Small and Pseudo-Small Sets for Markov Chains

by

Gareth O. Roberts* and Jeffrey S. Rosenthal**

(April 2000; last revised December 2000.)

In this paper we examine the relationship between small sets and their generalisation, pseudo-small sets. We consider conditions which imply the equivalence of the two notions, and give examples where they are definitely different. We give further examples where sets are both pseudo-small and small, but the minorisation constants implied by the two notions are different. Applications of recent computable bounds results are given and extended. We also give a result linking the ideas of monotonicity and minorisation. Specifically we demonstrate that if a non-monotone chain satisfies a minorisation condition, and furthermore is stochastically dominated by a monotone chain which satisfies a Lyapunov drift condition, then a probability construction exists which incorporates both the bounding process and the minorisation condition.

Keywords. Coupling, convergence rates, small set, minorisation condition, total variation distance.

1. Introduction.

A well-known concept in the theoretical study of Markov chains on general state spaces is the property of *small set*, or *minorisation condition.* This is the property that all of the transition probability distributions $P^{n_0}(x, \cdot)$, for some fixed $n_0 \in \mathbf{N}$ and for all x in some subset C, all have a certain non-zero component in common. In symbols, $P^{n_0}(x, \cdot) \geq \epsilon\, \nu(\cdot)$ for all $x \in C$. Such small sets arise often in the theoretical study of Markov chain Monte Carlo (MCMC) algorithms; see e.g. Smith and Roberts (1993), Tierney (1994), Gilks, Richardson and Spiegelhalter (1996), and Roberts and Rosenthal (1998).

* Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, England. Internet: `g.o.roberts@lancaster.ac.uk`.

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

If the small set C is hit infinitely often with probability 1, then it allows for the construction of *regeneration times* (cf. Athreya and Ney, 1978; Nummelin, 1978, 1984; Asmussen, 1987; Mykland, Tierney, and Yu, 1995). If the mean interarrival time is finite, then such regeneration times guarantee the existence of a *stationary distribution*, and furthermore provide some control (in terms of hitting times of C) over how quickly the chain converges to this stationary distribution.

Small sets also allow for the construction of *couplings* (cf. Lindvall, 1992; Meyn and Tweedie, 1993; Rosenthal, 1995a; see Appendix for details), whereby two different copies of the chain become equal with positive probability (since they may both update from the same distribution $\nu(\cdot)$). Such couplings also provide bounds on convergence rates to stationary distributions (though without guaranteeing the *existence* of a stationary distribution), through the *coupling inequality.*

Finally, since small sets provide a condition which holds for *all* elements of C simultaneously, they can also be used to construct *coalescence* (see e.g. Murdoch and Green, 1998), whereby copies of the chain started at *all* elements of the state space *all* become equal simultaneously. This is especially important in exact sampling schemes such as *coupling from the past* (Propp and Wilson, 1996) and *Fill's algorithm* (Fill, 1998; Fill, Machida, Murdoch, and Rosenthal, 1999).

In a different direction, small sets can be used to construct *shift-couplings* (cf. Aldous and Thorrison, 1993; Roberts and Rosenthal, 1997), whereby two copies of the chain become equal at two *different* times. This can be used to provide bounds on the convergence rate to stationarity of *ergodic average* distributions.

A notion related to but weaker than small set is that of *pseudo-small set*, or *pseudo-minorisation condition.* This notion was introduced formally in Roberts and Rosenthal (1996), though the idea underlying it may have been observed earlier. The idea here is that every pair of points $(x, y) \in C \times C$ has a component in common, but that common component may vary depending on the pair chosen. In symbols, $P^{n_0}(x, \cdot) \geq \epsilon\, \nu_{xy}(\cdot)$ and $P^{n_0}(y, \cdot) \geq \epsilon\, \nu_{xy}(\cdot)$ for all pairs (x, y), though here ν_{xy} depends on the choice of x and y.

The pseudo-minorisation condition does *not* immediately provide notions such as regeneration, the existence of a stationary distribution, coalescence, or a shift-coupling construction. However, the pseudo-minorisation condition is perfectly adequate for ordinary (pairwise) coupling constructions, which always consider just two chains at a time (see Appendix for details). That simple observation provides the basis for the current paper.

This paper is organised as follows. In Section 3, we develop analogues of previous convergence-rate results in terms of pseudo-small sets. In Section 4, we provide specialised versions of these results for countable and absolutely-continuous chains. Section 5 presents a number of different examples of small and pseudo-small sets. In Section 6, we consider the relationship between small and pseudo-small sets, and prove that for ϕ-irreducible, aperiodic Markov chains with countably-generated σ-algebras, all pseudo-small subsets are in fact small (though perhaps with much worse values of n_0 and ϵ). Section 7 considers what can go wrong when assumptions of ϕ-irreducibility and aperiodicity are relaxed. Finally, Section 8 presents a result which applies convergence results for stochastically monotone chains to non-monotone chains which are instead *bounded* by monotone chains.

The paper closes with an Appendix which reviews the traditional pairwise coupling construction based on small sets, and describes how the construction can be modified to be used for pseudo-small sets.

2. Definitions.

Let $\{X_n\}$ be a Markov chain on a state space $\mathcal{X}$, having transition probabilities $P(x,\cdot)$. We begin with a standard definition.

Definition. A set $C \subseteq \mathcal{X}$ is *small* (or, (n_0, ϵ, ν)-small) if there is $n_0 \in \mathbf{N}$, $\epsilon > 0$, and a probability measure ν, such that

$$P^{n_0}(x,\cdot) \;\geq\; \epsilon\nu(\cdot)\,, \qquad x \in C\,. \tag{1}$$

The existence of small sets for ϕ-irreducible Markov chains is proved in Jain and Jameson (1967) and Orey (1971); see Meyn and Tweedie (1993) for a modern exposition. (Recall that a Markov chain is *ϕ-irreducible* if there is a non-zero measure ϕ on $\mathcal{X}$, such that for any subset A with $\phi(A) > 0$, there is positive probability of hitting A starting from any $x \in \mathcal{X}$. See e.g. Meyn and Tweedie, 1993, for this and other basic Markov chain definitions.)

Small sets have many uses. The one most relevant to the current paper is for *pairwise coupling constructions.* Briefly, we can construct two copies of the Markov chain (one started in the arbitrary initial distribution, the other started in the stationary distribution) such that, each time they are both in the small set C, they have probability ϵ of coupling n_0 iterations later. For formal details, see the Appendix.

We now introduce a related, but weaker, notion than that of a small set.

Definition. A set $C \subseteq \mathcal{X}$ is *pseudo-small* (or, $(n_0, \epsilon, \{\nu_{xy}\})$-pseudo-small) if there is $n_0 \in \mathbf{N}$ and $\epsilon > 0$ such that for all $(x, y) \in C \times C$, there is a probability measure ν_{xy} with

$$P^{n_0}(x, \cdot) \wedge P^{n_0}(y, \cdot) \geq \epsilon \nu_{xy}(\cdot) \, . \tag{2}$$

(Note that (2) is shorthand for the two equations $P^{n_0}(x, A) \geq \epsilon \nu_{xy}(A)$ and $P^{n_0}(y, A) \geq \epsilon \nu_{xy}(A)$ for all measurable sets A.)

Obviously, any small set is also pseudo-small, with the same n_0 and ϵ, and with $\nu_{xy} = \nu$ for all pairs (x, y). The primary motivation for pseudo-small sets is that the usual pairwise coupling construction for small sets can be used essentially without change for pseudo-small sets (see Appendix). This means that any result proved using the small set pairwise coupling construction has an immediate analogues for pseudo-small sets, as we now explore.

3. General pseudo-small convergence results.

As described above, for any convergence bounds involving the ordinary (pairwise) coupling construction as in the Appendix, the bound remains true if a minorisation condition is replaced by a corresponding pseudo-minorisation condition. Thus, any coupling-based convergence result which uses a small set can immediately be "transformed" into a corresponding result involving pseudo-small sets.

In particular, because of the purely coupling proof of the following result for small sets (cf. Rosenthal, 1993; Meyn and Tweedie, 1993, Theorem 16.2.4), we have

Proposition 1. *Let $P(x,\cdot)$ be the transition probabilities for a Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. If the entire state space $\mathcal{X}$ is (n_0,ϵ)-pseudo-small, then*

$$\|P^n(x,\cdot)-\pi(\cdot)\| \;\le\; (1-\epsilon)^{\lfloor n/n_0\rfloor}\,, \qquad n\in\mathbf{N}\,,$$

independent of the initial value $x\in\mathcal{X}$.

Here and throughout, $\lfloor r\rfloor$ is the greatest integer not exceeding the real number r, and $\|P^n(x,\cdot)-\pi(\cdot)\| \equiv \sup_{A\subseteq\mathcal{X}} |P^n(x,A)-\pi(A)|$ represents the *total variation distance* between the actual distribution of the Markov chain (after n steps, when initially started at the point x), and the stationary distribution $\pi(\cdot)$.

Similarly, by transforming Proposition 1 of Cowles and Rosenthal (1998; based on Theorem 12 of Rosenthal, 1995a; see also Roberts and Tweedie, 1999), we obtain

Proposition 2. *Let $P(x,\cdot)$ be the transition probabilities for a Markov chain on a state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose for some function $V:\mathcal{X}\to[0,\infty)$, some $\lambda<1$ and $\Lambda<\infty$, some $\epsilon>0$, some positive integers m and k_0, and some $d>\frac{2\Lambda}{1-\lambda}$, we have the drift condition*

$$\mathbf{E}(V(X_m)\mid X_0=x)\le\lambda V(x)+\Lambda\,,\qquad x\in\mathcal{X}\,,$$

and also that the set $\{x\in\mathcal{X}; V(x)\le d\}$ is (mk_0,ϵ)-pseudo-small. Then for any $0<r<1$ and $M>0$, we have

$$\|\mathcal{L}(X_k)\;-\;\pi\|\;\le\;(1-\epsilon)^{\lfloor rk/mk_0\rfloor}\;+\;C_0\,(\alpha A)^{-1}\left(\alpha^{-(1-rk_0)}A^r\right)^{\lfloor k/m\rfloor}\,,\qquad k\in\mathbf{N}\,,$$

where

$$\alpha^{-1}\;=\;\frac{1+2M\Lambda+M\lambda d}{1+Md};\quad A\;=\;1+2(\lambda Md+M\Lambda)\,;\quad C_0\;=\;\left(1+\frac{M\Lambda}{1-\lambda}+M\mathbf{E}(V(X_0))\right)\,.$$

If furthermore it is known that $V(x)\ge1$ for all $x\in\mathcal{X}$, then it suffices that $d>\frac{2\Lambda}{1-\lambda}-1$, and these values may be improved slightly to

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

$$C_0 \;=\; \frac{M}{2}\left(\frac{\Lambda}{1-\lambda} + \mathbf{E}(V(X_0))\right) + (1-M)\,.$$

For example, taking $M = 1$ in the $V(x) \geq 1$ case, we obtain the simplification

$$\alpha^{-1} \;=\; \lambda + \frac{2\Lambda}{d+1}\,; \quad A \;=\; \lambda d + \Lambda\,; \quad C_0 \;=\; \frac{1}{2}\left(\frac{\Lambda}{1-\lambda} + \mathbf{E}(V(X_0))\right)\,.$$

Recall now that a Markov chain is *stochastically monotone* with respect to an ordering $\preceq$ on $\mathcal{X}$ if, for all fixed $\mathbf{z}$, we have that $\mathbf{P}(\mathbf{X}_1 \preceq \mathbf{z} | \mathbf{X}_0 = \mathbf{x}_1) \geq \mathbf{P}(\mathbf{X}_1 \preceq \mathbf{z} | \mathbf{X}_0 = \mathbf{x}_2)$ whenever $\mathbf{x}_1 \preceq \mathbf{x}_2$. For such chains, if $X_0 \succeq Y_0$, then it is possible (cf. Kamae et al., 1977; Lindvall, 1992, p. 134) to simultaneously construct $\{X_k\}$ and $\{Y_k\}$ so that $X_k \succeq Y_k$ for all k. Intuitively, for stochastically monotone chains and small sets of the form $C = \{x \in \mathcal{X};\ x \preceq c\}$, it is easier to prove convergence bounds, since if $X_k \succeq Y_k$ for all k, then $(X_k, Y_k) \in C \times C$ whenever $X_k \in C$.

Therefore, transforming Theorem 2.2 of Roberts and Tweedie (2000; which builds on the work of Lund and Tweedie, 1996 and Lund, Meyn, and Tweedie, 1996), we obtain the following. We write $\nu(V)$ for the expected value of V with respect to ν, and write $\mathbf{E}_x^\pi(V)$ for the expected value of V with respect to the *stochastic majorant* (with respect to $\preceq$) of a point mass at x and the stationary distribution $\pi(\cdot)$, defined by

$$\mathbf{E}_x^\pi(\mathbf{1}_{(-\infty,y]}) \;=\; \min\left[\mathbf{1}_{(-\infty,y]}(x),\ \pi\left((-\infty,y]\right)\right]\,.$$

Proposition 3. *Let $P(x,\cdot)$ be the transition probabilities for a stochastically monotone Markov chain on a totally ordered state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Let $C \subseteq \mathcal{X}$ be $(1,\epsilon)$-pseudo-small, where $C = \{x \in \mathcal{X};\ x \preceq c\}$ for some fixed $c \in \mathcal{X}$, and let $V : \mathcal{X} \to [1,\infty)$ be such that*

$$\mathbf{E}(V(X_1) \mid X_0 = x) \;\leq\; \lambda V(x) + b\mathbf{1}_C(x)\,, \qquad x \in \mathcal{X}$$

for some $\lambda < 1$ and $0 \leq b < \infty$. Then for $n > \log \mathbf{E}_x^\pi(V) \,/\, \log(\lambda^{-1})$, we have

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; K(n + \eta - \xi)\rho^n\,, \qquad n \in \mathbf{N}\,.$$

Here

$$K = \frac{e\epsilon(1-\epsilon)^{-\xi/\eta}}{\eta},$$

$$\xi = \frac{\log \mathbf{E}_x^\pi(V)}{\log(\lambda^{-1})}, \qquad \eta = \frac{\log\left(\frac{\lambda s + b - \epsilon}{\lambda(1-\epsilon)}\right)}{\log(\lambda^{-1})},$$

$s = \sup\{V(z);\ z \preceq x\}$, *and* $\rho = (1-\epsilon)^{\eta^{-1}}$.

In summary, any total variation distance convergence result based on pairwise coupling, which makes use of small sets, can immediately be transformed into a corresponding result using the weaker condition of pseudo-small sets.

Remark. Of course, there are many Markov chains which do not converge at all in total variation distance, so that neither small nor pseudo-small sets can be used in the above manner. Rather, other distance measures and other convergence techniques must be used; see e.g. Su (1998).

4. Countable pseudo-small state spaces.

For Markov chains on countable state spaces, certain more explicit formulae are available. We begin with the standard

Proposition 4. *Consider a Markov chain on a countable state space $\mathcal{X}$, and let $C \subseteq \mathcal{X}$ be any subset. Then for any $n_0 \in \mathbf{N}$, the subset C is (n_0, ϵ_{n_0})-small with*

$$\epsilon_{n_0} \;=\; \sum_{y\in\mathcal{X}} \inf_{x\in C} P^{n_0}(x,\{y\}) .$$

(Of course, we may have $\epsilon_{n_0} = 0$ for all $n_0 \in \mathbf{N}$.) Furthermore, C is not (n_0, ϵ')-small for any $\epsilon' > \epsilon_{n_0}$.

Proof. Define the probability measure $\nu(\cdot)$ by

$$\nu(\{z\}) \;=\; \frac{\inf_{x\in C} P^{n_0}(x,\{z\})}{\sum_{y\in\mathcal{X}} \inf_{x\in C} P^{n_0}(x,\{y\})} .$$

Then it is verified that $P^{n_0}(x,\cdot) \geq \epsilon_{n_0}\,\nu(\cdot)$ for all $x \in C$, with ϵ_{n_0} as above.

Furthermore, if there were some $\epsilon' > \epsilon_{n_0}$ and some other probability measure $\nu'(\cdot)$ such that $P^{n_0}(x,\cdot) \geq \epsilon'\,\nu'(\cdot)$ for all $x \in C$, then we could find $z \in \mathcal{X}$ with $\epsilon'\,\nu'(\{z\}) > \epsilon_{n_0}\,\nu(\{z\})$. But $\epsilon_{n_0}\,\nu(\{z\}) = \inf_{x\in C} P^{n_0}(x,\{z\})$, so this gives a contradiction. ∎

Using the concept of pseudo-small sets, we can improve the above result to

Proposition 5. *Consider a Markov chain on a countable state space $\mathcal{X}$, and let $C \subseteq \mathcal{X}$ be any subset. Then for any $n_0 \in \mathbf{N}$, the subset C is (n_0, ϵ_{n_0})-pseudo-small with*

$$\epsilon_{n_0} \;=\; \inf_{x,y\in C} \sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})]\,.$$

(Of course, we may have $\epsilon_{n_0} = 0$ for all $n_0 \in \mathbf{N}$.) Furthermore, C is not (n_0, ϵ')-pseudo-small for any $\epsilon' > \epsilon_{n_0}$.

Proof. For $x, y \in C$, define the probability measure $\nu_{xy}(\cdot)$ by

$$\nu_{xy}(\{w\}) \;=\; \frac{\min[P^{n_0}(x,\{w\}), P^{n_0}(y,\{w\})]}{\sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})]}\,.$$

Then it is verified that $P^{n_0}(x,\cdot) \geq \epsilon_{n_0}\,\nu_{xy}(\cdot)$ and $P^{n_0}(y,\cdot) \geq \epsilon_{n_0}\,\nu_{xy}(\cdot)$ for all $x, y \in C$, with ϵ_{n_0} as above.

Furthermore, if there were some $\epsilon' > \epsilon_{n_0}$ and some other probability measures $\nu'_{xy}(\cdot)$ such that $P^{n_0}(x,\cdot) \geq \epsilon'\,\nu'_{xy}(\cdot)$ and $P^{n_0}(y,\cdot) \geq \epsilon'\,\nu'_{xy}(\cdot)$ for all $x, y \in C$, then we could find $x, y \in C$ with $\epsilon' > \sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})]$. We could then find $w \in \mathcal{X}$ with $\nu'_{xy}(\{w\}) \geq \nu_{xy}(\{w\})$. It follows that

$$\epsilon'\,\nu'_{xy}(\{w\}) \geq \epsilon'\,\nu_{xy}(\{w\}) > \sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})]\,\nu_{xy}(\{w\})$$

$$= \min[P^{n_0}(x,\{w\}), P^{n_0}(y,\{w\})]\,,$$

giving a contradiction. ∎

Combining the above proposition with Proposition 1, we obtain the following corollary (which also follows from Dobrushin, 1956, pp. 71 and 334).

Corollary 6. *Consider a Markov chain on a countable state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Then for any $n_0 \in \mathbf{N}$, and any $x \in \mathcal{X}$, we have*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\leq\; (1-\epsilon_{n_0})^{\lfloor n/n_0 \rfloor}\,, \qquad n \in \mathbf{N}\,,$$

where

$$\epsilon_{n_0} \;=\; \inf_{x,y\in\mathcal{X}} \sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})]\,.$$

(Again, we may have $\epsilon_{n_0} = 0$ for all $n_0 \in \mathbf{N}$.)

As a special case of Corollary 6, we obtain an alternate proof of a special case of a result of P. Bickel.

Corollary 7. *(Bickel, 1999) Consider a Markov chain on a finite state space $\mathcal{X}$, with $|\mathcal{X}| = k$. For $n_0 \in \mathbf{N}$ and $x \in \mathcal{X}$, let $h_{n_0}(x) = \#\{y \in \mathcal{X}\,;\; P^{n_0}(x,\{y\}) > 0\}$. Suppose that for some $n_0 \in \mathbf{N}$, we have $h_{n_0}(x) > \frac{k}{2}$ for all $x \in \mathcal{X}$. Then the Markov chain is uniformly ergodic.*

Proof. Since $h_{n_0}(x) > \frac{k}{2}$ for all $x \in \mathcal{X}$, it follows easily that

$$\sum_{z\in\mathcal{X}} \min[P^{n_0}(x,\{z\}), P^{n_0}(y,\{z\})] > 0 \qquad \text{for all } x,y \in \mathcal{X}\,.$$

Hence, with ϵ_{n_0} as in the previous corollary, we have $\epsilon_{n_0} > 0$. The result now follows from the previous corollary. ∎

Remark. Of course, Corollary 7 can also be proved by showing that the chain is irreducible and aperiodic, and then using standard theory. In fact, Bickel (1999) proves more than Corollary 7, showing that the chain's convergence rate can be controlled by $\max_j \sum_{i=1}^{|\mathcal{X}|} |P^{n_0}(i,\{j\}) - \mathrm{median}_k P^{n_0}(k,\{j\})|$.

By analogy to Corollary 6, we obtain a similar result for continuous spaces when the transition distributions all have a density. (We omit the proof.)

Proposition 8. *Consider a Markov chain on a general state space $\mathcal{X}$, with stationary distribution $\pi(\cdot)$. Suppose that for some $n_0 \in \mathbf{N}$, we have that $P^{n_0}(x, dy) = f_{n_0,x}(y) m(dy)$ for all $x \in \mathcal{X}$, for some fixed σ-finite measure $m(\cdot)$ and some density functions $f_{n_0,x}$. Then*

$$\|P^n(x,\cdot) - \pi(\cdot)\| \;\le\; (1-\epsilon_{n_0})^{\lfloor n/n_0 \rfloor}\,, \qquad n \in \mathbf{N}\,,$$

where

$$\epsilon_{n_0} \;=\; \inf_{x,y\in\mathcal{X}} \int \min[f_{n_0,x}(z),\; f_{n_0,y}(z)]\; m(dz)\,.$$

5. Examples.

It is sometimes straightforward to construct pseudo-small sets, but far more difficult to directly construct small sets. Alternatively, sometimes it is easy to construct both, but the pseudo-small construction gives much better convergence bounds. This section presents a number of different examples to illustrate this.

Example #1: A simple illustration.

For a simple example, consider the (non-reversible) Markov chain on $\mathcal{X} = \{1, 2, 3\}$ with stationary distribution the uniform distribution $U(\cdot)$ on $\mathcal{X}$, and with transition probabilities given by

$$P \;=\; \begin{pmatrix} 1/2 & 1/2 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}\,.$$

Then the entire space $\mathcal{X}$ is $(1, \frac{1}{2})$-pseudo-small (with $\nu_{xy} = \delta_{z(x,y)}$ for appropriate points $z(x,y)$). Hence, pseudo-smallness gives a convergence bound of

$$\|P^n(x,\cdot) - U(\cdot)\| \;\le\; 0.5^n\,, \qquad n \in \mathbf{N}\,,$$

On the other hand, $\mathcal{X}$ is *not* $(1,\epsilon)$-small for any $\epsilon > 0$. Instead, the chain is only $(2, \frac{3}{4})$-small (with $\nu = U$). Hence, smallness gives a convergence bound of

$$\|P^n(x,\cdot) - U(\cdot)\| \;\le\; 0.25^{\lfloor n/2 \rfloor}\,, \qquad n \in \mathbf{N}\,.$$

Of course, this second bound is virtually as good as the first (and indeed, both are essentially optimal as can be seen by computing the eigenvalues of P). However, the second bound required computing with 2-step transition probabilities, not just 1-step transition probabilities, and that may be infeasible on more complicated examples.

Example #2: 1-pseudo-small but only 3-small.

For another example, define a transition matrix P on the state space $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ by

$$3P \;=\; \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Here the entire state space $\mathcal{X}$ is $(1, \frac{1}{3})$-pseudo-small by inspection of Figure 1.

Figure 1. A graphical illustration of the possible transitions for Example #2, of probability 1/3 each. (Not shown are the 1/3 probabilities of staying in the *same* state.) Inspection reveals that from any 2 states, there is some state which can be reached in 1 step from each. However this Markov chain is neither 1-small nor 2-small.

However, by inspection $\mathcal{X}$ is *not* n_0-small for $n_0 = 1$ or 2. For $n_0 = 3$, $\mathcal{X}$ is $(3, \frac{24}{27})$-

small, by taking $\nu(\cdot)$ to be uniform and noting that $P^3(x,\{y\}) \geq \frac{4}{27}$ for all $x, y \in \mathcal{X}$.

Remark. It is no coincidence that this example (which requires tripling n_0 to move from pseudo-smallness to smallness) is not reversible, while the previous example (which required just *doubling* n_0 to move from pseudo-smallness to smallness) is reversible; see Proposition 13 below.

Example #3: Random-scan Gibbs sampler.

Consider the Random Scan Gibbs sampler for an everywhere-positive probability distribution $\pi(\cdot)$ on the state space $\mathcal{X} = \{0,1\}^d$, i.e. the vertices of a d-dimensional hypercube (so that $|\mathcal{X}| = 2^d$). Specifically, given $X_k = (x_1, \ldots, x_d)$, this Markov chain chooses $X_{k+1} = (z_1, \ldots, z_d)$ by (a) choosing I_{k+1} uniform on the index set $\{1, 2, \ldots, d\}$; (b) setting $z_i = x_i$ for $i \neq I_{k+1}$; and (c) for $i = I_{k+1}$, choosing z_i to be 0 or 1 conditionally independently, according to the probabilities

$$\mathbf{P}(z_i = 0) \;=\; \frac{\pi(\{(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d)\})}{\pi(\{(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d)\}) + \pi(\{(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)\})} \tag{3}$$

and

$$\mathbf{P}(z_i = 1) \;=\; \frac{\pi(\{(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)\})}{\pi(\{(x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d)\}) + \pi(\{(x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d)\})}\,.$$

We let $M = \min_{x \in \mathcal{X}} \pi(\{x\}) / \max_{x \in \mathcal{X}} \pi(\{x\})$, so that $0 < M \leq 1$, and so that $\frac{M}{1+M} \leq \mathbf{P}(z_i = 0) \leq \frac{1}{1+M}$ in (3).

For this Markov chain, the entire state space $\mathcal{X}$ is both small and pseudo-small, but the constants are different.

Regarding smallness, $\mathcal{X}$ is clearly not n_0-small for any $n_0 < d$, since it is impossible for the chain to update all its components and hence move to the opposite corner of the hypercube in this time. For $n_0 = d$, the chain is $(d, \epsilon, \pi(\cdot))$-small, where $M^d\, d!\, d^{-d} \leq \epsilon \leq M^{-d}\, d!\, d^{-d}$. To see this, note that there is probability $d!\, d^{-d}$ that $I_1, \ldots, I_d$ are all distinct, and given that they are, the chance of ending up at a given site $x \in \mathcal{X}$ after d steps is between $M^d \pi(\{x\})$ and $M^{-d} \pi(\{x\})$. In fact, if $\pi = \pi_1 \times \ldots \times \pi_d$ is of product measure form – e.g. uniform – then the coordinates move independently, so the factors of M are not required, and we have $\epsilon = d!\, d^{-d}$ exactly.

Now consider pseudo-smallness, which involves minorising from just two states in $\mathcal{X}$ at a time, x and y say. The worst case is clearly when x and y are at opposite corners of $\mathcal{X}$. We shall assume for formulaic simplicity that d is even. We shall compute the $(d/2)$-pseudo-smallness constant by imagining running chains from x and y simultaneously, coupled so that when one chain is updating site i, the other chain is updating site $d+1-i$. Consider making $d/2$ updates. In order for all sites to have been updated by one chain or other after $d/2$ updates, it is necessary for the chain started from x (say) to not repeat visiting any one site i or its "complement" $d+1-i$. This happens with probability $(d/2)!\, 2^{d/2} \,/\, d^{d/2}$. In each of the first d updates ($d/2$ from each chain), there is a probability at least $M/(1+M)$ of matching the other chain in that coordinate. It follows that the chain is $(d/2, \tilde{\epsilon})$-pseudo-small where

$$\tilde{\epsilon} \;\geq\; \frac{M^d}{(1+M)^d}\, \frac{(d/2)!\, 2^{d/2}}{d^{d/2}} \;.$$

In the case where π is uniform (so $M = 1$), Stirling's formula gives (for large d) that $\epsilon \approx (2\pi d)^{1/2} e^{-d}$ for the d-smallness constant, while $\tilde{\epsilon} \approx (\pi d)^{1/2} e^{-d/2}$ for the $(d/2)$-pseudo-smallness constant. This indicates that $\tilde{\epsilon} > \epsilon$ for sufficiently large d, even though it requires fewer steps ($d/2$ versus d) to achieve. We conclude that, for this example, pseudo-smallness gives a clear speed-up in terms of convergence bounds, as compared to smallness convergence bounds.

On the other hand, for much larger values of n_0, we can obtain expressions for ϵ and $\tilde{\epsilon}$ in terms of Sterling numbers of the second kind (see for example Sloane, 2000). Here we merely make a remark applicable again to the uniform case. Note that for the uniform case, the Markov chain is equivalent to a random walk on the set $\mathcal{X}$ regarded as an additive (abelian) group. For such random walks on groups, very precise results are known about the eigenvalues and the convergence rate, cf. Diaconis (1988).

Indeed, by a simple expression for the probability that all sites have been updated after n_0 steps for large n_0, we obtain that X is $(n_0, 1 - d((d-1)/d)^{n_0})$-small. Hence, an upper bound on the Markov chain's geometric rate of convergence is $1 - d^{-1}$. However this is the actual geometric rate of convergence, computed as in Diaconis (1988) (see also Rosenthal, 1995b). Thus, the approximations used here, although fairly simple, are actually giving asymptotically the correct rate of convergence. Moreover, the pseudo-small bounds are even better, though the size of their improvement over the small set bounds diminishes as $n_0 \to \infty$.

Remark. We note that the results from Diaconis (1988) cannot be applied if $\pi(\cdot)$ is not uniform.

Example #4: Convergence of multi-dimensional diffusions.

Another example is provided by the diffusion convergence bounding of Roberts and Rosenthal (1996). Consider a diffusion process $\{\mathbf{X}_s\}_{s\geq 0}$ defined on $\mathbf{R}^k$ by $d\mathbf{X}_s = \mu(\mathbf{X}_s)ds + d\mathbf{B}_s$, for some function $\mu : \mathbf{R}^k \to \mathbf{R}$, where $\{\mathbf{B}_s\}_{s\geq 0}$ is standard k-dimensional Brownian motion. In dimensions larger than 1, it is very difficult to directly construct a coupling for such a diffusion that works for all points in a subset simultaneously. However, to construct a coupling for just two points at a time is relatively straightforward.

Based on that fact, Roberts and Rosenthal (1996) proved the following result. To state it, for $\mathbf{c}, \mathbf{d} \in \mathbf{R}^k$, we say a set $S \subseteq \mathbf{R}^k$ is a "$[\mathbf{c}, \mathbf{d}]$-medium set" if $c_i \leq \mu_i(x) \leq d_i$ for all $x \in S$, for $1 \leq i \leq k$.

Theorem 9. *Let $\{\mathbf{X}_s\}$ be a multi-dimensional diffusion process defined by $d\mathbf{X}_s = \mu(\mathbf{X}_s)ds + d\mathbf{B}_s$. Suppose C is contained in $\prod_{i=1}^{k}[\alpha_i, \beta_i]$, and let $D = \sup_{x,y\in C}\|x-y\|_2$ be the L^2 diameter of C. Let $S = \prod_{i=1}^{k}[a_i, b_i]$, where $a_i < \alpha_i < \beta_i < b_i$ for each i, and suppose S is a $[\mathbf{c}, \mathbf{d}]$-medium set. Set $L = \|\mathbf{d}-\mathbf{c}\|_2 \equiv \left(\sum_{i=1}^{k}(d_i - c_i)^2\right)^{1/2}$. Then for any $t > 0$, there exists an $\epsilon > 0$ such that C is (t, ϵ)-pseudo-small. Moreover, given any $t_0 > 0$, for all $t \geq t_0$, we have that C is (t, ϵ)-pseudo-small where*

$$\begin{aligned}\epsilon = {} & \Phi\left(\frac{-D - t_0 L}{\sqrt{4t_0}}\right) + e^{-DL/2}\Phi\left(\frac{t_0 L - D}{\sqrt{4t_0}}\right) \\ & - 2\sum_{i=1}^{k}\Phi\left(\frac{-(\alpha_i - a_i) - t_0 c_i}{\sqrt{t_0}}\right) - 2\sum_{i=1}^{k} e^{-2(\alpha_i - a_i)c_i}\Phi\left(\frac{t_0 c_i - (\alpha_i - a_i)}{\sqrt{t_0}}\right) \\ & - 2\sum_{i=1}^{k}\Phi\left(\frac{-(b_i - \beta_i) + t_0 d_i}{\sqrt{t_0}}\right) - 2\sum_{i=1}^{k} e^{2(b_i - \beta_i)d_i}\Phi\left(\frac{-t_0 d_i - (b_i - \beta_i)}{\sqrt{t_0}}\right).\end{aligned}$$

Here $\Phi(z) = \int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-s^2/2}ds$ is the cumulative distribution function of a standard normal distribution.

This result proves, in particular, that the subsets C described above are pseudo-small sets for the multi-dimensional diffusion considered. (It also gives a quantitative bound on the value of ϵ, though this bound may not be positive for all values of the parameters.) In light of Theorem 12 below, we see now that the pseudo-small sets constructed in Theorem 9 are in fact small. (This fact is not otherwise obvious, though it can also be proved using general Markov chain theory.) Indeed, aperiodicity and ϕ-irreducibility are both clearly satisfied. Furthermore, since the above result is stated for discrete time jumps $t > 0$, we see that Theorem 12 applies directly and there is no concern over the fact that the discrete-time process arises from an underlying process which is in continuous time.

On the other hand, we have no control over the smallness constants (as opposed to pseudo-small constants) for these sets C. Thus, to obtain quantitative bounds on the convergence rates of these diffusions, it is essential to use the pseudo-small construction. For details, see Roberts and Rosenthal (1996).

Example #5: One-dimensional monotone shifts.

The previous two examples were highly multi-dimensional, and indeed it appears that pseudo-smallness may be most advantageous in high dimensions.

At the "opposite extreme", consider a one-dimensional example, where the transition densities are just monotone shifts of each other. Specifically, let $\mathcal{X} \subseteq \mathbf{R}$, let $f : \mathbf{R} \to [0, \infty)$ be unimodal and have Lebesgue integral 1, let $\phi : \mathcal{X} \to \mathcal{X}$ be monotone (i.e. $\phi(x) \leq \phi(y)$ whenever $x \leq y$), and define Markov chain transition probabilities on $\mathcal{X}$ by

$$P(x, dy) \;=\; f(y - \phi(x))\, dy\,.$$

Then we have

Proposition 10. *For a one-dimensional monotone shift Markov chain as above, if a subset $C \subseteq \mathcal{X}$ is $(1, \epsilon)$-pseudo-small, then C is $(1, \epsilon)$-small with the same value of ϵ. That is, one-step pseudo-smallness provides absolutely no improvement over one-step smallness in this case.*

Proof. Let m be the mode of f, so that $m+\phi(x)$ is the mode of $P(x,\cdot)$. Let $L = \inf_{x\in C}(m+\phi(x))$ and $R = \inf_{x\in C}(m+\phi(x))$ be the left and right extremes, respectively, of the modes from C.

Now, by the unimodality of f, we have that for $a \leq x \leq b$,

$$f(y-\phi(a)) \leq f(y-\phi(x))\,, \qquad y \leq m+\phi(x)\,,$$

and

$$f(y-\phi(b)) \leq f(y-\phi(x))\,, \qquad y \geq m+\phi(x)\,.$$

Hence,

$$\inf_{a\leq x\leq b} f(y-\phi(x)) \;=\; \min\left[f(y-\phi(a)),\; f(y-\phi(b))\right]\,.$$

It follows from this that

$$\int \inf_{\substack{x\in C\\ a\leq x\leq b}} f(y-\phi(x))\,dy \;=\; \int \min\left[f(y-\phi(a)),\; f(y-\phi(b))\right]\,dy\,. \tag{4}$$

Now, in (4), the left-hand side is the largest ϵ such that the set $C\cap[a,b]$ is $(1,\epsilon)$-small, while the right-hand side is the largest ϵ such that the set $\{a,b\}$ is $(1,\epsilon)$-pseudo-small. (Compare Proposition 8.) Hence, if C has a minimal state a and maximal state b, then we can choose these values of a and b in (4), to see that the 1-pseudo-small constant for C is no larger than the 1-small constant, thus giving the result.

If C does not have minimal and/or maximal states, then the result still follows from (4), by instead choosing sequences $\{a_n\}\subseteq C$ and/or $\{b_n\}\subseteq C$, with $a_n \searrow \inf(C)$ and/or $b_n \nearrow \sup(C)$. For such a choice, the left-hand side of (4) converges to the 1-small constant for C, while the right-hand side converges to an upper bound on the 1-pseudo-small constant. It again follows that the 1-pseudo-small constant for C is no larger than the 1-small constant. ∎

Remark. One might think that Proposition 10 could be generalised to *any* stochastically monotone chain (not just monotone shifts), but this is false. For a simple example, consider simple symmetric random walk on $\mathcal{X} = \{1, 2, 3\}$ with holding boundaries, so that

$$2P = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} .$$

This Markov chain is clearly stochastically monotone, and is even "realisably monotone" in the sense of Fill and Machida (1999). Furthermore, the subset $\mathcal{X}$ is clearly $(1, \frac{1}{2})$-pseudo-small. However, $\mathcal{X}$ is clearly not $(1, \epsilon)$-small for any $\epsilon > 0$. The problem here is that, while the chain is realisably monotone, there is no way to realise monotonicity *and* realise the coupling implied by being $(1, \frac{1}{2})$-pseudo-small (i.e. have the chains started at 1 and 3 become equal with probability $\frac{1}{2}$) *simultaneously.* Indeed, as soon as we ensure that those two chains become equal with probability $\frac{1}{2}$, we can no longer maintain the monotonic structure, and hence can conclude nothing about smallness.

6. Relation of pseudo-small to small.

As discussed in the introduction, small sets have many uses. One is the pairwise coupling construction described in the Appendix, where pseudo-small sets do just as well. Others include regenerations and coalescence, where pseudo-small sets *cannot* be used in place of small sets.

Because of this complex relationship, we now explore further the implications of pseudo-small sets for smallness. We shall show (Theorem 12 below) that, for a ϕ-irreducible and aperiodic Markov chain, pseudo-smallness does indeed imply smallness (though perhaps with much worse values for n_0 and ϵ).

We begin with a lemma. (Recall that a state space is *countably generated* if its σ-algebra is the smallest σ-algebra containing some fixed countable collection of measurable sets; this is a very weak condition which is satisfied by virtually all examples of interest.)

Lemma 11. *Let S be an (n_0, ϵ)-pseudo-small subset for a ϕ-irreducible, aperiodic Markov chain on a countably generated state space $\mathcal{X}$. Then there is a small set $C \subseteq \mathcal{X}$ and $\delta > 0$ such that $P^{n_0}(x, C) \geq \delta$ for all $x \in S$.*

Proof. Choose any fixed point $x_0 \in \mathcal{X}$.

By Meyn and Tweedie (1993, Proposition 5.2.4 (ii)), we can write $\mathcal{X}$ as a countable union of small sets. But for an aperiodic, ϕ-irreducible chain, a finite union of small sets is petite and hence small (cf. Meyn and Tweedie, 1993, Proposition 5.5.5 and Theorem 5.5.7).

We conclude that we can find small sets which are arbitrarily large. In particular, there is a small set $C \subseteq \mathcal{X}$ with $P^{n_0}(x_0, C) \geq 1 - \frac{\epsilon}{2}$.

Now, let $x \in S$. Then there is $\nu = \nu_{x_0,x}$ with $P^{n_0}(x_0, \cdot) \geq \epsilon\nu(\cdot)$ and $P^{n_0}(x, \cdot) \geq \epsilon\nu(\cdot)$. This implies that we can write $P^{n_0}(x_0, \cdot) = (1-\epsilon)R(\cdot) + \epsilon\nu(\cdot)$, where $R(\cdot) = \frac{1}{1-\epsilon}(P^{n_0}(x_0, \cdot) - \epsilon\nu(\cdot))$ is also a probability measure. In particular, $(1-\epsilon)R(C) + \epsilon\nu(C) = P^{n_0}(x_0, C) \geq 1 - \frac{\epsilon}{2}$, whence

$$\epsilon\nu(C) \geq (1 - \frac{\epsilon}{2}) - (1-\epsilon)R(C) \geq (1 - \frac{\epsilon}{2}) - (1 - \epsilon) = \frac{\epsilon}{2}\,.$$

Hence, $\nu(C) \geq \frac{1}{2}$. By the pseudo-minorisation condition, it then follows that $P^{n_0}(x, C) \geq \epsilon\,\nu(C) \geq \frac{\epsilon}{2}$, for $x \in S$. The result therefore follows with $\delta = \epsilon/2$. ■

We can now prove

Theorem 12. *Let S be a pseudo-small subset for a ϕ-irreducible, aperiodic Markov chain on a countably generated state space $\mathcal{X}$. Then S is also small (though not necessarily with the same values of n_0 and ϵ).*

Proof. Let S be (n_0, ϵ)-pseudo-small.

By Lemma 11, there is an (n_1, ϵ_1, ν_1)-small set C and $\delta > 0$ with $P^{n_0}(x, C) \geq \delta$ for all $x \in S$.

But this implies that $P^{n_0+n_1}(x, \cdot) \geq \delta\epsilon_1\nu_1(\cdot)$ for all $x \in S$. Hence, S is $(n_0 + n_1, \delta\epsilon_1, \nu_1)$-small. ■

In the special case of finite *reversible* chains, we can make the connection between pseudo-small and small even more explicit, as follows.

Proposition 13. *Let C be n_0-pseudo-small for an irreducible, reversible Markov chain on a finite state space. Then C is $2n_0$-small.*

Proof. Recall that reversibility means that $P^{n_0}(x, \{y\})\pi(\{x\}) = P^{n_0}(y, \{x\})\pi(\{y\})$ for all states x and y, where $\pi(\cdot)$ is a stationary distribution. By irreducibility, $\pi(\cdot)$ is unique, and furthermore $\pi(\{x\}) > 0$ for all x. We conclude that

$$P^{n_0}(x, \{y\}) > 0 \qquad \text{if and only if} \qquad P^{n_0}(y, \{x\}) > 0\,. \tag{5}$$

Now, let $x, y \in C$. By pseudo-smallness, there is $\epsilon > 0$ and $\nu_{xy}(\cdot)$ with $P^{n_0}(x, \cdot) \geq \epsilon\nu_{xy}(\cdot)$ and $P^{n_0}(y, \cdot) \geq \epsilon\nu_{xy}(\cdot)$. Choose any state z with $\nu_{xy}(\{z\}) > 0$. Then $P^{n_0}(x, \{z\}) > 0$ and $P^{n_0}(y, \{z\}) > 0$. From (5), this implies that $P^{n_0}(z, \{y\}) > 0$, whence $P^{2n_0}(x, \{y\}) \geq P^{n_0}(x, \{z\})\, P^{n_0}(z, \{y\}) > 0$.

We conclude that $P^{2n_0}(u, \{v\}) > 0$ for all $u, v \in C$. Since C is finite, it now follows (cf. Proposition 4) that C is $2n_0$-small. ■

Of course, Proposition 13 does not hold for non-reversible chains, cf. Example #2 above.

More generally, we have the following.

Theorem 14. *Let C be $(n_0, \epsilon, \{\nu_{xy}\})$-pseudo-small for a ϕ-irreducible, reversible (with respect to $\pi(\cdot)$) Markov chain on a general state space $\mathcal{X}$. Then C is $(2n_0,\, \epsilon^2 K\pi(C),\, \pi\, \big|_C)$-small, where $\left[\pi\, \big|_C\right](dw) = \pi(dw)\mathbf{1}_C(w)\,/\,\pi(C)$, and where*

$$K \;=\; \inf_{x,y\in C} \mathbf{E}_{\nu_{xy}}\left[\frac{d\nu_{xy}}{d\pi}\right]\,. \tag{6}$$

(Of course, we may have $K = 0$.)

Proof. We compute using reversibility and pseudo-smallness that, for $x, y \in C$,

$$\begin{aligned} P^{2n_0}(x,dy) &= \int_{z\in\mathcal{X}} P^{n_0}(x,dz)P^{n_0}(z,dy) \\ &= \int_{z\in\mathcal{X}} P^{n_0}(x,dz)P^{n_0}(y,dz)\frac{\pi(dy)}{\pi(dz)} \\ &\geq \epsilon^2 \int_{z\in\mathcal{X}} \nu_{xy}(dz)\nu_{xy}(dz)\frac{\pi(dy)}{\pi(dz)} \\ &= \epsilon^2\pi(dy)\int_{z\in\mathcal{X}} \nu_{xy}(dz)\frac{\nu_{xy}(dz)}{\pi(dz)} \\ &= \epsilon^2\pi(dy)\mathbf{E}_{\nu_{xy}}(\frac{d\nu_{xy}}{d\pi})\,. \end{aligned}$$

Hence, for $x \in C$, we have

$$P^{2n_0}(x,dy) \;\geq\; \epsilon^2 K\pi(dy)\mathbf{1}_C(y) \;=\; \epsilon^2 K\pi(C)\left[\pi\,\big|_C\right](dy)\,,$$

with K as in (6). The result follows. ∎

7. Issues of irreducibility and aperiodicity.

Theorem 12 and other results above relied on having a ϕ-irreducible, aperiodic chain. We now consider to what extent these results are affected when the assumptions of ϕ-irreducibility and aperiodicity are dropped.

We begin with some simple counter-examples.

Proposition 15. *For a Markov chain which is reducible or periodic, a subset C may be pseudo-small without being small.*

Proof. Consider the matrix

$$P \;=\; \begin{pmatrix} 0 & 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & 0 & 0 & .5 & .5 \\ 0 & 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}\,.$$

Then the chain is reducible (but aperiodic), and $\{1,2,3\}$ is pseudo-small but is not small.

If instead we take

$$P \;=\; \begin{pmatrix} 0 & 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & 0 & 0 & .5 & .5 \\ 0 & 0 & 0 & .5 & .5 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

then the chain is periodic (but ϕ-irreducible where $\phi(\cdot)$ is, say, a point-mass at the point 6), and $\{1,2,3\}$ is pseudo-small but is not small. ∎

If the *entire* state space is pseudo-small, then some positive results can be obtained. We have

Lemma 16. *If the entire state space is pseudo-small, then the chain is aperiodic.*

Proof. If the chain were periodic, then we would have disjoint non-empty subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ (for some $d \geq 2$) with $P(x_i, \mathcal{X}_{i+1 \bmod d}) = 1$ for $x_i \in \mathcal{X}_i$. It follows that $P^{n_0}(x_1, \cdot) \wedge P^{n_0}(x_2, \cdot) = 0$ for $x_i \in \mathcal{X}_i$, for any $n_0 \in \mathbf{N}$. This contradicts pseudo-smallness. ∎

Recall now that a point $j \in \mathcal{X}$ on a countable state space is *recurrent* if $\mathbf{P}_j(X_n = j$ for some $n \geq 1) = 1$. Then we also have

Lemma 17. *If the entire state space is pseudo-small and countable, with at least one recurrent point, then the chain is ϕ-irreducible.*

Proof. Let $j \in \mathcal{X}$ be recurrent. If the chain were *not* ϕ-irreducible, then we must have $P_i(\tau_j < \infty) = 0$ for some $i \in \mathcal{X}$ (otherwise we could take $\phi = \delta_j$).

Suppose that for some $k \in \mathcal{X}$ and $n_0 \in \mathbf{N}$ we had $P^{n_0}(i, \{k\}) > 0$ and $P^{n_0}(j, \{k\}) > 0$. Since j is recurrent, this implies $P(k, \{j\}) = 1$. But since $P_i(\tau_j < \infty) = 0$, this implies $P(k, \{j\}) = 0$.

We conclude that there is no $k \in \mathcal{X}$ and $n_0 \in \mathbf{N}$ with $P^{n_0}(i, \{k\}) > 0$ and $P^{n_0}(j, \{k\}) > 0$. This means that $P^{n_0}(i, \cdot) \wedge P^{n_0}(j, \cdot) = 0$. In particular, the entire state space is not pseudo-small. ■

From these two lemmas, we obtain

Theorem 18. *Consider a Markov chain (not necessarily ϕ-irreducible or aperiodic) on a countable state space, having at least one recurrent point (which always holds if the state space is finite). If the entire state space is pseudo-small, then the entire state space is small. (Hence, a stationary distribution exists and the chain is uniformly ergodic.)*

Proof. Let $\mathcal{X}$ be (n_0, ϵ)-pseudo-small. By Lemma 16, the chain is aperiodic. By Lemma 17, the chain is ϕ-irreducible. Hence, the result follows from Theorem 12. ■

Remark. We believe that Lemma 17 and Theorem 18 continue to hold for general chains, not just for countable chains with at least one recurrent point, though we are not yet able to prove this.

8. Convergence via a monotone dominating chain.

We consider now the situation in which a Markov chain P_1 is *stochastically dominated* by another Markov chain P_2, where P_1 has known minorisation properties and P_2 is stochastically monotone, and we wish to combine this information so that we can use standard stochastically-monotone-convergence-bounds of previous authors (e.g. Proposition 3 above). We shall see that it is possible to construct multiple copies of the P_1 chain so that they all simultaneously respect the dominance of the P_2 chain. This will allow us to establish both pseudo-smallness and smallness conditions for the P_1 chain, thereby allowing for quantitative bounds which make use of the stochastic monotonicity.

More formally, let $P_1(x, \cdot)$ and $P_2(x, \cdot)$ be two Markov chains defined on a totally ordered state space $\mathcal{X}$. Let $\pi(\cdot)$ be a stationary distribution for P_1. Suppose that for some fixed $c \in \mathcal{X}$, the subset $C = \{z \in \mathcal{X};\ z \preceq c\}$ is (n_0, ϵ, ν)-small for P_1. Suppose further that P_2 is stochastically monotone, and that P_2 *stochastically dominates* P_1, in the sense that

$$P_2(x, \{z \in \mathcal{X};\ z \succeq y\}) \ \geq\ P_1(x, \{z \in \mathcal{X};\ z \succeq y\}) \,, \qquad x, y \in \mathcal{X} \,.$$

Finally, suppose that a drift condition is satisfied by P_2; specifically, there is a function $V : \mathcal{X} \to [1, \infty)$ and $\lambda < 1$ and $b < \infty$ such that

$$\int V(y)\, P_2(x, dy) \ \leq\ \lambda V(x) + b\mathbf{1}_C(x) \,, \qquad x \in \mathcal{X} \,. \tag{7}$$

We would like to get convergence bounds on $\|P_1^n(x, \cdot) - \pi(\cdot)\|$. We know a small set C for P_1, however P_1 is *not* assumed to be stochastically monotone. Can we make use of the stochastic monotonicity of the dominating chain P_2, to allow us to use the bounds of Proposition 3 anyway?

At first this may appear to be fairly straightforward. Indeed, since P_2 stochastically dominates P_1, we see by the nature of C that the P_1 chain will be in C whenever the P_2 chain is in C. Thus, it appears that we can use the drift condition (7) to bound the return times of the P_1 chain to C, and then use the smallness of C to get couplings for the P_1 chain once it is in C, thus producing regeneration times of the P_1 chain.

The problem with this approach is that it is not obvious that we can preserve the ordering of the two chains, while at the same time giving two different copies of the P_1 chain the option of coupling (with probability ϵ) or not coupling (with probability $1 - \epsilon$) when they're both in C.

We wish to show that it is indeed possible to preserve the ordering in this sense. We begin with a lemma. For measures R_1 and R_2 on $\mathcal{X}$, we write $R_1 \overset{\text{st.}}{\leq} R_2$ to mean that $R_1(\{z \in \mathcal{X};\ z \succeq y\}) \leq R_2(\{z \in \mathcal{X};\ z \succeq y\})$ for all $y \in \mathcal{X}$.

Lemma 19. *Let R_1, R_2, and ν be probability measures, and let $0 < \epsilon \leq 1$. Suppose $R_1 \overset{\text{st.}}{\leq} R_2$, and also $R_1(\cdot) \geq \epsilon\nu(\cdot)$. Then there is a probability measure Q_1 such that $R_2(\cdot) \geq \epsilon Q_1(\cdot)$ and $\nu \overset{\text{st.}}{\leq} Q_1$, and furthermore $R_1 - \epsilon\nu \overset{\text{st.}}{\leq} R_2 - \epsilon Q_1$.*

Proof. Construct a probability space containing random variables X and I such that $P(I=1) = \epsilon = 1 - \mathbf{P}(I=0)$, such that $\mathcal{L}(X \mid I=1) = \nu$, and $\mathcal{L}(X) = R_1$. (This is essentially the standard minorisation construction, cf. the Appendix.) By the upward kernel construction of Strassen's Theorem (see for example Lindvall, 1992, Chapter 4, Section 3), this probability space can be extended to support a further random variable Y such that $\mathcal{L}(Y) = R_2$ and $\mathbf{P}(X \leq Y) = 1$. We then set $Q_1 = \mathcal{L}(Y \mid I=1)$. The conclusions follow by inspection, since $R_1 - \epsilon\nu = \mathcal{L}(X \mid I=0)$ and $R_2 - \epsilon Q_1 = \mathcal{L}(Y \mid I=0)$. ∎

Using this lemma, we prove a second, more substantial lemma. Its conclusion is similar to that of Lemma 19, except that this time we find a single Q which works for *all* R_1 simultaneously.

Lemma 20. *Let R_2 and ν be probability measures, and let $0 < \epsilon \leq 1$. Then there is a probability measure Q such that $R_2 \geq \epsilon Q$ and $\nu \overset{\text{st.}}{\leq} Q$, and furthermore for all R_1 with $R_1 \overset{\text{st.}}{\leq} R_2$ and $R_1(A) \geq \epsilon\nu(A)$ for all A, we have and $R_1 - \epsilon\nu \overset{\text{st.}}{\leq} R_2 - \epsilon Q$.*

Proof. For $x \in \mathcal{X}$, write $(-\infty, x)$ as shorthand for the subset $\{z \in \mathcal{X};\ z \prec x\}$. In terms of this, define a measure M on $\mathcal{X}$ by

$$M(dx) \;=\; \min\left[R_2(dx),\ \epsilon\nu(dx) + \epsilon\nu((-\infty,x)) - M((-\infty,x))\right] .$$

Formally, this means that if $\epsilon\nu((-\infty,x)) - M((-\infty,x)) > 0$ then M has positive mass $\epsilon\nu((-\infty,x)) - M((-\infty,x))$ at x, while if $\epsilon\nu((-\infty,x)) - M((-\infty,x)) = 0$ then $M(dx) = \min\left[R_2(dx),\ \epsilon\nu(dx)\right]$. (Intuitively, the first argument of the min ensures that $M \leq R_2$, and the second ensures that $M \overset{\text{st.}}{\geq} \epsilon\nu$.) We then set $Q = \epsilon^{-1}M$, so that $M = \epsilon\, Q$.

Intuitively, M is the minimal (in the stochastic ordering sense, not in the sense of minorisation) probability distribution which is a minorisation for R_2 and which also stochastically dominates ν.

More formally, given R_1, let Q_1 be as in Lemma 19. Then we see by inspection that Q_1 satisfies the properties

$$Q_1(dx) \leq M(dx) + M\left((-\infty,x)\right) - Q_1\left((-\infty,x)\right) .$$

(This is really two inequalities in one: It bounds the atomic components, and furthermore if the atomic components are equal then it bounds the continuous components.) It follows that

$$\epsilon\, Q_1\left((-\infty, x)\right) \leq M\left((-\infty, x)\right)\,, \qquad x \in \mathcal{X}\,. \tag{8}$$

Hence, in particular, $M(\mathcal{X}) \geq \epsilon\, Q_1(\mathcal{X}) = \epsilon$. Furthermore, since $R_1 - \epsilon\nu \overset{\text{st.}}{\leq} R_2 - \epsilon Q_1$, it follows from (8) that also $R_1 - \epsilon\nu \overset{\text{st.}}{\leq} R_2 - M$.

It now follows that the Q constructed above satisfies all of the required properties. This completes the proof. ■

This lemma allows us to conclude, finally, the key result of this section.

Corollary 21. *Let $P_1(x, \cdot)$ and $P_2(x, \cdot)$ be two Markov chains defined on a totally ordered state space $\mathcal{X}$. Suppose that P_2 is stochastically monotone, and that P_2 stochastically dominates P_1. Suppose $C = \{z \in \mathcal{X};\ z \preceq c\}$ is a small set for P_1, with minorising measure $\epsilon\nu$. Then Markov chains $\{X_n^{1,x}\}_{x \leq a}$ and $\{X_n^{2,a}\}$ can be defined, for each $x \in \mathbf{R}$, so that $\{X_n^{i,x}\}$ follows the transitions P_i, and $X_0^{i,x} = x$, and the $X_n^{1,x}$ regenerate simultaneously according to $\epsilon\nu$, and such that $X_n^{1,x} \preceq X_n^{2,a}$ for all n.*

This corollary shows that it is possible to allow multiple copies of the P_1 chain to either regenerate (with probability ϵ) or not (with probability $1 - \epsilon$), while at the same time ensuring that the P_2 chain stochastically dominates each copy of the P_1 chain at all times. This allows for the standard monotone-chain coupling construction, using either small or pseudo-small sets, in the proof of the analogue of Proposition 3. Hence, this allows us to conclude the bounds of Proposition 3 in our more general situation, as follows.

Theorem 22. *Let $P_1(x, \cdot)$ and $P_2(x, \cdot)$ be the transition probabilities for two Markov chains on a totally ordered state space $\mathcal{X}$. Let $\pi(\cdot)$ be a stationary distribution for P_1. Suppose that for some $c \in \mathcal{X}$, the set $C = \{z \in \mathcal{X};\ z \preceq c\}$ is $(1, \epsilon, \nu)$-small for P_1. Suppose further that P_2 stochastically dominates P_1, that P_2 is stochastically monotone, and that P_2 satisfies (7) for some $\lambda < 1$ and $0 \leq b < \infty$. Then for $n > \log \mathbf{E}_x^\pi(V)\, /\, \log(\lambda^{-1})$, we* *have*

$$\|P_1^n(x,\cdot) - \pi(\cdot)\| \;\leq\; K(n+\eta-\xi)\rho^n \;, \qquad n \in \mathbf{N}\,,$$

where K, ξ, η, s, and ρ are as in Proposition 3.

Remarks.

(i) If $P_2 = P_1$, then Theorem 22 reduces directly to Proposition 3.

(ii) If the function V is itself monotone with respect to the ordering $\preceq$, i.e. if $V(x) \leq V(y)$ whenever $x \preceq y$, then (7) and the stochastic dominance immediately implies a corresponding drift condition for P_1. We can thus apply Proposition 3 directly, without requiring Theorem 22 at all. However, if V is not assumed to be monotone in this sense, then such a direct approach does not appear to be possible.

Acknowledgements. We thank Wilfrid Kendall for suggesting the connection to Fill and Machida (1999), and thank the anonymous referees for very helpful reports.

APPENDIX: Constructing pairwise couplings from small sets.

Small sets have many uses. The one most relevant to the current paper is for *pairwise coupling constructions*, which we now describe. This construction is standard; for further details, see e.g. Meyn and Tweedie (1993), Rosenthal (1993, 1995a), and Roberts and Tweedie (1999).

Given a Markov chain $P(x,\cdot)$ on a state space $\mathcal{X}$, with initial distribution $\nu(\cdot)$, stationary distribution $\pi(\cdot)$, and (n_0,ϵ,ν)-small set $C \subseteq \mathcal{X}$, we proceed as follows. We construct initial random variables $X_0 \sim \nu(\cdot)$ and $Y_0 \sim \pi(\cdot)$ arbitrarily (say, independently). Then, inductively for $k = 1, 2, \ldots$, given values of X_{kn_0} and Y_{kn_0}, we construct $X_{(k+1)n_0}$ and $Y_{(k+1)n_0}$ by:

1. If $X_{kn_0} = Y_{kn_0}$, then we simply choose

$$X_{(k+1)n_0} = Y_{(k+1)n_0} \sim P^{n_0}(X_{kn_0},\cdot)\,.$$

2. If $X_{kn_0} \neq Y_{kn_0}$, then:

(a) If $(X_{kn_0}, Y_{kn_0}) \in C \times C$, then we flip an independent coin having probability ϵ of coming up heads, and then:

(i) If the coin is heads, we choose

$$X_{(k+1)n_0} = Y_{(k+1)n_0} \sim \nu(\cdot)$$

and set $T = (k+1)n_0$.

(ii) If the coin is tails, we choose

$$X_{(k+1)n_0} \sim \frac{1}{1-\epsilon}\left[P^{n_0}(X_{kn_0}, \cdot) - \epsilon\,\nu(\cdot)\right]$$

and

$$Y_{(k+1)n_0} \sim \frac{1}{1-\epsilon}\left[P^{n_0}(Y_{kn_0}, \cdot) - \epsilon\,\nu(\cdot)\right]\,,$$

conditionally independently (say).

(b) If $(X_{kn_0}, Y_{kn_0}) \notin C \times C$, then we choose

$$X_{(k+1)n_0} \sim P^{n_0}(X_{kn_0}, \cdot)$$

and

$$Y_{(k+1)n_0} \sim P^{n_0}(Y_{kn_0}, \cdot)\,,$$

conditionally independently (say).

The stopping time T is thus defined in step 2(a)(i) above; we set $T = \infty$ if case 2(a)(i) never arises. To complete the construction (if $n_0 > 1$), we "fill in" the values X_m and Y_m, for m not a multiple of n_0, as follows. For $k = 0, 1, 2, \ldots$, we choose $X_{kn_0+1}, X_{kn_0+2}, \ldots X_{kn_0+n_0-1}$ jointly from their Markov chain distribution, conditional on the constructed values of X_{kn_0} and $X_{(k+1)n_0}$. For $k = 0, 1, 2, \ldots, T/n_0$, we choose $Y_{kn_0+1}, Y_{kn_0+2}, \ldots Y_{kn_0+n_0-1}$ conditionally independently (say), jointly from their Markov chain distribution, conditional on the constructed values of Y_{kn_0} and $Y_{(k+1)n_0}$. For $m > T$, we simply set $Y_m = X_m$.

The key points of this construction are that

(I) For all m, we have

$$\mathcal{L}(X_{m+1} \,|\, X_0, \ldots, X_m, Y_0, \ldots, Y_m) \;=\; P(X_m, \cdot)$$

and

$$\mathcal{L}(Y_{m+1} \,|\, X_0, \ldots, X_m, Y_0, \ldots, Y_m) \;=\; P(Y_m, \cdot)\,.$$

(In the language of Rosenthal, 1997, this says that we have constructed a *faithful coupling* of the two chains.)

(II) Because of (I), stationarity of $\pi(\cdot)$, and the fact that $Y_0 \sim \pi(\cdot)$, we have

$$\mathcal{L}(Y_m) \;=\; \pi(\cdot)\,, \qquad m = 0, 1, 2, \ldots$$

(III) If $T \leq m$, then

$$X_m \;=\; Y_m\,.$$

In the language of coupling theory, property (III) says that T is a *coupling time* (see e.g. Lindvall, 1992). Hence, the *coupling inequality* says that

$$\|\mathcal{L}(X_n) - \pi(\cdot)\| \;\leq\; \mathbf{P}(T > n)\,.$$

This fact can be used to bound the total variation distance between the distribution of the Markov chain after n steps, and the stationary distribution $\pi(\cdot)$.

Now, if C is $(n_0, \epsilon, \{\nu_{xy}\})$-pseudo-small instead of being small, then the above construction can still be used. One merely has to replace $\nu(\cdot)$ by $\nu_{X_{kn_0} Y_{kn_0}}(\cdot)$ throughout item 2(a) above. Aside from this, the construction works without change. Indeed, that is the key observation of the current paper.

REFERENCES

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

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

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

P. Bickel (1999), Personal communication.

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

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

R.L. Dobrushin (1956), Central limit theorem for nonstationary Markov chains. Th. Prob. Appl. **1**, 65–80 and 329–383.

J.A. Fill (1998), An interruptible algorithm for perfect sampling via Markov chains. Ann. Appl. Prob. **8**, 131–162.

J.A. Fill and M. Machida (1999), Stochastic Monotonicity and Realizable Monotonicity. Preprint.

J.A. Fill, M. Machida, D.J. Murdoch, and J.S. Rosenthal (1999), Extension of Fill's perfect rejection sampling algorithm to general chains. Random Structures and Algorithms, to appear.

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

N. Jain and B. Jamison (1967), Contributions to Doeblin's theory of Markov processes. Z. Wahrsch. Verw. Geb. **8**, 19–40.

T. Kamae, U. Krengel, and G.L. O'Brien (1977), Stochastic inequalities on partially ordered spaces. Ann. Prob. **5**, 899–912.

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

R.B. Lund and R.L. Tweedie (1996), Geometric convergence rates of stochastically ordered Markov chains. Math. Oper. Research **21**, 182–194.

R.B. Lund, S.P. Meyn, and R.L. Tweedie (1996), Computable exponential convergence rates for stochastically ordered Markov processes. Ann. Appl. Prob. **6**, 218-237.

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

D.J. Murdoch and P. Green (1998), Exact Sampling from a Continuous State Space. Scandinavian J. Stat. **25,** 483–502.

P.A. Mykland, L. Tierney, and B. Yu (1995), Regeneration in Markov chain samplers. J. Amer. Stat. Assoc. **90**, 233–241.

E. Nummelin (1978), Uniform and ratio limit theorems for Markov renewal and semi-regenerative processes on a general state space. Ann. Inst. Henri Poincaré Series B **14**, 119–143.

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.G. Propp and D.B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. Random Structures and Algorithms **9**, 223–252.

G.O. Roberts and J.S. Rosenthal (1996), Quantitative bounds for convergence rates of continuous time Markov processes. Electronic J. Prob. **1**, Paper no. 9, 1–21.

G.O. Roberts and J.S. Rosenthal (1997), Shift-coupling and convergence rates of ergodic averages. Comm. in Stat. – Stochastic Models **13**, 147–165.

G.O. Roberts and J.S. Rosenthal (1998), Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). Canadian J. Stat. **26**, 5–31.

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), Rates of convergence for stochastically monotone and continuous time Markov models. J. Appl. Prob. **37**, 359–373.

J.S. Rosenthal (1993), Rates of Convergence for Data Augmentation on Finite Sample Spaces. Ann. Appl. Prob. **3**, 819–839.

J.S. Rosenthal (1995a), Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. J. Amer. Stat. Assoc. **90** (1995), 558–566.

J.S. Rosenthal (1995b), Convergence rates of Markov chains. SIAM Review **37**, 387–

405.

J.S. Rosenthal (1997), Faithful couplings of Markov chains: now equals forever. Adv. Appl. Math. **18**, 372–381.

N. J. A. Sloane (2000), The On-Line Encyclopedia of Integer Sequences. Published electronically at `http://www.research.att.com/∼njas/sequences/`.

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.

F.E. Su (1998), Convergence of random walks on the circle generated by an irrational rotation. Trans. Amer. Math. Soc. **350**, 3717–3741.

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