+1} = X_{A_{n+1}} \sim \nu(\cdot)$,

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

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

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 $X_{A_n+1}, X_{A_n+2}, \ldots, X_{A_{n+1}-1}$ according to the transition kernel P, conditional on the values of A_n, A_{n+1}, X_{A_n}, and $X_{A_{n+1}}$.

[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 A_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, $X_{A_k} = X'_k$ for all k, and $A_k = (k - N_k) + T_{N_k}$ for all k.

Lemma 8. *On the event* $\{N_{m+j} = j\}$, *we have* $X_{m+T_j} = X'_{m+j}$.

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

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

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

$$= P[X_{m+T_j} \neq X''_{m+j},\ N_k \geq j] + P[X_{m+T_j} \neq X''_{m+j},\ N_k \leq j - 1]. \qquad (10)$$

By Lemma 8,

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

$$\leq \ \mathbf{P}[I_1 = I_2 = \ldots = I_j = 0] \ = \ (1-\epsilon)^j \,, \tag{11}$$

which bounds the first term in (10).

Also,

$$P[X_{m+T_j} \neq X''_{m+j},\ N_{m+j} \leq j-1] \ \leq \ P[N_{m+j} \leq j-1] \,. \tag{12}$$

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 (5), 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 $\alpha A_\mu > 1$,

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

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

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

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

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

$$= \ \alpha^{-m-j} (\alpha A_\mu)^{j-1} \mathbf{E}[h(X'_0, X''_0)] \,. \tag{13}$$

The result now follows by plugging (11) and (13) into (10). ∎

Remark. If we could replace (12) by

$$P[X_{m+T_j} \neq X''_{m+j},\ N_{m+j} < j] \ \leq \ P[d_{m+j} = 0,\ N_{m+j} < j] \,, \tag{14}$$

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., 2001), leading to Proposition 7 above. However, (14) is not true for general β_i. Indeed, in general if $N_{m+j} < j$ then we will not have $X_{m+T_j} = X'_{m+j}$. Hence, we might have $X'_{m+j} = X''_{m+j}$ and $d_{m+j} = 1$, even though $X_{m+T_j} \neq X''_{m+j}$. 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 $X'_{m+j} = X_{m+T_j}$ no matter what; however, this then invalidates e.g. the drift condition (5), (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 (14) 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 Antonietta Mira and two anonymous referees for helpful comments and corrections.

REFERENCES

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

J. Besag and P.J. Green (1993), Spatial statistics and Bayesian computation. J. Royal Stat. Soc. B **55**, 25–37.

S.P. Brooks, P. Giudici, and G.O. Roberts (2001), Efficient construction of reversible jump MCMC proposal distributions. Preprint.

K.S. Chan and C.J. Geyer (1994), Discussion of Tierney (1994). Ann. Stat. **22**, 1747–1758.

J.B. Conway (1985), A course in functional analysis. Springer, New York.

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 (2001), Quantitative convergence rates for inhomogeneous Markov chains. Preprint.

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

P.J. Green (1995), Reversible jump MCMC computation and Bayesian model determination. Biometrika **82**, 711–732.

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, to appear.

C. Kipnis and S.R.S. Varadhan (1986), Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. Comm. Math. Phys. **104**, 1–19.

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.

A. Mira and C. Geyer (1999), Ordering Monte Carlo Markov chains. Technical Report No. 632, School of Statistics, University of Minnesota.

A. Mira (2001), Ordering and improving Monte Carlo Markov chain performance. Preprint.

G. Norman and V.S. Filinov (1969). Investigation of Phase Transitions by a Monte Carlo Method. High Temperature **7**, 216–222.

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.

C.J. Preston (1977), Spatial birth-death processes. Bull. Int. Statist. Inst. **46**, 371–391.

M. Reed and B. Simon (1972), Methods of modern mathematical physics. Volume I: Functional analysis. Academic Press, New York.

G.O. Roberts and J.S. Rosenthal (1997a), 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 (1997b), Geometric ergodicity and hybrid Markov chains. Elec. Comm. Prob. **2**, 13–25.

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.

G.O. Roberts and R.L. Tweedie (2000), Geometric L2 and L1 convergence are equivalent for reversible Markov chains. J. Appl. Prob., to appear.

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://markov.utstat.toronto.edu/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.

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

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.