Asymptotic Variance and Convergence Rates of Nearly-Periodic MCMC Algorithms

by

Jeffrey S. Rosenthal*

(October 4, 2001; last revised September 17, 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, including a trans-dimensional MCMC example, a card-shuffling example, and several antithetic Metropolis algorithms.

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 $P^\mu = \sum_m \mu\{m\} P^m$ for probability distributions μ on the non-negative integers. We prove various results about the spectra and quantitative convergence rates of such chains. We also consider a number of examples, including a trans-dimensional MCMC example, a card-shuffling example, and several antithetic Metropolis algorithms.

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 elsewhere, 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$(2n, 1/2) \approx n$. Hence, we call $\overline{P}^{2n}$ the *binomial modification* of P^n.

We thus see that minor modifications to the original, (which is 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

$$P(x, A) \;\geq\; \epsilon\, \nu(A)\,, \qquad x \in \mathcal{X}, \quad A \subseteq \mathcal{X} \tag{2}$$

(abbreviated as $P(x, \cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$), for some probability measure $\nu(\cdot)$ on $\mathcal{X}$. 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 later in this paper.

Finally, we note that the general idea of considering a random number of iterations is not new. For example, if $T_n \sim \mathrm{Unif}\{1, 2, \ldots, n\}$ (as opposed to $B_n \sim \mathrm{Binomial}(2n,\, 1/2)$), 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 significantly 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

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

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 $\sigma(P_0) \subseteq [-1, 1)$ (as discussed in e.g. 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. We begin with a result about asymptotic variance. (All proofs are given in the Appendix.)

Proposition 1. *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*

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

We conclude from this Proposition that the quantity

$$interval(P) \equiv 1 - \Lambda(P_0) \equiv 1 - \sup_{\lambda \in \sigma(P_0)} \lambda \tag{3}$$

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

We next turn to 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, and write $\|\nu\|_{L^2(\pi)} = \int_{\mathcal{X}} (\frac{d\nu}{d\pi})^2\, d\pi$ for $L^2(\pi)$ distance (with $\|\nu\|_{L^2(\pi)} = \infty$ if ν is not absolutely continuous with respect to π). Then we have the following.

Proposition 2. *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*

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

Proposition 2 says that for large n, we roughly have

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

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

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

$$gap(P) \equiv 1 - r(P_0) \equiv 1 - \sup_{\lambda \in \sigma(P_0)} |\lambda| \tag{4}$$

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

Comparing (4) with (3), we see that they differ only by the absolute values signs; and this distinction characterises the difference between good distributional convergence, and good asymptotic variance, properties of Markov chains.

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

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 3. *If P is reversible, then gap$(\overline{P}^2) = \zeta(interval(P))$, and $r(\overline{P}_0^{2n}) = \big(1 - \zeta(interval(P))\big)^n$. [Hence, if $interval(P)$ is small, then gap$(\overline{P}^2) \approx interval(P)$ and $r(\overline{P}_0^{2n}) \approx \big(\Lambda(P_0)\big)^n$. On the other hand, if $interval(P) \approx 2$ as for an extremely antithetic chain, then gap$(\overline{P}^2) \approx 1$, indicating extremely fast convergence.]*

It follows from Theorem 3 that the convergence rate properties of $\overline{P}^2$ are essentially the same as the asymptotic variance properties of P. That is, the simple modification of using $\overline{P}^{2n}$ instead of P^n gives us distributional convergence which is as fast as would be indicated by the asymptotic variance properties. The chain $\overline{P}^{2n}$ is the *binomial modification* of P^n.

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 $\{X_n\}$ itself. Set $\theta_n(\lambda) = \frac{1}{2}\lambda^n + \frac{1}{2}\lambda^{n+1} = \frac{1}{2}\lambda^n(1+\lambda)$, so that if $\lambda \approx 1$ then $\theta_x(\lambda) \approx \lambda^n$, while if $\lambda \approx -1$ then $\theta_n(\lambda) \approx 0$. Then we have the following.

Theorem 4. *If P is reversible, then*

$$r(\widehat{P}_0^n) \;=\; \sup_{\lambda \in \sigma(P_0)} |\theta_n(\lambda)| \;\leq\; \max\left(\frac{n^n}{2(n+1)^{n+1}},\; \theta_n\left(\Lambda(P_0)\right)\right).$$

[Hence, if $interval(P)$ is small, then ignoring the $\frac{n^n}{2(n+1)^{n+1}}$ term, we will have $r(\widehat{P}_0^n) \approx \big(\Lambda(P_0)\big)^n$. Also, if $interval(P) \approx 2$ as for an extremely antithetic chain, then $r(\widehat{P}_0^n) \approx 0$ indicating extremely fast convergence.] However, if $-\frac{n}{n+1} \in \sigma(P_0)$, then $r(\widehat{P}_0^n) \geq \frac{n^n}{2(n+1)^{n+1}}$.

Theorem 4 thus says that, if P has good distributional convergence properties, then ignoring the $\frac{n^n}{2(n+1)^{n+1}}$ term, $\widehat{P}^n$ will have good asymptotic variance properties. However, the $\frac{n^n}{2(n+1)^{n+1}}$ term is problematic for $\widehat{P}^n$, since as $n \to \infty$ it is asymptotically equal to $\frac{1}{2en}$ which is sub-exponential. This could be a problem if $\sigma(P_0)$ contains points arbitrarily close to -1, which could be equal (or nearly equal) to $-\frac{n}{n+1}$ for arbitrarily large n.

Remark 5. A comparison of Theorems 3 and 4 indicates that if $\Lambda(P_0) \approx 1$ as it usually would be, then to first order as $n \to \infty$, and ignoring the $\frac{n^n}{2(n+1)^{n+1}}$ term, both $\overline{P}^{2n}$ and $\widehat{P}^n$ have spectral radius $\big(\Lambda(P_0)\big)^n$. That is, each of these modified schemes has distributional convergence rate approximately as good as the asymptotic variance rate of the original chain P, even if the original chain is periodic. Hence, in some sense $\overline{P}^{2n}$ and $\widehat{P}^n$ are "equally good". However, as noted above, the $\frac{n^n}{2(n+1)^{n+1}}$ term could be problematic for $\widehat{P}^n$. Furthermore, we shall see in Example 4 below that $\overline{P}^{2n}$ is more "flexible" than $\widehat{P}^n$, and is therefore better in some sense.

The above results indicate that the modified chains $\overline{P}^{2n}$ and $\widehat{P}^n$ provide good distributional convergence rates in the reversible case. However, to analyse non-reversible chains, or to exploit the specific structure of certain reversible chains, additional modifications may be called for. We therefore generalise our definitions as follows.

As a generalisation of $\overline{P}$, we shall consider the chain $P^\mu \equiv \sum_m \mu\{m\} P^m$, where μ is some fixed probability measures on the non-negative integers (and where $P^0 = I$). That is, $(P^\mu)^n$ corresponds to taking $T_1 + \ldots + T_n$ steps according to P, where $\{T_i\}$ are i.i.d. $\sim \mu$ and are chosen independently of the chain itself. Equivalently, $(P^\mu)^n$ corresponds to taking T steps according to P, where $T \sim \mu^{*n}$, the n-fold convolution of μ with itself. (In the language of Meyn and Tweedie (1993), P^μ is a *sampled chain*.)

Thus, $\overline{P} = P^\mu$ 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 $\mu\{0\} = \mu\{1\} = \mu\{2\} = 1/3$ (cf. Example 3 below). Similarly, to ensure that the chain always moves at least once, we might consider $\mu\{1\} = \mu\{2\} = 1/2$ (cf. Example 1 below).

As a generalisation of $\widehat{P}^n$, we shall consider $\widehat{{}^\mu P^n} \equiv P^\mu P^n$. That is, $\widehat{{}^\mu P^n}$ corresponds to taking one step according to P^μ, followed by n steps according to P. Equivalently, it corresponds to taking $n + T$ steps according to P, where $T \sim \mu$ is chosen independently of the chain.

Thus, $\widehat{P}^n = P^\mu P^n$ for the special case $\mu\{0\} = \mu\{1\} = 1/2$. In fact, running $\widehat{{}^\mu P^n}$ on an initial distribution ρ is precisely equivalent to running P^n on the initial distribution ρP^μ. Hence, modifications such as $\widehat{P}$ (as opposed to $\overline{P}$) correspond merely to choosing a more intelligent initial distribution. Below we shall consider both $(P^\mu)^n$ and $\widehat{{}^\mu P^n}$, however we shall concentrate more on $(P^\mu)^n$, since initial distributions are really a separate topic, and since by Remark 5, $(P^\mu)^n$ is in some ways better than $\widehat{{}^\mu P^n}$ anyway.

Remark 6. In the special case where $\mu\{0\} = \mu\{1\} = \ldots = \mu\{d-1\} = 1/d$ for some integer $d \geq 2$, we see that $\widehat{{}^{\mu}P^{n}} = \frac{1}{d}\sum_{i=n}^{n+d-1} P^i$, and corresponds to choosing uniformly from among d consecutive values of the original chain. This modification is well-known to be a way of guaranteeing convergence (though without specified quantitative rate) of a Markov chain of period d, see e.g. p. 31 of Orey (1971), or p. 71 of Hoel et al. (1972). (For more on this special case, see Theorem 9 below.) Hence, our $\widehat{{}^{\mu}P^{n}}$ is a generalisation of a well-known Markov chain modification.

5. Convergence rate bounds.

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 7. *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 $P^{m_0}(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$ as in (2), where $\epsilon > 0$ and where $\nu(\cdot)$ is any probability measure on $\mathcal{X}$. Then*

$$\|P^m(x,\cdot) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^{\lfloor m/m_0 \rfloor}\,,$$

where $\lfloor r \rfloor$ means the greatest integer not exceeding r.

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, and let $P^{\mu} = \sum_m \mu\{m\} P^m$ as before. The following result follows immediately by applying Proposition 7 to the chain P^{μ} instead of the chain P.

Theorem 8. *Suppose $(P^{\mu})^{m_0}(x,\cdot) \geq \epsilon\nu(\cdot)$ for all $x \in \mathcal{X}$, where $\epsilon > 0$ and where $\nu(\cdot)$ is any probability measure on $\mathcal{X}$. Then for all $x \in \mathcal{X}$,*

$$\|(P^{\mu})^m(x,\cdot) - \pi(\cdot)\|_{TV} \;\equiv\; \|\mathcal{L}(X_{T_m} \mid X_0 = x) - \pi(\cdot)\|_{TV} \;\leq\; (1-\epsilon)^{\lfloor m/m_0 \rfloor},$$

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

Theorem 8 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, practical mechanism for obtaining a sample from a given distribution $\pi(\cdot)$, whenever an MCMC algorithm with good asymptotic variance properties is available (even if the algorithm is periodic or nearly so). Some examples applying Theorem 8 are presented in Section 6. We also note that minorisation conditions can sometimes be approximately verified numerically even when they are analytically intractable; see e.g. Cowles and Rosenthal (1998), Cowles (2001).

Now, Theorem 8 can only be applied if P^{μ} is *uniformly ergodic*, i.e. satisfies a minorisation condition on the entire state space $\mathcal{X}$. On the other hand, there is also a great deal of interest in convergence rates for non-uniformly ergodic chains (see e.g. Meyn and Tweedie, 1994; Rosenthal, 1995b; Roberts and Tweedie, 1999; Jones and Hobert, 2001). We shall consider corresponding results about P^{μ} for non-uniform chains in subsequent research.

Regarding $\widehat{{}^{\mu}P^{n}}$, convergence results are somewhat more awkward since the near-periodicity part is dealt with only at the first iteration, rather than repeatedly at each iteration as with $(P^{\mu})^n$. However, in the *exactly* periodic case, with μ as in Remark 6, we have the following simple result.

Theorem 9. *Suppose a Markov chain $P(x,\cdot)$ has stationary distribution $\pi(\cdot)$ and period $d \geq 2$, so the state space $\mathcal{X}$ can be partitioned as $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \ldots \cup \mathcal{X}_d$, with $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ for $1 \leq i \leq d-1$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$. Suppose that for each i there is a probability measure $\nu_i(\cdot)$ on $\mathcal{X}$, such that for some fixed $\epsilon > 0$,*

$$P^{m_0}(x,\cdot) \geq \epsilon\, \nu_i(\cdot)\,, \qquad x \in \mathcal{X}_i\,, \quad 1 \leq i \leq d-1\,. \tag{5}$$

Let $\mu\{0\} = \mu\{1\} = \ldots = \mu\{d-1\} = 1/d$. *Then*

$$\|\widehat{{}^{\mu}P^{m}}(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} \;\le\; (1-\epsilon)^{\lfloor m/m_0 \rfloor}\,, \qquad x \in \mathcal{X}\,.$$

Note that condition (5) requires a separate minorisation condition on each piece $\mathcal{X}_i$, but does *not* require a single minorisation condition valid on all of $\mathcal{X}$. Furthermore, if $P(x,\cdot) \ge \epsilon\,\nu_j(\cdot)$ for all $x \in \mathcal{X}_j$ for some *fixed* j, then it follows easily that (5) is satisfied with the same ϵ, but with $m_0 = d$. On the other hand, it is non-trivial to generalise Theorem 9 to bound the distributional convergence of $\widehat{{}^{\mu}P^{n}}$ when the chain is *nearly* periodic but not *exactly* periodic; for in that case, it does not follow that $\pi(\mathcal{X}_i) = 1/d$ for each i, which is required for the proof.

Remark. As observed in Roberts and Rosenthal (2000), small-set conditions of the form $P(x,\cdot) \ge \epsilon\nu(\cdot)$ for all $x \in C$, can be replaced by *pseudo-small* conditions of the form $P(x,\cdot) \ge \epsilon\nu_{xy}(\cdot)$ and $P(y,\cdot) \ge \epsilon\nu_{xy}(\cdot)$ for all $x, y \in C$, where ν_{xy} can depend on the specific pair (x,y), without affecting any bounds which use coupling (such as Theorems 8 and 9 above). So, this provides an immediate generalisation of the above results; though for ease of exposition, we do not pursue that here.

6. 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 $x \in [0,1]$, $P(x,\cdot) = \mathrm{Unif}[1,2]$, while for $x \in (1,2]$, $P(x,\cdot) = \mathrm{Unif}[0,1]$. This chain is reversible with respect to $\pi(\cdot) = \mathrm{Unif}[0,2]$.

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 $x \le 1$ 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, $interval(P) = 1$ while $gap(P) = 0$. 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 3, we see that $gap(\overline{P}) = \zeta(1) = 3/4$, i.e. the binomial-modified chain $\overline{P}^m$ converges to $\pi(\cdot)$ extremely quickly, as does $\widehat{P}^m$. (Note, however, that 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.)

To apply Theorem 8, we see that we cannot have $P^{k_0}(x, \cdot) \geq \epsilon\nu(\cdot)$ with $\epsilon = 1$, for all $x \in \mathcal{X}$ for any k_0 and $\nu(\cdot)$. Rather than settle for $\epsilon < 1$ here, we resort to a trick by setting $\mu\{1\} = \mu\{2\} = 1/2$. We then have $P^\mu(x, \cdot) = \pi(\cdot)$ for all $x \in \mathcal{X}$, so we can take $\epsilon = 1$ in Theorem 8, to get that $\|(P^\mu)^m(x, \cdot) - \pi(\cdot)\|_{TV} = 0$ for any $m \geq 1$ and all $x \in \mathcal{X}$.

We conclude that in this example, the binomially-modified chain $\overline{P}$ converges at rate $(1/4)^n$, and in fact the chain P^μ with $\mu\{1\} = \mu\{2\} = 1/2$ converges exactly in just one iteration, even though the original chain is periodic.

Example 2. *A nearly-periodic chain.*

Again let $\mathcal{X} = [0, 2]$, and suppose now that we only know there are some $\delta_1, \delta_2 > 0$ such that for $x \in [0, 1]$, $P(x, \cdot) \geq \delta_1 \mathrm{Unif}[1, 2]$, while for $x \in (1, 2]$, $P(x, \cdot) \geq \delta_2 \mathrm{Unif}[0, 1]$. (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 $\pi(\cdot) = \mathrm{Unif}[0, 2]$.) 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 8. Indeed, we have $P(x, \cdot) \geq \delta_2 \mathrm{Unif}[0, 1]$ for $x \in (1, 2]$, and $P^2(x, \cdot) \geq \delta_1\delta_2 \mathrm{Unif}[0, 1]$ for $x \in [0, 1]$. Similarly $P(x, \cdot) \geq \delta_1 \mathrm{Unif}[1, 2]$ for $x \in [0, 1]$, and $P^2(x, \cdot) \geq \delta_1\delta_2 \mathrm{Unif}[1, 2]$ 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 $\nu(\cdot) = \mathrm{Unif}[0, 2]$.

Hence, by Theorem 8, $\|(P^\mu)^m(x, \cdot) - \pi(\cdot)\|_{TV} \leq (1 - \delta_1\delta_2)^m$. This provides a bound on how many iterations of P^μ should be done (or equivalently, what *random* number of 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 $\pi(\cdot) = \mathrm{Unif}(\mathcal{X})$. 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 $B_n \sim \mathrm{Binomial}(2n, 1/2)$. 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.

For this choice of μ, we can take $\epsilon = 1$ in Theorem 9, and this shows that the chain $\widehat{{}^{\mu}P^n}$ also converges to stationarity in just one step. Thus, for this example, $(P^\mu)^n$ and $\widehat{{}^{\mu}P^n}$ work equally well, when μ is uniform on $\{0, 1, 2, \ldots, D - 1\}$.

Example 4. *A misspecified μ distribution.*

Consider the previous example with $D = 3$, but suppose we have erroneously set $\mu\{0\} = \mu\{1\} = 1/2$. That is, suppose the chain has period 3, but we mistakenly thought it had near-periodicity problems corresponding to period 2.

In this case, we will have $\|\widehat{{}^{\mu}P^n}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} = 1/3$, for any $x \in \mathcal{X}$ and any $n \geq 0$. This is because $\widehat{{}^{\mu}P^n}(x, \cdot)$ will always be concentrated equally on some two of the three points in $\mathcal{X}$. Hence, in particular, $\|\widehat{{}^{\mu}P^n}(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \not\to 0$, so that $\widehat{{}^{\mu}P^n}(x, \cdot)$ does not converge in distribution at all.

On the other hand, consider $(P^\mu)^n$. We see by inspection that, say, $(P^\mu)^2(x, \cdot) \geq \epsilon\, \pi(\cdot)$ for all $x \in \mathcal{X}$ if $\epsilon = 3/4$, since e.g. $(P^\mu)^2(1, \cdot)$ has distribution $(1/4, 1/2, 1/4)$. Hence, it follows from Theorem 8 that

$$\|(P^\mu)^n(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \;\leq\; (1 - 3/4)^{\lfloor n/2 \rfloor} \;=\; (1/4)^{\lfloor n/2 \rfloor},$$

so that $\|(P^\mu)^n(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \to 0$, and quite quickly at that.

We conclude from this that $(P^\mu)^n$ is more "flexible" than $\widehat{{}^\mu P^n}$, since it converges quickly even if μ is misspecified.

Example 5. *A trans-dimensional Metropolis-Hastings algorithm.*

Consider the chain of Lemma 1 of Brooks, Guidichi, and Roberts (2002). This is a very simple example of a trans-dimensional 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 $\int_0^1 f(y)dy = 1$. (Here e is a "0-dimensional" single point, with $e \notin [0, 1]$.) 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., 2002) 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 8, 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

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

Example 6. *A multi-dimensional antithetic Metropolis algorithm.*

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^n, $\overline{P}^{2n}$, and $\widehat{P}^n$. For each of the six combinations, we ran 100,000 separate runs, each for $n = 20$ 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 summarise our results in Table 1.

	P^n	$\overline{P}^{2n}$	$\widehat{P}^n$
Q_1	9.999944	10.000033	9.999950
Q_2	8.127410	0.038961	0.048713

Table 1. Means of the quantity $X_{20,1}$ (which should have mean zero in stationarity), under each of the proposals Q_1 and Q_2, and for each of the schemes P^n, $\overline{P}^{2n}$, and $\widehat{P}^n$. Here $n = 20$ and $\mu\{0\} = \mu\{1\} = 1/2$. Note that the means are computed only from the final iteration n, as opposed to averaging over all iterations from 0 to n, to examine the distribution at time n rather than estimator properties.

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, and fails to mix properly. 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 $\overline{P}^{2n}$ and $\widehat{P}^n$ are useful to produce good distributional convergence from nearly-periodic original chains P^n.

Example 7. *A Bayesian posterior distribution.*

Consider the following statistical model. Suppose that conditional on Z, the variables $\{Y_i\}_{i=1}^J$ are conditionally i.i.d., with $P(a < Y_i - Z \leq b) = \int_a^b C\, e^{-\sqrt{|x|}}\, dx$ for $a < b$, for appropriate normalising constant C. Take the prior distribution $Z \sim$ Cauchy, so $P(a < Z \leq b) = \int_a^b \frac{dx}{\pi\,(1+x^2)}$). Let $\pi(\cdot)$ be the posterior distribution of Z, conditional on the observed data $Y_1, \ldots, Y_J$; that is, $\pi(\cdot) = \mathcal{L}(Z \,|\, Y_1, \ldots, Y_J)$. The formula for $\pi(\cdot)$ is

$$\pi(dz) \;\propto\; \frac{dz}{\pi\,(1+z^2)} \prod_{i=1}^{J} e^{-\sqrt{|z-Y_i|}}\,.$$

Thus, $\pi(\cdot)$ is a simple example of a Bayesian posterior distribution, for which MCMC algorithms are very widely used.

Suppose, as in Example 6, that we run a Metropolis algorithm for $\pi(\cdot)$, with the antithetic proposal $Q(z,\cdot) = N(-z,\, \sigma^2)$. This ensures good asymptotic variance properties of the algorithm. But how quickly does this algorithm converge to $\pi(\cdot)$ in distribution?

For many values of the data $\{Y_i\}$, the Metropolis algorithm will not have any near-periodicity problems. However, suppose that the data happens to be approximately *symmetric* around 0. In that case, $\pi(\cdot)$ will also be approximately symmetric, so there will be very few rejections. In this case, it is possible that the algorithm could jump back and forth between very positive and very negative values, exhibiting near-periodic behaviour and therefore poor distributional convergence properties. Use of either $\overline{P}$ or $\widehat{P}$ should alleviate this problem.

Suppose for definiteness that $J = 80$, with $Y_1 = \ldots = Y_{40} = -50$, and $Y_{41} = \ldots = Y_{80} = +50$. We begin our Markov chain at $X_0 = -40$, and run $\{X_n\}$ as a Metropolis algorithm with proposal $Q(z,\cdot) = N(-z,\, \sigma^2)$ where $\sigma = 0.01$, for 20 iterations, under the three schemes P^{20}, $\overline{P}^{40}$, and $\widehat{P}^{20}$.

Under this scheme, we compute (by running the algorithm for 100,000 repetitions and averaging) that with P^{20}, we have $E(X) \doteq -10.25$, which is very far from its mean of 0, indicating poor distributional convergence. On the other hand, with $\overline{P}^{40}$, we have $E(X) \doteq 0.113$ which is very close to 0. Similarly, with $(\widehat{P})^{20}$, we have $E(X) \doteq -0.258$ which is again very close to 0.

We conclude from this that, for this Bayesian model and data, the schemes $\overline{P}$ and $\widehat{P}$ both have excellent distributional convergence properties, even though the original chain P has major problems of near-periodicity, and even though computing $\overline{P}^{40}$ and $\widehat{P}^{20}$ requires virtually no additional computation compared to P^{40}.

Example 8. *Card-shuffling by random transpositions.*

Diaconis and Shashahani (1981), see also Diaconis (1988), consider the following model for shuffling a deck of cards (of size D). The deck's state is represented as an element σ of the symmetric group S_D of permutations of $\{1, 2, \ldots, D\}$. Given a state σ_n at time n, the state σ_{n+1} at time $n+1$ is chosen as follows. With probability p, we do nothing, so that $\sigma_{n+1} = \sigma_n$. Otherwise, with probability $1-p$, we choose a pair (i,j) uniformly at random from the set $\{(i,j); 1 \le i < j \le D\}$. We then *transpose* cards i and j in the deck, leaving the other cards fixed, so that $\sigma_{n+1}(i) = \sigma_n(j)$, $\sigma_{n+1}(j) = \sigma_n(i)$, and $\sigma_{n+1}(k) = \sigma_n(k)$ for $k \neq i, j$.

This model defines a Markov chain on S_D, whose stationary distribution $\pi(\cdot)$ is the uniform distribution on S_D. Diaconis and Shashahani (1981) prove that, if $p = 1/D$, then $\|P^{\frac{1}{2}D\log D + cD}(\sigma, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \le a e^{-2c}$ for some constant a, implying that roughly $\frac{1}{2}D\log D + O(D)$ iterations are necessary to converge to $\pi(\cdot)$. On the other hand, if $p = O(D^{-2})$, then more than $O(D^2)$ iterations are needed for convergence (see Diaconis, 1988, p. 44), and in general the closer p gets to 0, the closer to periodic the chain becomes.

Suppose now that $p = 0$, so that there is no holding probability at all. In that case, the *sign* of the permulation σ_n keeps alternating, so that $\|P^n(\sigma, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \ge 1/2$ for all n (regardless of σ), and P^n does not converge in distribution at all.

However, use of either $\overline{P}^{2n}$ or $\widehat{P}^n$ again solves this convergence problem. We ran each of the three schemes for $D = 52$ (as for an ordinary deck of cards), beginning in the identity state (corresponding to a completely sorted deck), and with $n = 1143 \doteq \frac{1}{2} D \log D + 20D$. For each scheme, we computed (by averaging 100,000 repetitions) the mean values of $\sigma(1)$ (the position of the Ace of Spades), and of $\mathrm{sign}(\sigma)$. We compared these to the known means under $\pi(\cdot)$. The results are summarised in Table 2.

	$\pi(\cdot)$	P^n	$\overline{P}^{2n}$	$\widehat{P}^n$
$\sigma(1)$	26.5	26.526790	26.465390	26.541590
$\mathrm{sign}(\sigma)$	0	-1.000000	0.001180	0.005020

Table 2. Means of the quantities $\sigma(1)$ (the position of the Ace of Spades), and of $\mathrm{sign}(\sigma)$ (the permutation's sign), in the stationary distribution $\pi(\cdot)$, and under each of the schemes P^n, $\overline{P}^{2n}$, and $\widehat{P}^n$. Here $D = 52$, and $n = 623$, and $\mu\{0\} = \mu\{1\} = 1/2$.

We see from Table 2 that $\sigma(1)$ converges well under all schemes. However, $\mathrm{sign}(\sigma)$ fails to converge under P^n, but converges well under $\overline{P}^{2n}$ and $\widehat{P}^n$. Hence, once again, $\overline{P}$ and $\widehat{P}$ manage to alleviate the periodicity problems inherent in P, with very little additional computational effort.

Example 9. *A random-effects model.*

Finally, we consider the random-effects model studied analytically in Rosenthal (1996), using the baseball data of Efron and Morris (1975) and Morris (1983). Here m and A are parameters with prior distributions flat and InverseGamma$(-1, 2)$, respectively. Then $\theta_i \,|\, m, A \sim N(m, A)$, with $\{\theta_i\}_{i=1}^K$ conditionally independent. Finally, $Y_i \,|\, \{\theta_j\}_{j=1}^K \sim N(\theta_i, V)$ with $\{Y_i\}_{i=1}^K$ conditionally independent, and with $V = 0.00434$ estimated directly from the data. We condition on the baseball data $\{Y_i\}$, with $K = 18$, as presented in Table 1 of Morris (1983). This gives rise to a posterior distribution $\pi(\cdot)$ on $\mathbf{R}^{20}$, corresponding to the 20-tuple $(A, m, \theta_1, \ldots, \theta_K)$.

We run a Gibbs sampler on $(A, m, \theta_1, \ldots, \theta_K)$ for this $\pi(\cdot)$. It was proved in Rosenthal (1996) that, if we begin with $\theta_i = \frac{1}{K} \sum_j Y_j$ for all i, then this Markov chain satisfies

$\|P^{140}(x,\cdot) - \pi(\cdot)\|_{\mathrm{TV}} < 0.01$, i.e. it converges in distribution after 140 iterations. So, there are no problems with distributional convergence of the original chain. However, we here examine the distributional convergence of the $\overline{P}$ and $\widehat{P}$ schemes as well. The results are summarised in Table 3.

	P^n	$\overline{P}^{2n}$	$\widehat{P}^n$
A	0.320066	0.319072	0.319450
m	0.265295	0.266219	0.266159
θ_1	0.392684	0.393402	0.393401
θ_5	0.312400	0.312379	0.312134
θ_{18}	0.149678	0.149867	0.149684

Table 3. Means of various quantities, for the Gibbs sampler for the random-effects model, for each of the schemes P^n, $\overline{P}^{2n}$, and $\widehat{P}^n$. Here $n = 140$ and $\mu\{0\} = \mu\{1\} = 1/2$. Note that the means are approximately equal under all three schemes.

We see from Table 3 that, in this case, all variables converge well under all three of the schemes. Hence, for this model, it is not necessary to use $\overline{P}$ or $\widehat{P}$ to achieve good distributional convergence. On the other hand, no harm is done either, indicating that there is no problem with using $\overline{P}$ or $\widehat{P}$ even if your chain turns out not to have any near-periodicity problems at all.

7. Discussion and 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 a random number of iterations) can produce chains which in addition 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 6 above) to first modify the transitions of a given chain to create an antithetic chain with the same stationary distribution, and then to randomly modify (as described herein) the number of iterations of the antithetic chain to create a chain with excellent convergence properties.

From the practical point of view, the proposed Markov chain modifications raise a number of issues. When should a user bother with such modifications? If you do use them, should you use $(P^\mu)^n$ or $\widehat{{}^\mu P^n}$? And, what choice of μ should be made?

We would answer these questions as follows. For routine uses of MCMC algorithms (especially for Metropolis algorithms with many rejections), periodicity-related problems might not arise. So, it is not necessary to use these modifications on a routine basis. However, even when there are no periodicity-related problems, these modifications do no harm (cf. Example 9 herein), and there is very little additional work required to use them (e.g. just sampling from one binomial distribution), so there is no particular reason *not* to use them on a routine basis, either.

More significantly, if there is any hint or possibility of periodicity-related problems, then these modifications should be used. Indeed, they might well provide greatly improved distributional convergence (cf. Examples 1 through 8 herein).

As for which of $(P^\mu)^n$ and $\widehat{{}^\mu P^n}$ to use: Often they are both equally good at overcoming periodicity problems (cf. Example 3), in which case either one could be used. However, as discussed in Remark 5, $(P^\mu)^n$ has better guaranteed convergence properties. Furthermore, as seen in Example 4, $(P^\mu)^n$ is more flexible if μ is incorrectly specified. Thus, overall we recommend $(P^\mu)^n$ over $\widehat{{}^\mu P^n}$, though often either one will suffice.

As for the choice of μ: If it is known that the chain is approximately periodic with period d, then it makes sense to let μ be uniform on $\{0, 1, \ldots, d-1\}$ as in Remark 6. Or, if it is know that the chain has special structure that can be exploited, then it is wise to choose μ on that basis (cf. Example 1 herein). However, in the absense of such additional information, we recommend the simplest case of choosing $\mu\{0\} = \mu\{1\} = 1/2$, so that P^μ reduces to the binomial modification $\overline{P}$. This modification is simple to implement, and is flexible enough to handle a variety of periodicity-related problems (cf. Example 4 herein).

In conclusion, we feel that the proposed Markov chain modifications protect against periodicity-related distributional convergence problems, while doing no harm and requiring minimal effort to implement. Thus, they appear to be worthwhile, to ensure good distributional convergence in various MCMC applications.

Appendix: Proofs of Theoretical Results

Proof of Proposition 1. 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

$$\langle g,\, h(P_0)g\rangle = \int_{\sigma(P_0)} h(\lambda) E_{g,P_0}(d\lambda)\,, \tag{6}$$

for every bounded Borel-measurable function $h : \mathbf{R} \to \mathbf{R}$. In particular, setting $h(P_0) \equiv 1$ in (6), we see that

$$\langle g,\, g\rangle = \pi(g^2) = \int_{\sigma(P_0)} E_{g,P_0}(d\lambda)\,. \tag{7}$$

Then it is known (Kipnis and Varadhan, 1986; see also Geyer, 1992; Chan and Geyer, 1994; Mira and Geyer, 1999) that

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

Now, since $\lambda \to \frac{1+\lambda}{1-\lambda}$ is an increasing function for $\lambda \in \sigma(P_0) \subseteq [-1, 1)$, we have from the above 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 (7). Also $\Lambda < 1$, so $1 + \Lambda < 2$. ■

Proof of Proposition 2. It follows from Roberts and Rosenthal (1997b) (cf. Roberts and Tweedie, 2000, Theorem 3) that

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

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

We have

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

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

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

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

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

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

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

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

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

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

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

The result follows. ■

Proof of Theorem 3. We have that

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

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

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

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,

$$r(\overline{P}_0^2) = \sup_{\lambda \in \sigma(\overline{P}_0^2)} |\lambda| = \sup_{\lambda \in \sigma(\overline{P}_0^2)} |\eta(\lambda)| = \sup_{\lambda \in \sigma(P_0)} \eta(\lambda) = \eta\left(\sup_{\lambda \in \sigma(P_0)} \lambda\right).$$

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

Proof of Theorem 4. 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 equally then follows by taking absolute values and supremums. The inequality and final statement follow since $\sup_{-1\leq x\leq 0} |\theta_n(x)| \;=\; \frac{n^n}{2(n+1)^{n+1}}$, with the sup occurring at $x = -\frac{n}{n+1}$, and $\theta_n(x)$ is non-negative and increasing for $x \geq 0$. ■

Proof of Theorem 9. It follows from the hypotheses that $\pi(\mathcal{X}_i) = 1/d$ for all i. Choose $Y_0 \sim \pi(\cdot)$, and let j be such that $Y_0 \in \mathcal{X}_j$. Then let $0 \leq T \leq d-1$ be such that $P^T(x, \mathcal{X}_j) = 1$. It then follows that $T \sim \mu(\cdot)$.

We now let $X_0 = x$, and then run $\{X_{n+T}\}_{n=0}^\infty$ and $\{Y_n\}_{n=0}^\infty$ jointly. It follows from the minorisation condition that we can force them to have probability ϵ of becoming equal, once every m_0 iterations. We conclude that we can define them jointly to ensure that $P(X_{m_0 n+T} \neq Y_{m_0 n}) \leq (1-\epsilon)^n$. It then follows from the standard *coupling inequality* (see e.g. Lindvall, 1992; or Rosenthal, 1995a, Appendix) that $\|\mathcal{L}(X_{m_0 n+T}) - \mathcal{L}(Y_{m_0 n})\|_{\mathrm{TV}} \leq (1-\epsilon)^n$. But $\mathcal{L}(X_{m_0 n+T}) = \widehat{\mu P^{m_0 n}}$, and by stationarity $\mathcal{L}(Y_{m_0 n}) = \pi(\cdot)$. The result follows by setting $n = \lfloor m/m_0 \rfloor$. ∎

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 (2002), Efficient construction of reversible jump MCMC proposal distributions. J. Royal Stat. Soc. B, to appear.

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.

P. Diaconis (1988), Group Representations in Probability and Statistics. IMS Lecture Series volume **11**, Institute of Mathematical Statistics, Hayward, California.

P. Diaconis and M. Shashahani (1981), Generating a Random Permutation with Random Transpositions. Z. Wahrscheinlichkeitstheorie Verw. Gebiete **57**, 159–179.

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.

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

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.

P.G. Hoel, S.C. Port, and C.J. Stone (1972), Introduction to Stochastic Processes. Waveland Press, Prospect Heights, IL.

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.

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

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.

S. Orey (1971), Lecture notes on limit theorems for Markov chain transition probabilities. Van Nostrand Reinhold, London.

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 (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.