Convergence of Pseudo-Finite Markov Chains

by

Jeffrey S. Rosenthal

School of Mathematics, University of Minnesota, Minneapolis, MN 55455, U.S.A.

(November, 1992)

Note: After completing this paper, it was discovered that similar ideas had been studied previously by others (see A.H. Hoekstra and F.W. Steutel, *Linear Alg. Appl.* **60** (1984), 65–77; J.Th. Runnenburg and F.W. Steutel, *Ann. Math. Stat.* **33** (1962), 1483–1484). Thus, this paper was withdrawn at that point.

1. Introduction.

One important subject in the study of Markov chains is the question of their convergence to a stationary distribution, and in particular the rate of this convergence. For random walks on finite groups, great progress has been made in determining precise convergence rates in many cases, including for ordinary "riffle" card-shuffling [BD]. See [D] for background, examples, and references. For random walks on compact Lie groups, there has been some recent progress; see [R1]. For more general Markov chains, the notion of Harris recurrence (see [A], [AN], [N]) has proven useful in obtaining rates of convergence (see e.g. [T], [R2], [R3]).

Finite state-space Markov chains remain the simplest case to study, because their convergence can be analyzed directly in terms of the finite spectrum of their transition kernel; see e.g. [DS]. In this paper, we identify a class of Markov chains, which we call "pseudo-finite", which are "essentially" finite, in the following sense. There is a *finite* state space Markov chain that captures all of the important information about the pseudo-finite chain, and convergence-rate questions about the pseudo-finite chain can be answered in terms of the finite chain. Thus, to understand the pseudo-finite chain it is completely sufficient to understand the finite chain, an apparently easier problem.

The definition of pseudo-finite is as follows.

Definition. A Markov chain $\{\theta_k\}$ on a general state space Θ is *pseudo-finite* if the transition probabilities $P(\theta, \cdot)$ satisfy

$$P(\theta, \cdot) = \sum_{j=1}^{n} f_j(\theta) P_j(\ \cdot\ )$$

for some finite n, some probability distributions $P_1, \ldots, P_n$ on Θ, and some (non-random) measurable weight functions $f_1, \ldots, f_n : \Theta \to [0,1]$ with $\sum_j f_j(\theta) = 1$ for all $\theta \in \Theta$.

We shall obtain results which reduce the question of convergence of pseudo-finite Markov chains to a simple calculation (Corollary 2), and give a bound on the rate of convergence in terms of an associated finite Markov chain (Proposition 1).

Of course, pseudo-finite Markov chains are a very special case of Markov chains. Thus, this work should be seen as a small step towards understanding the convergence rate of general state space Markov chains, but by no means the complete picture. On the other hand, Proposition 3 below shows that a very large class of Markov chains are "almost" pseudo-finite in a certain natural sense.

Our interest in pseudo-finiteness arose in an application [R2] to Bayesian statistics. The Markov chains considered there were pseudo-finite, with the P_j being various beta distributions, and the $f_j(\theta)$ being related to distributions of sums of binomial distributions. The notion of pseudo-finiteness helped the author's analysis of these chains (though it was not mentioned explicitly). See [R2] for details.

2. Results.

The reason for the terminology "pseudo-finite" is given by

Proposition 1. *Given a pseudo-finite Markov chain $\{\theta_k\}_{k=0}^{\infty}$ as above, with initial distribution (at time $k = 0$) given by $\mathcal{L}(\theta_0) = \nu$, define the* finite *Markov chain $\{y_k\}_{k=1}^{\infty}$ on $\mathcal{Y} = \{1, 2, \ldots, n\}$ by*

(a) The initial distribution (at time $k = 1$) for y_1 is given by

$$Prob\,(y_1 = j) = E_\nu(f_j)\ ,$$

the expected value of f_j under the distribution ν; and

(b) The transition matrix for $\{y_k\}$ is given by

$$Prob\,(y_{k+1} = j \mid y_k = i) = T_{ji} = E_{P_i}(f_j)\ ,$$

the expected value of f_j under the distribution P_i.

Then $\{\theta_k\}$ and $\{y_k\}$ are equivalent in the sense that

(1) If $A_k^{(j)} = Prob(y_k = j)$, then for $k \geq 1$,

$$\mathcal{L}(\theta_k) = \sum_{j=1}^{n} A_k^{(j)} P_j(\,\cdot\,)\ ;$$

(2) *If* $A = \left\{A^{(j)}\right\}_{j=1}^{n}$ *is a stationary distribution for* $\{y_k\}$ *(where* $A^{(j)} = Prob(y = j)$*), then* $\pi(\cdot) = \sum_{j=1}^{n} A^{(j)} P_j(\cdot)$ *is stationary for* $\{\theta_k\}$*;*

(3) *If* $\pi(\cdot)$ *is a stationary distribution for* $\{\theta_k\}$*, and the* $P_j(\,\cdot\,)$ *are linearly independent, then* $\left\{A^{(j)} = E_\pi(f_j)\right\}$ *is stationary for* $\{y_k\}$*;*

(4) *If* A *and* π *are as in (2), then for* $k \geq 1$*,*

$$\|\mathcal{L}(\pi_k) - \pi\|_\Theta \leq \|\mathcal{L}(y_k) - A\|_{\mathcal{Y}} \ ,$$

where $\|\ \cdot\ \|_\Theta$ *and* $\|\ \cdot\ \|_{\mathcal{Y}}$ *are total variation distance on* Θ *and* $\mathcal{Y}$*, respectively.*

Proof. For (1), we proceed by induction on k. For $k = 1$,

$$\begin{aligned}\mathcal{L}(\theta_1) = \int P(\theta, \cdot)\nu(d\theta) &= \int \sum_j f_j(\theta) P_j(\cdot)\nu(d\theta) \\ &= \sum_j E_v(f_j) P_j(\cdot) = \sum_j A_1^{(j)} P_j(\cdot) \ .\end{aligned}$$

Once (1) is known for k, for $k+1$ we have

$$\begin{aligned}\mathcal{L}(\theta_{k+1}) = \int P(\theta, \cdot) Prob(\theta_k \in d\theta) &= \int \sum_j f_j(\theta) P_j(\cdot) \sum_i A_k^{(i)} P_i(d\theta) \\ &= \sum_{i,j} E_{P_i}(f_j) A_k^{(i)} P_j(\cdot) \\ &= \sum_{i,j} T_{ji} A_k^{(i)} P_j(\cdot) = \sum_j A_{k+1}^{(j)} P_j(\cdot) \ .\end{aligned}$$

Statement (2) follows immediately from statement (1).

For (3),

$$\begin{aligned}\pi(\cdot) = \int P(\theta, \cdot)\pi(d\theta) &= \int \sum_j f_j(\theta) P_j(\cdot)\pi(d\theta) \\ &= \sum_j E_\pi(f_j) P_j(\cdot) = \sum_j A^{(j)} P_j(\cdot) \ .\end{aligned}$$

Hence, iterating this expression once,

$$\begin{aligned}\sum_j A^{(j)} P_j(\cdot) &= \int P(\theta, \cdot) \sum_i A^{(i)} P_i(d\theta) \\ &= \int \sum_j f_j(\theta) P_j(\cdot) \sum_i A^{(i)} P_i(d\theta) \\ &= \sum_{i,j} E_{P_i}(f_j) A^{(i)} P_j(\cdot) \\ &= \sum_j \left(\sum_i T_{ji} A^{(i)} \right) P_j(\cdot) \ .\end{aligned}$$

By linear independence of the P_j, we must have $A^{(j)} = \sum_i T_{ij} A^{(i)}$.

For (4), we note that for any $S \subseteq \Theta$,

$$\begin{aligned}
&|Prob(\theta_k \in S) - \pi(S)| \\
&\quad = |\sum_j A_k^{(j)} P_j(S) \; - \; \sum_j A^{(j)} P_j(S)| \qquad \text{using (1)} \\
&\quad = |E_{A_k}(\phi) - E_A(\phi)| \qquad\qquad \text{where } \phi : \mathcal{Y} \to [0,1] \;\text{ by }\; \phi(j) = P_j(S) \\
&\quad \leq \sup_{\phi : \mathcal{Y} \to [0,1]} |E_{A_k}(\phi) - E_A(\phi)| \;=\; \|A^{(k)} - A\|_{\mathcal{Y}} \; .
\end{aligned}$$

■

This proposition shows that convergence of pseudo-finite Markov chains can be understood in terms of the theory of *finite* chains. For example, it is well known that a strictly positive finite Markov chain (i.e. one with $P(x,y) > 0$ for all x, y) converges exponentially quickly to a (strictly positive) unique stationary distribution. This immediately implies

Corollary 2. *Let $\{\theta_k\}$ be a pseudo-finite Markov chain, with $f_j(\theta)$ and $P_j(\cdot)$ the associated weight functions and distributions. Suppose $E_{P_i}(f_j) > 0$ for each i, j. Then $\{\theta_k\}$ has a unique invariant distribution to which it converges exponentially quickly.*

Proposition 1 (4) shows that to study the rate of convergence of a pseudo-finite Markov chain $\{\theta_k\}$, one need only consider an equivalent, finite Markov chain $\{y_k\}$. We illustrate this with a simple example.

Example. Consider a Markov chain $\{\theta_k\}$ defined on the unit interval [0,1] with $\theta_0 = 1/3$, and with the following transition mechanism: given θ_k, we choose θ_{k+1} by
(a) with probability $\theta_k/2$, choosing θ_{k+1} uniform on $[0, \frac{1}{2}]$;
(b) with probability $(\theta_k)^2/2$, choosing θ_{k+1} uniform on $[\frac{1}{2}, 1]$;
(c) with probability $1 - \theta_k/2 - (\theta_k)^2/2$, choosing θ_{k+1} from the beta distribution $Beta(2,2)$.
To analyze this Markov chain directly on [0,1] would be somewhat involved; however, using the notion of pseudo-finiteness it is very easy. Using the notation in the definition, we have $n = 3$, $f_1(\theta) = \theta/2$, $f_2(\theta) = \theta^2/2$, and $f_3(\theta) = 1 - \theta/2 - \theta^2/2$. Also P_1 is uniform on $[0, \frac{1}{2}]$, P_2 is uniform on $[\frac{1}{2}, 1]$, and $P_3 = Beta(2,2)$. Thus the matrix T_{ji} is given by

$$T_{ji} = E_{P_i}(f_j) = \begin{pmatrix} \frac{1}{8} & \frac{3}{8} & \frac{1}{2} \\ \frac{1}{24} & \frac{7}{24} & \frac{3}{10} \\ \frac{5}{6} & \frac{1}{3} & \frac{1}{5} \end{pmatrix} \; .$$

Also the initial distribution of $\{y_k\}$ is

$$\mathcal{L}(y_1) = [f_j(\frac{1}{3})]_j = [\frac{1}{6}, \frac{1}{18}, \frac{7}{9}] \ .$$

We work numerically for simplicity. We compute that the eigenvalues of the matrix T are

$$\lambda_1 = 1; \qquad \lambda_2 = 0.06505; \qquad \lambda_3 = -0.44838 \ ,$$

with corresponding eigenvectors

$$v_1 = [0.3446, 0.2092, 0.4462]; \qquad v_2 = [0.3968, -1.3968, 1]; \qquad v_3 = [-0.6301, -0.3699, 1]$$

(so v_1 is the stationary distribution for $\{y_k\}$). In terms of these eigenvectors,

$$\mathcal{L}(y_1) = v_1 + 0.0302 v_2 + 0.3014 v_3 \ ,$$

so

$$\mathcal{L}(y_k) = v_1 + 0.0302(\lambda_2^{k-1}) v_2 + 0.3014(\lambda_3^{k-1}) v_3 \ .$$

Hence,

$$\|\mathcal{L}(y_k) - v_1\|_{\mathcal{Y}} = \frac{1}{2} \|0.0302(\lambda_2^{k-1}) v_2 + 0.3014(\lambda_3^{k-1}) v_3\|_{L^1} < (0.55)(\frac{1}{2})^k \text{ (say)}.$$

We thus conclude that our original chain $\{\theta_k\}$ has a unique stationary distribution given by

$$\pi(\cdot) = (0.3446) P_1(\cdot) + (0.2092) P_2(\cdot) + (0.4462) P_3(\cdot) \ ,$$

and that for $k \geq 1$,

$$\|\mathcal{L}(\theta_k) - \pi\|_{\Theta} < (0.55)(\frac{1}{2})^k \ .$$

■

We acknowledge that the applications of Proposition 1 will be somewhat limited, however they do sometimes arise naturally as the work in [R2] indicates.

We conclude with the observation that a large class of Markov chains is "almost" pseudo-finite. Indeed, we have

Proposition 3. *Let $\{\theta_k\}$ be a Markov chain on a state space Θ, with transition kernel $P(\theta, \cdot)$. Assume that Θ is compact, and that there is a measure ν on Θ for which $P(\theta, \cdot) <<$*

$\nu(\cdot)$ for all θ, say with density $f_\theta(\theta')$. Write $f(\theta,\theta')$ for $f_\theta(\theta')$. Assume further that f is a continuous function of two variables, except for possible jump-discontinuities on a finite rectangular grid. Then for any $\epsilon > 0$, there is a pseudo-finite Markov chain $Q(\theta,\cdot)$ on Θ, with

$$|Q(\theta,S) - P(\theta,S)| < \epsilon \ ,$$

for all $\theta \in \Theta$ and $S \subseteq \Theta$.

Proof. Since $f(\theta,\theta')$ is a rectangularly piecewise continuous function on the compact set $\Theta\times\Theta$, given $\epsilon > 0$ we can find a function f_0 on $\Theta\times\Theta$ of the form $f_0(\theta,\theta') = \sum_{i=1}^{n} g_i(\theta)h_i(\theta')$ for which

$$|f(\theta,\theta') - f_0(\theta,\theta')| < \epsilon \qquad \text{for all } \theta,\theta' \in \Theta \ .$$

(Indeed, we can take g_i and h_i to be step functions.) Define $Q(\cdot,\cdot)$ by

$$Q(\theta,\cdot) = \sum_i g_i(\theta)\,\nu_i(\cdot)$$

where $d\nu_i = h_i\,d\nu$. Then $Q(\cdot,\cdot)$ is pseudo-finite, and for any θ and S,

$$\begin{aligned} |P(\theta,S) - Q(\theta,S)| &= \left| \int_S P(\theta,d\theta') \ - \ \sum_i \int_S g_i(\theta)\,\nu_i(d\theta') \right| \\ &= \left| \int_S f(\theta,\theta')\nu(d\theta') \ - \ \int_S f_0(\theta,\theta')\,\nu(d\theta') \right| \\ &< \epsilon\,\nu(S) \ \le \ \epsilon \ , \end{aligned}$$

as required. ∎

Acknowledgements. I am very grateful to Persi Diaconis, Jim Fill, and James M^{c}Kernan for discussions and encouragement.

REFERENCES

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

[AN] 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.

[BD] D. Bayer and P. Diaconis (1992), *Trailing the Dovetail Shuffle to its Lair,* Ann. Prob. **2**, 294-313.

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

[DS] P. Diaconis and D. Stroock (1991), *Geometric bounds for eigenvalues of Markov chains,* Ann. Appl. Prob. **1**, 36-61.

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

[R1] J.S. Rosenthal (1991), *Random Rotations: Characters and Random Walks on $SO(N)$,* to appear in Annals of Probability.

[R2] J.S. Rosenthal (1991), *Rates of convergence for data augmentation on finite sample spaces,* to appear in Annals of Applied Probability.

[R3] J.S. Rosenthal (1991), *Rates of convergence for Gibbs sampling for variance component models,* preprint.

[T] L. Tierney (1991), *Markov Chains for Exploring Posterior Distributions,* Tech. Rep. 560, School of Statistics, University of Minnesota.