RATES OF CONVERGENCE FOR EVERYWHERE-POSITIVE MARKOV CHAINS

J.R. Baxter* and Jeffrey S. Rosenthal**

(January, 1994; revised March, 1994.)

0. Introduction

It is often useful to know that the distribution of a Markov process converges to a stationary distribution, and if possible to know how rapidly convergence takes place. Such rates of convergence are of particular interest when running stochastic algorithms such as Markov chain Monte Carlo (see Gelfand and Smith, 1990; Tierney, 1994), since they indicate how long the algorithm should be run before it gives satisfactory answers. Related convergence questions have been studied in an operator-theoretic context (Orey, 1962; Ornstein and Sucheston, 1970; Baxter, 1978), and more recently to obtain quantitative bounds by probabilistic methods (Meyn and Tweedie, 1993b; Rosenthal, 1993).

The Markov processes in applications such as Markov chain Monte Carlo often have the property that they are *everywhere-positive*, in the sense that there is a σ-finite reference measure with respect to which the transition kernel $P(x,\cdot)$ has an everywhere-positive density for each x. That such processes eventually converge to a stationary distribution (if one exists) follows from standard results (e.g. Tierney, 1994, Theorem 1). We investigate the rates of convergence for such processes here.

We will use total variation norm to measure the closeness of two measures. Theorem 1, stated below, establishes convergence in total variation norm at a geometric rate, for everywhere-positive processes such that the Markov transition operator is a compact operator on an appropriate Hilbert space. Theorem 1 was proved by Schervish and Carlin (1992) for the specific case of the Gibbs sampler, under additional assumptions; see also Liu, Wong, and Kong (1991a, 1991b). The proof we give here seems to be simpler as well as more general. This is presented in Section 1 below.

Unfortunately, the compactness assumption need not be satisfied in general, and it is easy to see that geometric convergence of the sort established in Theorem 1 is not always possible. As a partial replacement for Theorem 1, in Theorem 2 we provide an estimate giving a (non-geometric) quantitative rate of convergence valid for general Markov processes having everywhere-positive transition densities. This estimate is proved in Section 2.

Before stating our results more precisely we define some notation.

Let $(\mathcal{X},\mathcal{B})$ be a measurable space, and consider a Markov process with state space $(\mathcal{X},\mathcal{B})$. We will denote the Markov transition function as usual by P, so that $P(x,\cdot)$ is a

* School of Mathematics, University of Minnesota, Minneapolis, MN 55455, U.S.A. Internet: baxter@math.umn.edu

** Department of Statistics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Internet: jeff@utstat.toronto.edu

probability measure on $\mathcal{B}$ for each $x \in \mathcal{X}$, and $P(\cdot, A)$ is a $\mathcal{B}$-measurable function for each $A \in \mathcal{B}$.

The Markov operator associated with the Markov transition function P will also be denoted by P. We recall that P acts to the right on functions and to the left on measures, so that

$$\mu P(A) = \int P(x, A)\, \mu(dx), \quad Pf(x) = \int f(y)\, P(x, dy).$$

For background on general Markov chains, see Revuz (1984) and Meyn and Tweedie (1993a). For background on operator theory, see Reed and Simon (1972) or Rudin (1991).

Throughout this note we assume that there exists an invariant probability measure π for P, so that $\pi P = \pi$. A simple application of Jensen's Inequality then gives the following standard fact:

Lemma 1. *For every r, $1 \leq r \leq \infty$, P is a weak $L^r(\pi)$-contraction. That is,*

$$\|Pf\|_r \leq \|f\|_r$$

for every $f \in L^r(\pi)$.

Indeed, for $r < \infty$ we have

$$\begin{aligned} \int |Pf|^r \, d\pi &\leq \int \Big(\int |f(y)|\, P(x, dy) \Big)^r \, \pi(dx) \\ &\leq \int \int |f(y)|^r \, P(x, dy)\, \pi(dx) \\ &= \int |f(y)|^r \, \pi(dy), \end{aligned}$$

as claimed. For $r = \infty$, we note first that $\pi P = \pi$ implies that for any π-null set A, $\{x : \ P(x, A) > 0\}$ is also a π-null set. Thus $f \leq g$ (π-a.e.) implies $Pf \leq Pg$ (π-a.e.) and the lemma follows.

It is easy to see that if ν is a probability measure which is absolutely continuous with respect to π, then νP is also absolutely continuous with respect to π. Let λ denote any σ-finite measure which is mutually absolutely continuous with respect to π. We will denote the density of π with respect to λ by φ. We can use λ as a reference measure, and express the operation of P on measures as an operation on densities with respect to λ, as follows. For any $f \in L^1(\lambda)$, let ν be the signed measure whose density with respect to λ is f. Define $T_\lambda f$ to be the density of νP with respect to λ. The operator T_λ represents the action of P on those measures which are absolutely continuous with respect to λ. We write S for T_π.

Since P is obviously a weak contraction with respect to total variation norm on measures, it is easy to see that T_λ is an $L^1(\lambda)$-contraction. Also, for any $f \in L^1(\lambda)$ and any bounded measurable function g, we have

$$\int (T_\lambda f) g \, d\lambda = \int f \ \ Pg \, d\lambda,$$

and hence T_λ agrees on $L^s(\lambda)$ with the adjoint of P on $L^r(\lambda)$, for any r with $1 \le r < \infty$ such that P is a bounded operator on $L^r(\lambda)$, where $1/r + 1/s = 1$. It follows in particular that

Corollary to Lemma 1 *T_π is a weak $L^s(\pi)$-contraction for every s, $1 \le s \le \infty$.*

For general λ, the operator T_λ is a weak contraction on $L^1(\lambda)$ but, unlike T_π, the operator T_λ is not in general a weak contraction on $L^s(\lambda)$ for $s > 1$. However, following Schervish and Carlin (1992), we shall have occasion to consider the space $L^2(\mu)$, where μ is the measure with density $1/\varphi$ with respect to λ. A simple computation shows that T_λ is a weak contraction with respect to $L^2(\mu)$.

Theorem 1 *Given a Markov process on $\mathcal{X}$ with transition function P and stationary distribution π, suppose there is a σ-finite measure λ, mutually absolutely continuous with respect to π, such that*

(i) the transition function P for the process is such that for each $x \in \mathcal{X}$ the measure $P(x, \cdot)$ is absolutely continuous with respect to λ, and there is a jointly measurable function $p(x, y)$ on $\mathcal{X} \times \mathcal{X}$ such that $p(x, \cdot)$ is the density of $P(x, \cdot)$ with respect to λ for each $x \in \mathcal{X}$, $p(x, \cdot) > 0$ λ-almost everywhere for each x, and

(ii) the restriction of the operator T_λ to $L^2(\mu)$ is a compact operator, where μ is defined by $\frac{d\mu}{d\lambda} = \left(\frac{d\pi}{d\lambda}\right)^{-1}$ as above.

Then

(a) there exists a constant α with $0 \le \alpha < 1$ such that $\|T_\lambda f\|_2 \le \alpha \|f\|_2$ for every $f \in L^2(\mu)$ with $\int f \, d\lambda = 0$, and

(b) for any initial probability measure ν, if for some time m the distribution νP^m has a density with respect to π which is square-integrable then νP^n converges to π geometrically fast in total variation norm as $n \to \infty$.

We will give a proof of a slightly more general version of Theorem 1 in the next section.

In Section 2 we will prove the next result, in which the compactness assumption is dropped.

Theorem 2 *Suppose that assumption (i) of Theorem 1 holds. For any real s with $s > 1$ and any $\varepsilon > 0$, there is a number β with $0 \le \beta < 1$, such that*

(a) if $f \in L^s(\pi)$ and $\int f \, d\pi = 0$, then for all n such that $\|S^{n-1} f\|_1 \ge \varepsilon \|f\|_s$ we have $\|S^n f\|_1 \le \beta^n \|f\|_1$, and

(b) If ν is an initial probability distribution and νP^m has a density g with respect to π for some m, where $g \in L^s(\pi)$, then $\|\nu P^{m+k} - \pi\| \le \beta^k \|\nu P^m - \pi\|$ for all $k \ge 0$ such that $\|\nu P^{m+k-1} - \pi\| \ge \varepsilon \|g - 1\|_s$.

Since $\|\nu P^{m+k} - \pi\|$ is nonincreasing, this theorem asserts that the convergence to the stationary distribution is geometric with convergence factor β until $\|\nu P^{m+k} - \pi\|$ is sufficiently small.

1. Compactness of the Markov Operator

In this section we prove Theorem 1, with a slightly more general version of the second assumption. The idea is that we can find a "worst" function f, and then argue that since T_λ is everywhere-positive, there must be some "cancellation" in obtaining $T_\lambda f$ from f, so that $T_\lambda f$ must have smaller norm that f.

We first reduce the theorem to the case in which $\lambda = \pi$. Indeed, the map $W : L^2(\mu) \to L^2(\pi)$ defined by $Wf = f/\varphi$ is an isometry such that $WT_\lambda W^{-1} = T_\pi = S$. Hence T_λ is compact if and only if S is. Furthermore, if assumption (i) holds for some λ which is mutually absolutely continuous with respect to π, then it must hold for λ replaced by π. Thus it is enough to establish the theorem in the case that $\lambda = \mu = \pi$, so that $T_\lambda = S$.

Let s be real with $1 < s < \infty$, and let V be a closed subspace of $L^s(\pi)$ such that $\int f \, d\pi = 0$ for every $f \in V$. (In Theorem 1 we take $s = 2$ and let V be the whole space of functions f with $\int f \, d\pi = 0$.) Let K be the set of $f \in V$ with $\|f\|_s \le 1$. We suppose that the operator S is compact on V in the sense that SK is a precompact subset of $L^s(\pi)$. Under this assumption we show

Lemma *There is a number α with $0 \le \alpha < 1$ such that $\|Sf\|_s \le \alpha \|f\|_s$ for every $f \in V$.*

Proof. The space V is convex and closed with respect to the norm topology, hence weakly closed. Since $L^s(\pi)$ is reflexive, the unit ball K is weakly compact by Alaoglu's Theorem. Since S is bounded, S is weakly continuous. Hence SK is weakly compact, and hence is closed. We have assumed that SK is precompact, hence it is in fact compact with respect to the norm topology. Let α denote the supremum of the numbers $\|Sf\|_s$, as f ranges over K. The compactness of SK implies at once that there is some element f in K with $\|Sf\|_s = \alpha$. In other words, the operator S on V assumes its norm on V.

Suppose $\alpha > 0$. We may write $f = g - h$, where g, h are the positive and negative parts of f. Since $p > 0$ π-almost everywhere, it is easy to see that $Sg \wedge Sh > 0$ π-almost everywhere, and hence in particular that $\|(Sf)^+\|_s < \|Sg\|_s$ and $\|(Sf)^-\|_s < \|Sh\|_s$. Thus

$$\begin{aligned}
\int |Sf|^s d\pi &= \int ((Sf)^+)^s \, d\pi + \int ((Sf)^-)^s \, d\pi \\
&< \int (Sg)^s \, d\pi + \int (Sh)^s \, d\pi \\
&\le \int g^s \, d\pi + \int h^s \, d\pi = \int |f|^s \, d\pi.
\end{aligned}$$

It follows that $\alpha < 1$ as claimed, proving the lemma.

The lemma implies part (a) of Theorem 1 at once. For part (b) we set $f = g - 1$, where g is the density of νP^m. Using the Cauchy-Schwarz inequality, and with α as in part (a), we have

$$\|\nu P^n - \pi\|_{\mathrm{TV}} = \frac{1}{2}\|S^{n-m} f\|_1 \le \frac{1}{2}\|S^{n-m} f\|_2 \le \frac{1}{2}\alpha^{n-m}\|f\|_2 \, ,$$

which goes to 0 exponentially quickly, establishing part (b).

Remarks

(I) Our argument clearly works for any operator S satisfying (i) that assumes its norm on V.

(II) In the case of the Gibbs sampler, Schervish and Carlin (1992) assume that S is Hilbert-Schmidt, i.e. that

$$\int p(x,y)^2\,\pi(dx)\,\pi(dy) < \infty.$$

As they note, the Hilbert-Schmidt property is a sufficient condition for the compactness of S. This might be the most convenient way to check whether Theorem 1 is applicable.

(III) The compactness assumption in Theorem 1 is indeed necessary. There are many known examples of non-geometric, everywhere-positive Markov chains. For one simple example, let $\mathcal{X}$ be the set of positive integers, let $P(i,\{j\}) = \frac{i}{2j}\frac{1}{2^{i\vee j}}$ for $i \neq j$, and take $\pi(\{i\})$ proportional to $\frac{1}{i^2}$. Then $\pi P = \pi$, and P is everywhere-positive, but it is easily seen that (with initial distribution concentrated at the point 1, say) the total variation distance to π goes down only at rate $1/k$.

2. An Estimate for Convergence

In this section we prove Theorem 2. The idea of the proof is that since $S^{n-1}f$ has large enough L^1-norm, when we apply S to it, its positive and negative parts will cancel each other out to some extent. We now proceed to quantify this cancellation.

For any $\delta > 0$, let

$$A(\delta) = \{(x,y) :\ p(x,y) \geq \delta\},$$

and for any $\delta > 0$ and any $y \in \mathcal{X}$ let

$$A_1(\delta, y) = \{x :\ p(x,y) \geq \delta\}.$$

Suppose that μ is any probability on $\mathcal{B}$. As a consequence of assumption (i), $\mu \times \pi(A(\delta)^c) \searrow 0$ as $\delta \searrow 0$.

Set

$$b = \mu \times \pi(A(\delta)^c).$$

Since

$$\mu \times \pi(A(\delta)^c) = \int \mu(A_1(\delta,y)^c)\,\pi(dy),$$

we see that

$$\pi(\{y :\ \mu(A_1(\delta,y)) \leq 1/2\}) \leq 2b.$$

Hence

$$\pi(\{y : \int p(x,y)\,\mu(dx) \leq \delta/2\}) \leq 2b. \tag{1}$$

In other words, the set on which the density of μP is smaller than $\delta/2$ has π-measure no larger than $2b$.

Now let u, v be any nonnegative functions in $L^1(\pi)$ such that $\int u\, d\pi = \int v\, d\pi = 1$. Let μ be the measure with density u with respect to π, and let ν be the measure with density v. By Hölder we have (recalling that $1/r + 1/s = 1$)

$$\mu \times \pi(A(\delta)^c) \leq \|u\|_s(\pi \times \pi(A(\delta)^c))^{1/r}, \quad \nu \times \pi(A(\delta)^c) \leq \|v\|_s(\pi \times \pi(A(\delta)^c))^{1/r}.$$

Recall that Su is simply the density of μP. Hence by (1) we have $Su \geq \delta/2$ on a set which has π-measure at least $1 - 2\|u\|_s(\pi \times \pi(A(\delta)^c))^{1/r}$, and a similar statement holds for Sv. Thus

$$\|(Su) \wedge (Sv)\|_1 \geq \frac{\delta}{2}(1 - 2(\|u\|_s + \|v\|_s)(\pi \times \pi(A(\delta)^c))^{1/r}).$$

By normalizing, we can extend this inequality to the case of any nonnegative measurable functions u, v such that $\int u\, d\pi = \int v\, d\pi$. In this case we have

$$\|(Su) \wedge (Sv)\|_1 \geq \frac{\delta}{2}(\|u\|_1 - 2(\|u\|_s + \|v\|_s)(\pi \times \pi(A(\delta)^c))^{1/r}).$$

Now let n be such that $\|S^{n-1}f\|_1 \geq \varepsilon\|f\|_s$, where f is the function described in Theorem 2. Let $u = (S^{n-1}f)^+$, $v = (S^{n-1}f)^-$. Since $\|S^{n-1}f\|_s \leq \|f\|_s$, we find that

$$\|(Su) \wedge (Sv)\|_1 \geq \frac{\delta}{2}(\|u\|_1 - 4\|f\|_s(\pi \times \pi(A(\delta)^c))^{1/r}).$$

Clearly $\|u\|_1 = (1/2)\|S^{n-1}f\|_1 \geq (\varepsilon/2)\|f\|_s$. Choose δ such that $4(\pi \times \pi(A(\delta)^c))^{1/r} \leq \varepsilon/4$. Then

$$\|(Su) \wedge (Sv)\|_1 \geq \frac{\delta\varepsilon}{8}\|f\|_s \geq \frac{\delta\varepsilon}{4}\|u\|_1$$

and hence $\|S^n f\|_1 \leq (1 - \delta\varepsilon/4)\|S^{n-1}f\|_1$. This proves part (a) of Theorem 2 with $\beta = 1 - \delta\varepsilon/4$. Part (b) follows by taking $f = g - 1$.

Remark Since our formula for β is in some sense explicit, it might be useful computationally. However, the estimation of the π-measure of sets required to choose δ seems difficult to carry out.

Acknowledgements. We thank Peter Rosenthal for helpful discussions, and thank the referee for useful comments. This work was partially supported by NSF and by NSERC.

References

J.R. Baxter (1978), Harmonic functions and mass cancellation. Trans. Amer. Math. Soc. **245**, 375-384.

A.E. Gelfand and A.F.M. Smith (1990), Sampling based approaches to calculating marginal densities. J. Amer. Stat. Assoc. **85**, 398-409.

S. Orey (1962), An ergodic theorem for Markov chains. Z. Wahrscheinlichkeitstheorie Verw. Gebeite **1**, 174-176.

D. Ornstein and L. Sucheston (1970), An operator theorem on L_1 convergence to zero with applications to Markov kernels. Ann. Math. Stat. **41**, 1631-1639.

J. Liu, W. Wong, and A. Kong (1991a), Correlation structure and the convergence of the Gibbs sampler, *I*. Tech. Rep. **299**, Dept. of Statistics, University of Chicago. Biometrika, to appear.

J. Liu, W. Wong, and A. Kong (1991b), Correlation structure and the convergence of the Gibbs sampler, *II*: Applications to various scans. Tech Rep. **304**, Dept. of Statistics, University of Chicago. J. Royal Stat. Sci. **(B)**, to appear.

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

S.P. Meyn and R.L. Tweedie (1993b), Computable bounds for convergence rates of Markov chains. Tech. Rep., Dept. of Statistics, Colorado State University.

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

D. Revuz (1984), Markov chains, 2$^{\text{nd}}$ ed. North-Holland, Amsterdam.

J.S. Rosenthal (1993), Minorization conditions and convergence rates for Markov chain Monte Carlo. Tech. Rep. **9321**, Dept. of Statistics, University of Toronto.

W. Rudin (1991), Functional Analysis, 2$^{\text{nd}}$ ed. McGraw-Hill, New York.

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

L. Tierney (1994), Markov chains for exploring posterior distributions. Ann. Stat., to appear.