2

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENT

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b.

2a. SECURITY CLASSIFICATION AUTH
NA

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for Public Release; Distribution Unlimited

2b. DECLASSIFICATION/DOWNGRADING ULE
NA

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
FSU Statistics Report M-868

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARO 27868.16-MA

6a. NAME OF PERFORMING ORGANIZATION
Florida State University

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
AFOSR/NM

6c. ADDRESS *(City, State and ZIP Code)*
Department of Statistics
Tallahassee, FL 32306-3033

7b. ADDRESS *(City, State and ZIP Code)*
Bldg. 410
Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL *(If applicable)*
NM

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
DAAL03-90-G-0103

8c. ADDRESS *(City, State and ZIP Code)*
Bldg. 410
Bolling AFB, DC 20332-6448

10. SOURCE OF FUNDING NOS.

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT NO.

11. TITLE *(Include Security Classification)*
A Proof of Convergence of the Markov Chain Simulation Method

12. PERSONAL AUTHOR(S)
Krishna B. Athreya, Hani Doss, and Jayaram Sethuraman

13a. TYPE OF REPORT
Technical

13b. TIME COVERED
FROM ________ TO ________

14. DATE OF REPORT *(Yr., Mo., Day)*
July 1992

15. PAGE COUNT
30

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Successive substitution sampling, calculation of posterior distributions, ergodic theorem.

ntify by block number)

400 277 3288

Abstract

The Markov chain simulation method has been successfully used in many problems, including some that arise in Bayesian statistics. We give a self-contained proof of the convergence of this method in general state spaces under conditions that are easy to verify. We also provide a result giving a geometric rate of convergence.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. □ DTIC USERS □

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Hani Doss

22b. TELEPHONE NUMBER *(Include Area Code)*
(904) 644-3218

22c. OFFICE SYMBOL
AFOSR/NM

A Proof of Convergence of the Markov Chain Simulation Method

Krishna B. Athreya[1]

Department of Statistics
Iowa State University
Ames, Iowa 50011

Hani Doss[2] and Jayaram Sethuraman[3]

Department of Statistics
Florida State University
Tallahassee, Florida 32306–3033

July 1992

FSU Technical Report No. 868
AFOSR Series D Technical Report No. 7
USARO Technical Report No. 125

[1] Research supported by National Science Foundation Grant NSFDMS-90-07182.

[2] Research supported by Air Force Office of Scientific Research Grant 90-0202.

[3] Research supported by Army Research Office Grant DAAL03-90-G-0103.

A Proof of Convergence of the Markov Chain Simulation Method

Krishna B. Athreya

Iowa State University

Hani Doss and Jayaram Sethuraman

Florida State University

July 1992

Abstract

The Markov chain simulation method has been successfully used in many problems, including some that arise in Bayesian statistics. We give a self-contained proof of the convergence of this method in general state spaces under conditions that are easy to verify. We also provide a result giving a geometric rate of convergence.

Key words and phrases: Successive substitution sampling, calculation of posterior distributions, ergodic theorem.

1 Introduction

Let π be a probability distribution on a measurable space $(\mathcal{X}, \mathcal{B})$, and suppose that we are interested in estimating characteristics of it such as $\pi(E)$ or $\int f d\pi$ where $E \in \mathcal{B}$ and f is a bounded measurable function. Even when π is fully specified one may have to resort to methods like Monte Carlo simulation methods, especially when π is not computationally tractable. For this one uses the available huge literature on generation of random variables from an explicitly or implicitly described probability distribution π. Generally these methods require $\mathcal{X}$ to be the real line or require that π have special features, such as a structure in terms of independent real valued random variables. When one cannot generate random variables with distribution π one has to be satisfied with looking for a sequence of random variables $X_1, X_2, \ldots$ whose distributions converge to π and using X_n with a large index n as an observation from π. An example is the classical Markov chain simulation method, which can be described as follows.

Let $P(x, A)$ be a transition probability function with the property that it has stationary distribution π, i.e.

$$\pi(C) = \int P(x, C)\pi(dx) \quad \text{for all } C \in \mathcal{B}. \tag{1.1}$$

We fix a starting point x_0, generate an observation X_1 from $P(x_0, \cdot)$, generate an observation X_2 from $P(X_1, \cdot)$, etc. This generates the Markov chain $x_0 = X_0, X_1, X_2, \ldots$. Let $P^n(x, \cdot)$ denote the distribution of X_n when the chain is started at x. If we can show that

$$\sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \text{for all } x \in \mathcal{X},$$

then by running the chain sufficiently long, we succeed in generating an observation X_n with distribution approximately π. Then, we may estimate π for example by generating G such chains in parallel, obtaining independent observations $X_n^{(1)}, \ldots, X_n^{(G)}$, or by running one (or a few) very long chains. In Section 3 we make some remarks on the advantages and disadvantages of these two methods.

The Metropolis algorithm and its variants produce Markov transition functions satisfying (1.1). This algorithm was originally developed for estimating certain distributions and expectations arising in statistical physics, but can also be used in Bayesian analysis; see Tierney (1991) for a review.

However, in the usual problems of Bayesian statistics, the most commonly used Markov chain is one that is used to estimate the unknown joint distribution $\pi = \pi_{X^{(1)},\ldots,X^{(p)}}$ of the (possibly vector-valued) random variables $(X^{(1)}, \ldots, X^{(p)})$ by updating the coordinates one at a time, as follows. We suppose that we know the conditional distributions $\pi_{X^{(i)}|\{X^{(j)}\, j \neq i\}}$, $i = 1, \ldots, p$ or at least that we are able to generate observations from these conditional distributions. If $X_m = (X_m^{(1)}, \ldots, X_m^{(p)})$ is the current state, the next state $X_{m+1} = (X_{m+1}^{(1)}, \ldots, X_{m+1}^{(p)})$ of the Markov chain is formed as follows. Generate $X_{m+1}^{(1)}$ from $\pi_{X^{(1)}|\{X^{(j)}\, j \neq 1\}}(\cdot, X_m^{(2)}, \ldots, X_m^{(p)})$, then $X_{m+1}^{(2)}$ from $\pi_{X^{(2)}|\{X^{(j)}\, j \neq 2\}}(X_{(m+1)}^{(1)}, \cdot, X_m^{(3)}, \ldots, X_m^{(p)})$, and so on until $X_{m+1}^{(p)}$ is generated from $\pi_{X^{(p)}|\{X^{(j)}\, j \neq p\}}(X_{(m+1)}^{(1)}, \ldots, X_{(m+1)}^{(p-1)}, \cdot)$. If P is the transition function that produces X_{m+1} from X_m, then it is easy to see that P satisfies (1.1).

This method is reminiscent of the simulation method described in Geman and Geman (1984). In that paper, p, the number of coordinate indices in the vector $(X^{(1)}, \ldots, X^{(p)})$, is usually of the order of $N \times N$ where $N = 256$ or higher. They assume that these indices form a graph with a meaningful neighborhood structure and that π is a Gibbs distribution, so that the conditional distributions $\pi_{X^{(i)}|\{X^{(j)}\ j\neq i\}}$, $i = 1, \ldots, p$ depend on much fewer that $p-1$ coordinates. They also assume that each random variable X_i takes only a finite number k of values and that π gives positive mass to all possible k^{N^2} values. Geman and Geman (1984) appeal to the ergodic theorem on Markov chains with a finite state space and prove that this simulation method works. They prove other interesting results on how this can be extended when a temperature parameter T (that can be incorporated into π) is allowed to vary. This may be the reason why the method described in the previous paragraph has come to be known as the Gibbs sampler. We consider this to be a misnomer, because no Gibbs distribution nor any graph with a nontrivial neighborhood structure supporting a Gibbs distribution is involved in this method; we will refer to it simply as successive substitution sampling.

We note that this algorithm depends on π only through the conditional distributions $\pi_{X^{(i)}|\{X^{(j)}\ j\neq i\}}$. Perhaps the first thought that comes to mind when considering this method is to ask whether or not, in general, these conditionals determine the joint distribution π. The answer is that in general they do not; we give an example in Remark 3 of Section 2.2. A necessary consequence of convergence of successive substitution sampling is that the joint distribution is determined by the conditionals. It is therefore clear that any theorem giving conditions guaranteeing convergence also gives, indirectly, conditions which guarantee that the conditionals determine the joint distribution π.

We now give a very brief description of how this method is useful in some Bayesian problems. We suppose that the parameter θ has some prior distribution, that we observe a data point Y whose conditional distribution given θ is $\mathcal{L}(Y \mid \theta)$, and that we wish to obtain $\mathcal{L}(\theta \mid Y)$, the conditional distribution of θ given Y. It is often the case that if we consider an (unobservable) auxiliary random variable Z, then the distribution $\pi_{\theta,Z} = \mathcal{L}(\theta, Z \mid Y)$ has the property that $\pi_{\theta|Z}$ $(= \mathcal{L}(\theta \mid Y, Z))$ and $\pi_{Z|\theta}$ $(= \mathcal{L}(Z \mid Y, \theta))$ are easy to calculate. Typical examples are missing and censored data problems. If we have a conjugate family of prior distributions on θ, then we may take Z to be the missing or the censored observations, so that $\pi_{\theta|Z}$ is easy to calculate. Successive substitution sampling then gives a random observation with distribution (approximately) $\mathcal{L}(\theta, Z \mid Y)$, and retaining the first coordinate gives an observation with distribution (approximately) equal to $\mathcal{L}(\theta \mid Y)$.

Another application arises when the parameter θ is high dimensional, and we are in a nonconjugate situation. Let us write $\theta = (\theta_1, \ldots, \theta_k)$, so that what we wish to obtain is $\pi_{\theta_1,\ldots,\theta_k}$. Direct calculation of the posterior will involve the evaluation of a k-dimensional integral, which may be difficult to accomplish. On the other hand, application of the successive substitution sampling algorithm involves the generation of one-dimensional random variables from $\pi_{\theta_i|\{\theta_j,\ j\neq i\}}$, which is available in closed form, except for a normalizing constant. There exist very efficient algorithms for doing this; see Zaman (1992).

Let us now return to the Markov chain simulation method. Let P be a transition probability function on the measurable space $(\mathcal{X}, \mathcal{B})$, i.e. P is a function on $\mathcal{X} \times \mathcal{B}$ such that for each $x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$, and for each $C \in \mathcal{B}$, $P(\cdot, C)$ is a measurable function on $(\mathcal{X}, \mathcal{B})$. Let $X_0, X_1, \ldots$ be a Markov chain with transition probability function P, i.e. $P(X_n \in C \mid X_{n-1} = x) = P(x, C)$, for $n = 1, 2, \ldots$. If $X_0 \equiv x$, we will say that the Markov chain starts at x and for any event C, $P(C \mid X_0 = x)$ will be denoted by $P_x(C)$. Similarly, for any bounded measurable function f defined on the Markov chain, $E(f \mid X_0 = x)$ will be denoted by $E_x(f)$. Let $P^n(x, C) = P(X_n \in C \mid X_0 = x)$. Suppose that π is a probability measure on $(\mathcal{X}, \mathcal{B})$ and is a stationary probability measure for the Markov chain, i.e. it satisfies (1.1).

When $\sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0$ ($\sup_{C \in \mathcal{B}} |\frac{1}{n} \sum_{j=1}^{n} P^j(x, C) - \pi(C)| \to 0$) for a set of starting points x which has probability 1 with respect to the stationary measure π, we will say that the Markov chain is *ergodic* (*mean ergodic*). The objective of this paper is to give theorems, whose conditions are very simple to check, and which guarantee ergodicity or mean ergodicity. (These are the minimum conditions required for success of Markov chain simulation method. Once such conditions are established, it is useful to also make a statement on the rate of convergence, if this is possible.) Before stating these theorems, we will need a few definitions concerning Markov chains. For any set $C \in \mathcal{B}$, let $N_n(C) = \sum_{m=1}^{n} I(X_m \in C)$ and $N(C) = \sum_{m=1}^{\infty} I(X_m \in C)$ be the number of visits to C by time n and the total number of visits to C, respectively. The expectations of $N_n(C)$ and $N(C)$, when the chain starts at x, are given by $G_n(x, C) = \sum_{m=1}^{n} P^m(x, C)$ and $G(x, C) = \sum_{m=1}^{\infty} P^m(x, C)$, respectively. Define $T(C) = \inf\{n : n > 0,\ X_n \in C\}$ to be the first time the chain hits C, after time 0. Note that $P_x(T(C) < \infty) > 0$ is equivalent to $G(x, C) > 0$.

The set $A \in \mathcal{B}$ is said to be *accessible* if

$$P_x(T(A) < \infty) > 0 \quad \text{for all } x \in \mathcal{X}.$$

Let ρ be a probability measure on $(\mathcal{X}, \mathcal{B})$. The Markov chain is said to be *ρ-irreducible* if every set A with $\rho(A) > 0$ is accessible. The set A is said to be *recurrent* if

$$P_x(T(A) < \infty) = 1 \quad \text{for all } x \in \mathcal{X}.$$

For the case where the σ-field $\mathcal{B}$ is separable, there is a very useful equivalent definition of ρ-irreducibility of a Markov chain. In this case, we can deduce from Theorem 2.1 of Orey (1971), on the existence of "C-sets", that ρ-irreducibility of a Markov chain implies that there exist a set $A \in \mathcal{B}$ with $\rho(A) > 0$, an integer n_0, and a number $\epsilon > 0$ satisfying

$$P_x(T(A) < \infty) > 0 \quad \text{for all } x \in \mathcal{X}, \tag{1.2}$$

and

$$x \in A,\ C \subset A \quad \text{imply} \quad P^{n_0}(x, C) \geq \epsilon \rho(C). \tag{1.3}$$

Let $\rho_A(C) = \frac{\rho(C \cap A)}{\rho(A)}$. This is well defined because $\rho(A) > 0$. The set function ρ_A is a probability measure satisfying $\rho_A(A) = 1$. Note that (1.2) simply states that A is an accessible set and this condition does not make reference to the probability measure ρ. Condition (1.3) states that uniformly in $x \in A$, the n_0-step transition probabilities from x into subsets of A are bounded below by ϵ times ρ. That (1.2) and (1.3) imply ρ_A-irreducibility is, of course, immediate. This alternative definition of ρ_A-irreducibility, which applies to nonseparable σ-fields as well, will be usually much easier to verify in Markov chain simulation problems. By replacing ρ by ρ_A, we can also assume with no loss of generality that ρ is a probability measure with $\rho(A) = 1$ when verifying Condition (1.3).

For any subset $\mathcal{M}$ of the positive integers, g.c.d.($\mathcal{M}$) will denote the greatest common divisor of the integers in $\mathcal{M}$.

The main results of this paper are the following two theorems, which are stated for general Markov chains. They give sufficient conditions for the Markov chain simulation method to be successful.

Theorem 1 *Suppose that the Markov chain $\{X_n\}$ with transition function $P(x, C)$ has an invariant probability measure π, i.e. (1.1) holds. Suppose that there is a set $A \in \mathcal{B}$, a probability measure ρ with $\rho(A) = 1$, a constant $\epsilon > 0$, and an integer $n_0 \geq 1$ such that*

$$\pi\Big\{x : P_x(T(A) < \infty) > 0\Big\} = 1, \tag{1.4}$$

and

$$P^{n_0}(x, \cdot) \geq \epsilon\rho(\cdot) \;\; \textit{for each } x \in A. \tag{1.5}$$

Suppose further that

$$g.c.d.\Big\{m \geq 1 : \textit{ there is an } \epsilon_m > 0 \textit{ such that } P^m(x, \cdot) \geq \epsilon_m\rho(\cdot) \;\; \textit{for each } x \in A\Big\} = 1. \tag{1.6}$$

Then there is a set D_0 such that

$$\pi(D_0) = 1 \;\; \textit{and} \;\; \sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \;\; \textit{for each } x \in D_0. \tag{1.7}$$

Theorem 2 *Suppose that the Markov chain $\{X_n\}$ with transition function $P(x, C)$ satisfies conditions (1.1), (1.4) and (1.5). Then*

$$\sup_{C \in \mathcal{B}} \Big|\frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r}(x, C) - \pi(C)\Big| \to 0 \;\; \textit{as } m \to \infty \;\; \textit{for } [\pi]\textit{-almost all } x, \tag{1.8}$$

and hence

$$\sup_{C \in \mathcal{B}} \Big|\frac{1}{n} \sum_{j=1}^{n} P^j(x, C) - \pi(C)\Big| \to 0 \;\; \textit{as } n \to \infty \;\; \textit{for } [\pi]\textit{-almost all } x. \tag{1.9}$$

Let $f(x)$ be a measurable function on $(\mathcal{X}, \mathcal{B})$ such that $\int \pi(dy)|f(y)| < \infty$. Then

$$P_x\Big\{\frac{1}{n} \sum_{j=1}^{n} f(X_j) \to \int \pi(dy)f(y)\Big\} = 1 \;\; \textit{for } [\pi]\textit{-almost all } x \tag{1.10}$$

and

$$\frac{1}{n} \sum_{j=1}^{n} E_x(f(X_j)) \to \int \pi(dy)f(y) = 1 \;\; \textit{for } [\pi]\textit{-almost all } x. \tag{1.11}$$

Variants of these theorems form a main core of interest in the Markov chain literature. However, most of this literature makes strong assumptions such as the existence of a recurrent set A and proves the *existence* of a stationary probability measure before establishing (1.7) and (1.8). Theorems 1 and 2 exploit the existence of a stationary probability measure, which is given to us "for free" in the Markov chain simulation method, and establish the ergodicity or mean ergodicity under minimal and easily verifiable assumptions. For example, we have already noted that in the context of the Markov chain simulation method, we really need to check only (1.4), (1.5), and (1.6). To show (1.4) in most cases one will establish that $P_x(T(A) < \infty) > 0$ for all x. Condition (1.6) is usually called the *aperiodicity* condition and is automatically satisfied if (1.5) holds with $n_0 = 1$. We also indicate the critical points in the proof where one can use additional information to obtain results on the rate of the convergence.

There is a long history on ergodic theorems for Markov chains. For the case where $\mathcal{X}$ is a finite space, it has long been known that if there is an integer k and a point $y \in \mathcal{X}$ such that $\min_{x\in\mathcal{X}} P^k(x, \{y\}) \geq \epsilon > 0$, then (1.7) holds at an exponential rate; see for instance p. 173 of Doob (1953). Other sufficient conditions for ergodicity are known for the case where $\mathcal{X}$ is a countable space. See e.g. Theorems 1.2 and 1.3 of Chapter 3 of Karlin and Taylor (1975).

In interesting problems, including those that arise in Bayesian statistics, the state space $\mathcal{X}$ generally is not countable. Early results on ergodicity of Markov chains on general state spaces used a condition known as the Doeblin condition; see Hypothesis (D′) on p. 197 of Doob (1953), which can be stated in an equivalent way as follows. There is a probability measure ϕ on $(\mathcal{X}, \mathcal{B})$, an integer k, and an $\epsilon > 0$ such that

$$P^k(x, C) \geq \epsilon\phi(C) \text{ for all } x \in \mathcal{X} \text{ and for all } C \in \mathcal{B}.$$

This is a very strong condition. It implies that there exists a stationary probability measure to which the Markov chain converges at a geometric rate, from any starting point.

Theorem 3 *Suppose that the Markov chain satisfies the Doeblin condition. Then there exists a unique invariant probability measure π such that for all n*

$$\sup_C |P^n(x, C) - \pi(C)| \leq (1 - \epsilon)^{(n/k)-1} \text{ for all } x \in \mathcal{X}.$$

A proof of this theorem may be found on p. 197 of Doob (1953). The Doeblin condition, though easy to state, is very strong and rarely holds in the problems that appear in the class of applications we are considering. We note that it is equivalent to the conditions of Theorem 1, with the set A of Theorem 1 replaced by $\mathcal{X}$. In its absence, one has to to impose the obvious conditions of irreducibility and aperiodicity and some other extra conditions, often times recurrence, to obtain ergodicity. Standard references in this area are Orey (1971), Revuz (1975) and Nummelin (1984). An exposition suitable for our purposes can be found in Athreya and Ney (1978). Theorem 4.1 of that paper may be stated as follows.

Theorem 4 *Suppose that there is a set $A \in \mathcal{B}$, a probability measure ρ concentrated on A, and an ϵ with $0 < \epsilon < 1$ such that*

$$P_x(T(A) < \infty) = 1 \text{ for all } x \in \mathcal{X},$$

and

$$P(x, C) \geq \epsilon\rho(C) \text{ for all } x \in A \text{ and all } C \in \mathcal{B}.$$

Suppose further that there is an invariant probability measure π. Then

$$\sup_{C\in\mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \text{ for all } x \in \mathcal{X}.$$

This theorem establishes ergodicity under the assumption of the existence of a stationary probability measure but also makes the strong assumption of the existence of a recurrent set A. It is always difficult to check that a set A is recurrent. Our main results, Theorems 1 and 2, weaken this recurrence condition to just the accessibility of the set A from $[\pi]$-almost all starting points x. We believe that this makes it routine to check the conditions of our theorem in Markov chain simulation problems.

Tierney (1991) gives sufficient conditions for convergence of Markov chains to their stationary distribution. The main part of his Theorem 1 may be stated as follows.

Theorem 5 *Suppose that the chain has invariant probability measure π. Assume that the chain is π-irreducible and aperiodic. Then (1.7) holds.*

The main difference between Theorems 1 and 5 is that in Theorem 1 the probability measure with respect to which irreducibility needs to be verified is not restricted to be the stationary measure. This distinction is more than cosmetic. To check π-irreducibility, one has to check that a certain condition holds for all sets which have positive probability under the stationary distribution. For certain Markov chain simulation problems in which the state space is very complicated, it is difficult or impossible to even identify these sets, since it is difficult to get a handle on the unknown π. An example of such a situation arose in the context of Bayesian nonparametrics in Doss (1991), where the state space was the set of all distribution functions. In that paper, the author was not able to get enough of a handle on π to identify those sets to which it gives positive probability. On the other hand, a convenient choice of ρ made it possible to check ρ-irreducibility through the equivalent Conditions (1.4) and (1.5).

Another application of Theorem 1 in the area of Bayesian nonparametrics appears in Escobar and West (1991).

We point out that Tierney (1991) does not give a detailed definition of aperiodicity, but refers the reader to Chapter 2.4 of Nummelin (1984) where an implicit definition of the period of a Markov chain is given. In the present paper, aperiodicity as constructively defined in (1.6), is extremely easy to check: If the n_0 appearing in (1.5) is 1, then (1.6) is automatic.

Results which give not only convergence of the Markov chain to its stationary distribution but also convergence at a geometric rate are obviously extremely desirable. Such results are given in Theorem 1 of Schervish and Carlin (1990) and in Proposition 1 of Tierney (1991). It is, however, important to keep in mind that checking conditions that ensure convergence at a geometric rate is usually an order of magnitude more difficult than checking the conditions needed for simple convergence, for example Theorems 1 and 5 in the present paper. This is because in cases where the dimension of the state space of the Markov chain is very high, it is usually extremely difficult to check the integrability conditions needed. This situation arises in Bayesian nonparametrics for example; see Doss (1991) for an illustration.

In addition, the Markov chain may converge but not at a geometric rate. This can happen even in very simple situations. An illustration is provided in the example below, which is due to T. Sellke. Let U be a random variable on $\mathbb{R}$ with distribution ν which we take to be the standard Cauchy distribution. Let the conditional distribution of V given U be the Beta distribution with parameters 2 and 2, shifted so that it is centered at U, and let $X = (U, V)$. If we start successive substitution sampling at $X_0 = (0,0)$, then it is easy to see that U_1 must be in the interval $(-1,1)$, and in fact, the value of U can change by at most one unit at each iteration. Thus, the distribution of U_n is concentrated in the interval $(-n,n)$. In particular,

$$\sup_{C\in\mathcal{B}}\left|P(U_n \in C \mid U_0 = 0) - \nu(C)\right| \geq \nu\{(-\infty,-n)\cup(n,\infty)\} \sim \left(\frac{2}{\pi}\right)\frac{1}{n},$$

so that the rate of convergence cannot be geometric. The distribution ν could have been taken to be any distribution whose tails are "thicker than those of the exponential distribution", and in fact, we can make the rate of convergence arbitrarily slow by taking the tails of ν to be sufficiently thick.

It is not difficult to see that if we select the starting point at random from a bounded density concentrated in a neighborhood of the origin, then this example provides a simple counterexample to Theorem 3 of Tanner and Wong (1987), which asserts convergence at a geometric rate.

This paper is organized as follows. Section 2 gives the proofs of Theorems 1 and 2 and also states and proves a theorem that gives, under additional conditions, convergence at a geometric rate. Section 3 discusses briefly some issues to consider when deciding how to use the output of the Markov chain to estimate π and functionals of π.

2 Ergodic Theorems for Markov Chains on General State Spaces

The proofs of Theorems 1 and 2 rest on the familiar technique of regenerative events in a Markov chain. See for instance Athreya and Ney (1978). In Section 2.1, we prove Proposition 1 in which we assume that the set A is a singleton α, so that ρ is the degenerate probability measure on $\{\alpha\}$. We also assume that the singleton α is an aperiodic state, a condition which is stated more fully as Condition (C) in Proposition 1 below. Under these simplified assumptions we establish the ergodicity of the Markov chain.

In Section 2.2 we establish Theorem 1 as follows. In Proposition 2 we show that, when $n_0 = 1$, under the conditions of Theorem 1, a general Markov chain can be reduced to one satisfying the above simplified assumptions of Proposition 1. This is done by enlarging the state space with an extra point Δ and extending the Markov chain to the enlarged space. We then show that this singleton set $\{\Delta\}$ satisfies the simplified assumptions of Proposition 1. From this it follows that the extended chain is ergodic. After this step we deduce that the original chain is also ergodic. Finally, we show how the condition $n_0 = 1$ can be discarded under the aperiodicity condition (1.6).

In Section 2.3 we prove Theorem 2 which asserts convergence of averages of transition functions and averages of functions of the Markov chain, without the aperiodicity assumption (1.6). The key step in the proof is to recognize that the Markov chain observed at time points which are multiples of n_0 is an embedded Markov chain satisfying the conditions of Proposition 2 and with a stationary probability distribution π_0 which is the restriction of π to the set A_0 defined by (2.30). In the Markov chain literature, mean ergodicity is usually obtained as an elementary consequence of ergodicity in the aperiodic case and the existence of a well-defined period and cyclically moving disjoint subclasses. Our proof circumvents, in a way which we believe is new, the need for well-defined periodicity and cyclically moving disjoint subclasses.

2.1 State Spaces with a Distinguished Point

Fix a point α in $\mathcal{X}$. For convenience, we will refer to this point as the distinguished point. We will often write just α for the singleton set $\{\alpha\}$. The number of visits to α, $N_n(\{\alpha\})$ and $N(\{\alpha\})$, will be denoted simply by N_n and N, respectively. The first time the chain visits α after time 0, namely $T(\{\alpha\})$, will be denoted simply by T. Let

$$C_0 = \Big\{x : P_x(T < \infty) = 1\Big\} = \Big\{x : P_x(T = \infty) = 0\Big\} \tag{2.1}$$

and

$$\mathcal{X}_0 = \Big\{x : P_x(T < \infty) > 0\Big\} \tag{2.2}$$

be the set of all states x from which α can be reached with probability 1 and the set of all states from which α is accessible, respectively.

Definition 1 The state α is said to be *transient* if $P_\alpha(T < \infty) < 1$ and *recurrent* if $P_\alpha(T < \infty) = 1$. The state α is said to be *positive recurrent* if $E_\alpha(T) < \infty$.

Proposition 1 *Suppose that the transition function $P(x, C)$ satisfies the following conditions:*

(A) π is a stationary probability measure for P.

(B) $\pi\Big\{x : P_x(T < \infty) > 0\Big\} = 1$.

Then

$$\pi(C_0) = 1 \quad \text{and} \quad \sup_{C\in\mathcal{B}} \Big|\frac{1}{n}\sum_{j=0}^{n} P^j(x, C) - \pi(C)\Big| \to 0 \quad \text{for each } x \in C_0. \tag{2.3}$$

Suppose in addition that

(C) g.c.d.$\{n : P^n(\alpha, \alpha) > 0\} = 1$.

Then

$$\sup_{C\in\mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \text{for each } x \in C_0. \tag{2.4}$$

The proof of this proposition is given after the remark following the proof of Lemma 3.

Lemma 1 *If Conditions (A) and (B) of Proposition 1 hold, then $\pi(\alpha) > 0$ and α is positive recurrent.*

Proof We first establish that $\pi(\alpha) > 0$. From Condition (A) it follows that $\pi(\alpha) = \int \pi(dx)P^n(x, \alpha)$ for $n = 1, 2, \ldots$, and hence

$$n\pi(\alpha) = \int \pi(dx)G_n(x, \alpha) \tag{2.5}$$

for all n. The Monotone Convergence Theorem and Condition (B) imply that

$$\lim n\pi(\alpha) = \int \pi(dx)G(x, \alpha) > 0, \tag{2.6}$$

and hence $\pi(\alpha) > 0$.

Let the Markov chain start at some $x \in \mathcal{X}$. Let $T_1 = T$ and $T_k = \inf\{n : n > T_{k-1},\ X_n = \alpha\}$ for $k = 2, 3, \ldots$, with the usual convention that the infimum of the empty set is ∞. If $N < \infty$ then only finitely many T_k's are finite. If $N = \infty$ then all the T_k's are finite. In the latter case, the Markov chain starts afresh from α at time T_k, and hence $T_k - T_{k-1},\ k = 2, 3, \ldots$ are independent and identically distributed with distribution H where $H(n) = P_\alpha(T \leq n)$. These facts, the Strong Law of Large Numbers and the inequality

$$\frac{N_n}{T_{N_n} + 1} \leq \frac{N_n}{n} \leq \frac{N_n}{T_{N_n}} \tag{2.7}$$

imply that

$$\frac{1}{n} N_n \to \frac{1}{E_\alpha(T)} I(N = \infty) \tag{2.8}$$

with probability 1 under P_x for each $x \in \mathcal{X}$.

From the Bounded Convergence Theorem, it follows that

$$\frac{1}{n} G_n(x, \alpha) = E_x\Big(\frac{1}{n} N_n\Big) \to \frac{1}{E_\alpha(T)} P_x(N = \infty) \quad \text{for each } x \in \mathcal{X}.$$

Divide both sides of (2.5) by n, take limits and compare with the above. By using the fact that π is a probability measure and applying the Bounded Convergence Theorem, we obtain

$$\pi(\alpha) = \frac{1}{E_\alpha(T)} \int \pi(dx) P_x(N = \infty). \tag{2.9}$$

Since $\pi(\alpha) > 0$, it follows that $\int \pi(dx) P_x(N = \infty) > 0$ and $E_\alpha(T) < \infty$, and hence α is positive recurrent. ◇

The arguments leading to the conclusion $\pi(\alpha) > 0$ in the above lemma, which were based on (2.5) and (2.6), and did not use the full force of Condition (B). The following corollary records that fact and will be used later in this paper.

Corollary 1 *Let π satisfy (A) of Proposition 1 and let $E \in \mathcal{B}$ be such that*

$$\pi\Big(\{x : G(x, E) > 0\}\Big) > 0.$$

Then $\pi(E) > 0$.

The fact that α is positive recurrent gives us a way of obtaining an explicit form for a finite stationary measure ν and show that it must be a multiple of π.

Lemma 2 *Let α be recurrent. Let*

$$\nu(C) = E_\alpha\Big(\sum_{j=0}^{T-1} I(X_j \in C)\Big) = \sum_{n=0}^{\infty} P_\alpha(X_n \in C,\ T > n) \tag{2.10}$$

be the expected number of visits to C between consecutive visits to α, beginning from α. Then ν is a stationary measure for $P(\cdot, \cdot)$ with $\nu(\mathcal{X}_0^c) = 0$, and is unique up to a multiplicative constant; more precisely,

$$\nu(\cdot) = \int P(x, \cdot) \nu(dx),$$

and if ν' is any other stationary measure with $\nu'(\mathcal{X}_0^c) = 0$, then

$$\nu'(C) = \nu'(\alpha)\nu(C) \quad \text{for all } C \in \mathcal{B}.$$

The measure ν also has the property

$$\nu(C_0^c) = 0.$$

Suppose that Conditions (A) and (B) of Proposition 1 hold, so that α is positive recurrent and π is a stationary probability measure for $P(\cdot,\cdot)$ with $\pi(\mathcal{X}_0^c) = 0$. Then

$$\nu(\mathcal{X}) = E_\alpha(T) < \infty$$

and

$$\pi(C) = \frac{\nu(C)}{E_\alpha(T)}$$

is the unique stationary probability measure with $\pi(C_0) = 1$.

Proof Since $\sum_{n=0}^{T-1} I(X_n = \alpha) = 1$ we have $\nu(\alpha) = 1 = P_\alpha(T < \infty)$. To show that $\nu(C_0^c) = 0$, notice that for all n

$$0 = P_\alpha(T = \infty) = E_\alpha\Big(P_\alpha(T = \infty \mid X_1, X_2, \ldots, X_n)\Big) = E_\alpha\Big(P_{X_n}(T = \infty)I(T > n)\Big).$$

From this it follows that

$$0 = P_\alpha\Big\{P_{X_n}(T = \infty)I(T > n) > 0\Big\} = P_\alpha\{X_n \in C_0^c,\, T > n\}$$

for each n. From the definition of ν in (2.10) it now follows that $\nu(C_0^c) = 0$. We now show that ν is a stationary measure. Let $f(x)$ be a bounded measurable function on $(\mathcal{X}, \mathcal{B})$. Then

$$\begin{aligned}
\int \nu(dx)f(x) &= \sum_{n=0}^{\infty} E_\alpha\Big(f(X_n)I(T > n)\Big) \\
&= f(\alpha) + \sum_{n=1}^{\infty}\Big(E_\alpha\Big(f(X_n)I(T > n-1)\Big) - E_\alpha\Big(f(X_n)I(T = n)\Big)\Big) \\
&= f(\alpha) + \sum_{n=1}^{\infty} E_\alpha\Big(E_\alpha\Big(f(X_n)I(T > n-1)\Big) \Bigm| X_0, X_1, \ldots, X_{n-1}\Big) \\
&\quad - \sum_{n=1}^{\infty} E_\alpha\Big(f(X_n)I(T = n)\Big) \\
&= f(\alpha) - f(\alpha)P_\alpha(T < \infty) + \sum_{n=1}^{\infty} E_\alpha\Big(E_{X_{n-1}}(f(X_n))I(T > n-1)\Big) \\
&= \sum_{n=1}^{\infty} E_\alpha\Big(\int P(X_{n-1}, dy)f(y)I(T > n-1)\Big) \\
&= \sum_{n=0}^{\infty} E_\alpha\Big(\int P(X_n, dy)f(y)I(T > n)\Big) \\
&= \int_{y\in\mathcal{X}}\Big(\int_{x\in\mathcal{X}} \nu(dx)P(x, dy)\Big)f(y).
\end{aligned}$$

where the fourth equality in the above follows from the Markov property. This shows that ν is a stationary measure.

Let ν' be any other stationary measure for $P(\cdot,\cdot)$ satisfying $\nu'(\mathcal{X}_0^c) = 0$. Fix $C \in \mathcal{B}$. Then for C such that $\alpha \notin C$,

$$\begin{aligned}
\nu'(C) &= \int \nu'(dx)P(x,C) \\
&= \nu'(\alpha)P_\alpha(X_1 \in C) + \int_{x\neq\alpha} \nu'(dx)P_x(X_1 \in C) \\
&= \nu'(\alpha)P_\alpha(X_1 \in C) + \int_{y\in\mathcal{X}}\int_{x\neq\alpha} \nu'(dy)P(y,dx)P_x(X_1 \in C) \\
&= \nu'(\alpha)P_\alpha(X_1 \in C) + \int \nu'(dy)P_y(X_2 \in C,\, T > 1) \\
&\vdots \\
&= \nu'(\alpha)\sum_{m=1}^{n} P_\alpha(X_m \in C,\, T > m-1) + \int \nu'(dy)P_y(X_{n+1} \in C,\, T > n) \\
&\geq \nu'(\alpha)\sum_{m=1}^{n} P_\alpha(X_m \in C,\, T > m-1) \\
&\geq \nu'(\alpha)\sum_{m=1}^{n} P_\alpha(X_m \in C,\, T > m)
\end{aligned}$$

for each n. In the last line above we used the fact that $\{X_m \in C,\ T > m-1\} = \{X_m \in C,\ T > m\}$, since $\alpha \notin C$. Thus $\nu'(C) \geq \nu'(\alpha)\nu(C)$ for all C since $\nu(\alpha) = 1$. Let $\lambda(C) = \nu'(C) - \nu'(\alpha)\nu(C)$. Then λ is a stationary nonnegative measure and $\lambda(\alpha) = 0$ since $\nu(\alpha) = 1$. Thus

$$0 = \lambda(\alpha) = \int G_n(x,\alpha)\lambda(dx) \to \int G(x,\alpha)\lambda(dx)$$

by the Monotone Convergence Theorem. Therefore $0 = \lambda(\mathcal{X}_0)$ since the integrand above, $G(x,A)$, is positive for all $x \in \mathcal{X}_0$ in view of Condition (B). This proves that

$$\nu'(C) = \nu'(\alpha)\nu(C), \tag{2.11}$$

which shows that ν is the unique stationary measure satisfying $\nu(\mathcal{X}_0^c) = 0$, up to a multiplicative constant.

We now assume that α is positive recurrent. Since $\sum_{n=0}^{T-1} I(X_n \in \mathcal{X}) = T$, we have $\nu(\mathcal{X}) = E_\alpha(T) < \infty$. Let π be a stationary probability measure satisfying $\pi(\mathcal{X}_0^c) = 0$. From (2.11), we have the equality

$$\pi(C) = \pi(\alpha)\nu(C).$$

From the earlier part of this proof it now follows that π is the unique stationary probability measure. ◇

One can consider general measurable functions $f(x)$ with $\int |f(y)|\pi(dy) < \infty$, instead of $I(x = \alpha)$ as was done in Lemmas 1 and 2. By reworking inequalities (2.7) and (2.8), showing that the end effects can be ignored and by using the law of large numbers for averages of i.i.d. random variables, we can obtain the following corollary.

Corollary 2 *Let Conditions (A) and (B) of Proposition 1 hold. Let $f(x)$ be a measurable function with $\int |f(y)|\pi(dy) < \infty$. Let*

$$A_f = \Big\{x : P_x\Big\{\frac{1}{n}\sum_{1\leq j\leq n} f(X_j) \to \int f(x)d\pi(x)\Big\} = 1\Big\}.$$

Then

$$\pi(A_f) = 1$$

and

$$\frac{1}{n}\sum_{j=1}^{n} E_x(f(X_j)) \to \int \pi(dx)f(x) \text{ for } [\pi]\text{-almost all } x.$$

Proof Using the definitions of the hitting times $\{T_k\}$ of α defined in Lemma 1, define

$$U = \sum_{j=1}^{\min(n,T_1)} f(X_j), \quad V_r = \sum_{j=T_r+1}^{T_r+1} f(X_j), \quad V_r^* = \sum_{j=T_r+1}^{T_r+1} |f(X_j)| \text{ and } W = \sum_{j=T_{N_n}+1}^{n} f(X_j).$$

From the simple bounds

$$|U| \leq \sum_{j=1}^{T_1} |f(X_j)| \quad \text{and} \quad |W| \leq V_{N_n+1}^* \leq \max_{1\leq r\leq N_{n+1}} V_r^*$$

and the fact that for each $x \in C_0$, under P_x, $V_1, V_2, \ldots$ are i.i.d. random variables with mean $E_\alpha(T)\int f(x)\pi(dx)$, and $E_x(T_1) < \infty$, we get $P_x(U/n \to 0) = P_x(W/n \to 0) = 1$. Thus,

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) = \frac{U}{n} + \frac{N_n}{n}\frac{1}{N_n}\sum_{r=1}^{N_n} V_r + \frac{W}{n} \to \int f(x)\pi(dx)$$

as $n \to \infty$ for $[\pi]$-almost all x. ◇

To get the convergence assertions (2.3) and (2.4) of Proposition 1 we need the following lemma from renewal theory.

Lemma 3 *Let $\{p_n, n = 0,1,\ldots\}$ be a probability distribution with $p_0 = 0$ and let $\mu = \sum_{n=1}^{\infty} p_n < \infty$. Let $\{\eta_i,\ i = 1,2,\ldots\}$ be a sequence of i.i.d. random variables with distribution $\{p_n\}$. Let $S_0 = 0$, $S_k = \sum_{n=j}^{k} \eta_j$ for $n \geq 1$. Define $\{p_n^{(k)}, n = 1,2,\ldots\}$, $k = 1,2,\ldots$ recursively by $p_n^{(1)} = p_n$, $p_n^{(k)} = \sum_{0\leq j\leq n} p_j^{(k-1)} p_{n-j} = P(S_k = n)$. For $n = 0,1,\ldots,$ define*

$$r_n = \sum_{k=0}^{\infty} p_n^{(k)}. \tag{2.12}$$

Then

(a) r_n is the unique solution of the so-called renewal equations

$$r_0 = 1,\ r_n = \sum_{j=1}^{n} r_{n-j}p_j, \quad n = 1,2,\ldots.$$

Furthermore,

(b) $\frac{1}{n}\sum_{j=0}^{n} r_j \rightarrow \frac{1}{\mu}$ *as* $n \rightarrow \infty$.
If the additional condition g.c.d.$\{n : p_n > 0\} = 1$ *holds, then*
(c) $r_n \rightarrow \frac{1}{\mu}$ *as* $n \rightarrow \infty$.

Proof It is easy to establish (a) by direct verification. To prove Part (b), we note that $\sum_{j=0}^{n} r_j = \sum_{k=0}^{\infty} P(S_k \leq n) = E(N(n))$ where $N(n) = \sup\{k : S_k \leq n\}$. By the Strong Law of Large Numbers and the inequalities

$$S_{N(n)} \leq n < S_{N(n)+1},$$

it follows that

$$\frac{N(n)}{n} \rightarrow \frac{1}{\mu} \text{ w.p. 1.}$$

Part (c) is the well known discrete renewal theorem for which there are many proofs in standard texts, some of which are purely analytic (see, e.g. Chapter XIII.10 in Feller (1950)) and others are probabilistic (see e.g. Chapter 2 of Hoel, Port, and Stone (1972)). ◇

Remark 1 The tail behavior of the probability distribution $\{p_n\}$ affects the rate of convergence of $|r_n - \frac{1}{\mu}|$. Here is an example of a result on rates of convergence. The following are equivalent:

$$\sum \exp(nt_0)\, p_n < \infty \text{ for some } t_0 > 0. \tag{2.13}$$

$$\sum \exp(nt_0)\, |r_n - r_{n+1}| < \infty \text{ for some } t_0 > 0. \tag{2.14}$$

$$\left|r_n - \frac{1}{\mu}\right| = O(\rho^n). \tag{2.15}$$

When these conditions hold, it can be asserted that $\exp(-t_0) < \rho < 1$. Similarly, if $\sum n^p p_n < \infty$ for some $p > 0$ then it is known that there is a θ with $0 < \theta < p$ such that

$$\left|r_n - \frac{1}{\mu}\right| = O(n^{-\theta}).$$

See e.g. Asmussen (1987) or Stone (1965).

Proof of Proposition 1 Let D be the collection of all measurable functions f on $(\mathcal{X}, \mathcal{B})$ with $\sup_y |f(y)| \leq 1$. Let $f \in D$. Then for any $x \in \mathcal{X}$,

$$E_x(f(X_n)) = E_x\Big(f(X_n)I(T > n)\Big) + \sum_{k=0}^{n} P_x(T = k)E_\alpha(f(X_{n-k})), \quad n = 0, 1, \ldots. \tag{2.16}$$

Let $v_n = E_\alpha(f(X_n))$, $a_n = E_\alpha\Big(f(X_n)I(T > n)\Big)$ and $p_n = P_\alpha(T = n)$, $n = 0, 1, \ldots$. Note that v_n and a_n also depend on the function f while p_n does not. Putting $x = \alpha$ in (2.16) we get the important identity

$$v_n = a_n + \sum_{k=0}^{n} p_k v_{n-k}. \tag{2.17}$$

It is not difficult to check that $v_n = \sum_{k=0}^n a_k r_{n-k}$ is the unique solution to (2.17) where r_n is as defined in (2.12). Thus

$$\frac{1}{n}\sum_{j=0}^{n} v_j = \frac{1}{n}\sum_{j=0}^{n}\sum_{k=0}^{j} a_k r_{j-k} = \frac{1}{n}\sum_{k=0}^{n} a_k R_{n-k} = \sum_{k=0}^{\infty} a_k \frac{R_{n-k}}{n} I(k \le n)$$

where $R_n = \sum_{j=0}^n r_j$. Also,

$$\frac{1}{\mu}\sum_{j=0}^{\infty} a_j = \frac{E_\alpha\left(\sum_{j=0}^{T-1} f(X_j)\right)}{E_\alpha(T)} = \frac{\int f d\nu}{E_\alpha(T)} = \int f d\pi.$$

Thus for $f \in D$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{j=0}^{n} v_j - \int f d\pi\right| &\le \sum_{k=0}^{\infty} |a_k| \left|\frac{R_{n-k}}{n} I(k \le n) - \frac{1}{\mu}\right| \\
&\le 2\sum_{j=m}^{\infty} |a_j| + \left(\sum_{j=0}^{\infty} |a_j|\right) \sup_{n-m \le k \le n} \left|\frac{R_k}{n} - \frac{1}{\mu}\right| \\
&\le 2\sum_{j=m}^{\infty} P_\alpha(T > j) + (E_\alpha(T)) \sup_{n-m \le k \le n} \left|\frac{R_k}{n} - \frac{1}{\mu}\right|
\end{aligned}$$

for any positive integer m. Note that for fixed m, $\sup_{n-m\le k \le n} |\frac{R_k}{n} - \frac{1}{\mu}| \to 0$ as $n \to \infty$ from Part (b) of Lemma 3, and $\sum_{j=m}^{\infty} P_\alpha(T > j) \to 0$ as $m \to \infty$, since α is positive recurrent. By first fixing m and letting $n \to \infty$, and then letting $m \to \infty$, we get

$$\left|\frac{1}{n}\sum_{j=0}^{n} v_j - \int f d\pi\right| \to 0 \text{ uniformly in } f \text{ as } n \to \infty. \tag{2.18}$$

Let $x \in C_0$. Let $w_n = E_x(f(X_n))$, $b_n = E_x\big(f(X_n)I(T > n)\big)$ and $g_n = P_x(T = n)$. Note that for a fixed x, $b_n \to 0$ as $n \to \infty$, uniformly in f and that g_n is a probability sequence which does not depend on f. Using equation (2.16) once again, we see that w_n satisfies the equation

$$w_n = b_n + \sum_{k=0}^{n} g_k v_{n-k}, \quad n = 0, 1, \ldots. \tag{2.19}$$

Using (2.18), we conclude that

$$\frac{1}{n}\sum_{j=0}^{n} w_j = \frac{1}{n}\sum_{j=0}^{n} b_j + \sum_{k=0}^{n} g_k \frac{1}{n}\sum_{j=0}^{n-k} v_j \to \int f d\pi$$

uniformly in f as $n \to \infty$. This establishes (2.3) of Proposition 1.

We now use Condition (C). Under this assumption, g.c.d.$\{n : P^n(\alpha, \alpha) > 0\} = 1$, and thus g.c.d.$\{n : p_n > 0\} = 1$; see for instance see the lemma on p. 29 of Chung (1967). Thus, from Part (c) of Lemma 3 we have $r_n \to \frac{1}{\mu}$. Repeating the arguments leading to (2.18) and (2.19) with this stronger result on r_n, we see that $v_n \to \int f d\mu$ and $w_n \to \int f d\mu$ uniformly in $f \in D$. This proves conclusion (2.4) and completes the proof of Proposition 1. ◇

2.2 Proof of Theorem 1 for General Markov Chains

We will now establish Theorem 1 under the condition that the n_0 appearing in (1.5) is 1. This is stated as Proposition 2 below, and though it is technically weaker, its proof contains the heart of the arguments needed to establish Theorem 1.

Proposition 2 *Suppose that $A \in \mathcal{B}$ and let ρ be a probability measure on $(\mathcal{X}, \mathcal{B})$ with $\rho(A) = 1$. Suppose that the transition function $P(x, C)$ of the Markov chain $\{X_n\}$ satisfies (1.1), (1.4), and (1.5) where the n_0 appearing in (1.5) is equal to 1. Then there is a set D_0 such that*

$$\pi(D_0) = 1 \quad \textit{and} \quad \sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \textit{for each } x \in D_0. \tag{2.20}$$

Proof The proof consists of adding a point Δ to $\mathcal{X}$, defining a transition function on the enlarged space and appealing to Proposition 1.

Consider the space $(\bar{\mathcal{X}}, \bar{\mathcal{B}})$, where $\bar{\mathcal{X}} = \mathcal{X} \cup \{\Delta\}$ and $\bar{\mathcal{B}}$ is the smallest σ-field containing $\mathcal{B}$ and $\{\Delta\}$. Let $\epsilon^* = \epsilon/2$, and define the transition probability function $\bar{P}(x, C)$ on $(\bar{\mathcal{X}}, \bar{\mathcal{B}})$ by

$$\bar{P}(x, C) = \begin{cases} P(x, C) & \text{if } x \in \mathcal{X} \setminus A,\ C \in \mathcal{B} \\ P(x, C) - \epsilon^* \rho(C) & \text{if } x \in A,\ C \in \mathcal{B} \\ \epsilon^* & \text{if } x \in A,\ C = \{\Delta\} \\ \int_A \rho(dz) \bar{P}(z, C) & \text{if } x = \Delta,\ C \in \bar{\mathcal{B}} \end{cases} \tag{2.21}$$

Also, define the probability measure $\bar{\pi}$ on $(\bar{\mathcal{X}}, \bar{\mathcal{B}})$ by

$$\bar{\pi}(C) = \begin{cases} \pi(C) - \epsilon^* \rho(C) \pi(A) & \text{if } C \in \mathcal{B} \\ \epsilon^* \pi(A) & \text{if } C = \{\Delta\} \end{cases} \tag{2.22}$$

We will now show that the transition probability function $\bar{P}(x, C)$ together with $\bar{\pi}$ and the distinguished point Δ satisfy Conditions (A), (B), and (C) of Proposition 1.

If $x \in A$ then $\bar{P}(x, \Delta) = \epsilon^* > 0$, so that $\bar{G}(x, \Delta) > 0$. If $x \in \mathcal{X} \setminus A$, we have $\bar{G}(x, \Delta) \geq \int_A G(x, dy) \bar{P}(y, \Delta) \geq \epsilon^* G(x, A) > 0$ in view of (1.4). Finally $\bar{P}(\Delta, \Delta) = \epsilon^* > 0$. This verifies both Conditions (A) and (C) of Proposition 1.

Next, for $C \in \mathcal{B}$, we have

$$\begin{aligned} \int_{\bar{\mathcal{X}}} \bar{\pi}(dx) \bar{P}(x, C) &= \int_{\mathcal{X}} \big(\pi(dx) - \epsilon^* \rho(dx) \pi(A)\big) \bar{P}(x, C) + \epsilon^* \pi(A) \int_{\mathcal{X}} \rho(dx) \bar{P}(x, C) \\ &= \int_{\mathcal{X}} \pi(dx) \bar{P}(x, C) \\ &= \int_{\mathcal{X}} \pi(dx) \big(P(x, C) - \epsilon^* \rho(C) I(x \in A)\big) \\ &= \pi(C) - \epsilon^* \rho(C) \pi(A) \\ &= \bar{\pi}(C). \end{aligned}$$

When $C = \{\Delta\}$, we have

$$\begin{aligned} \int_{\bar{\mathcal{X}}} \bar{\pi}(dx) \bar{P}(x, \Delta) &= \int_{\mathcal{X}} \big(\pi(dx) - \epsilon^* \rho(dx) \pi(A)\big) \big(\epsilon^* I(x \in A)\big) + \epsilon^* \pi(A) \int_{\mathcal{X}} \rho(dx) \epsilon^* I(x \in A) \\ &= \epsilon^* \pi(A) \\ &= \bar{\pi}(\Delta). \end{aligned}$$

This verifies Condition (B) of Proposition 1.

Thus Proposition 1 implies that there exists a set $\bar{D}_0 \in \bar{\mathcal{B}}$ such that

$$\bar{\pi}(\bar{D}_0) = 1 \quad \text{and} \quad \sup_{C \in \bar{\mathcal{B}}} |\bar{P}^n(x, C) - \bar{\pi}(C)| \to 0 \quad \text{for each } x \in \bar{D}_0. \tag{2.23}$$

To translate (2.23) as a result for $P^n(x, C)$ we define a function $v(x, C)$ on $\bar{\mathcal{X}} \times \mathcal{B}$ by

$$v(x, C) = \begin{cases} I(x \in C) & \text{if } x \in \mathcal{X} \\ \rho(C) & \text{if } x = \Delta \end{cases}$$

We may view $v(x, C)$ as a transition function from $\bar{\mathcal{X}}$ into $\mathcal{X}$. The following lemma shows how one can go from $P^n(x, C)$ to $\bar{P}^n(x, C)$ and back. The proof of Proposition 2 is continued after Lemma 5.

Lemma 4 *The transition functions $P(x, C)$, $\bar{P}(x, C)$ and $v(x, C)$ and the probability measures π and $\bar{\pi}$ are related as follows:*

$$P(x, C) = \int_{\bar{\mathcal{X}}} \bar{P}(x, dy) v(y, C) \quad \textit{for } x \in \mathcal{X},\ C \in \mathcal{B}, \tag{2.24}$$

$$\bar{P}(x, C) = \int_{\mathcal{X}} v(x, dy) \bar{P}(y, C) \quad \textit{for } x \in \bar{\mathcal{X}},\ C \in \bar{\mathcal{B}}, \tag{2.25}$$

$$P^n(x, C) = \int_{\bar{\mathcal{X}}} \bar{P}^n(x, dy) v(y, C) \quad \textit{for } x \in \mathcal{X},\ C \in \mathcal{B}, \tag{2.26}$$

and

$$\pi(C) = \int_{\bar{\mathcal{X}}} \bar{\pi}(dx) v(x, C) \quad \textit{for } C \in \mathcal{B}. \tag{2.27}$$

Proof These are proved by direct verification. For $x \in \mathcal{X}$, $C \in \mathcal{B}$, we have

$$\begin{aligned} \int_{\bar{\mathcal{X}}} \bar{P}(x, dy) v(y, C) &= \int_{\mathcal{X}} \bar{P}(x, dy) I(y \in C) + \epsilon^* I(x \in A) \rho(C) \\ &= P(x, C) - \epsilon^* I(x \in A) \rho(C) + \epsilon^* I(x \in A) \rho(C) \\ &= P(x, C). \end{aligned}$$

Similarly, for $x \in \bar{\mathcal{X}}$, $C \in \bar{\mathcal{B}}$, we get

$$\int_{\mathcal{X}} v(x, dy) \bar{P}(y, C) = \begin{cases} \bar{P}(x, C) & \text{if } x \in \mathcal{X} \\ \int \rho(dy) \bar{P}(y, C) = \bar{P}(\Delta, C) & \text{if } x = \Delta \end{cases}$$

We prove (2.26) by induction on n. For $n = 1$, this is just (2.24). Assume that (2.26) has been proved for $n - 1$.

For $x \in \mathcal{X}$, $C \in \mathcal{B}$, we have

$$\begin{aligned} \int_{\bar{\mathcal{X}}} \bar{P}^n(x, dy) v(y, C) &= \int_{z, y \in \bar{\mathcal{X}}} \bar{P}^{n-1}(x, dz) \bar{P}(z, dy) v(y, C) \\ &= \int_{z, y \in \bar{\mathcal{X}},\, w \in \mathcal{X}} \bar{P}^{n-1}(x, dz) v(z, dw) \bar{P}(w, dy) v(y, C) \\ &= \int_{z \in \bar{\mathcal{X}},\, w \in \mathcal{X}} \bar{P}^{n-1}(x, dz) v(z, dw) P(w, C) \\ &= \int_{w \in \mathcal{X}} P^{n-1}(x, dw) P(w, C) \\ &= P^n(x, C), \end{aligned}$$

where the second inequality follows from (2.25), the third follows from (2.24), and the fourth from the induction step.

Finally, for $C \in \mathcal{B}$, we notice that

$$\int_{\mathcal{X}} \bar{\pi}(dx)v(x,C) = \int_{\mathcal{X}} \Big(\pi(dx) - \epsilon^*\pi(A)\rho(dx)\Big)v(x,C) + \epsilon^*\pi(A)\rho(C) = \pi(C).$$

This completes the proof of the lemma. ◇

The next lemma shows that $\bar{\pi}$ dominates ρ.

Lemma 5 *Let $C \in \mathcal{B}$. Then*

$$\bar{\pi}(C) = 0 \;\; \textit{implies that} \;\; \rho(C) = 0. \tag{2.28}$$

Proof From the careful choice of $\epsilon^* = \epsilon/2$ used to define $\bar{P}(x,C)$ in definition (2.21), we have

$$\bar{P}(x,C) = P(x,C) - \epsilon^*\rho(C) > \epsilon^*\rho(C) \;\; \text{whenever } x \in A \text{ and } C \in \mathcal{B}. \tag{2.29}$$

Applying Lemma 2 to the Markov chain $\{\bar{\mathcal{X}}, \bar{\mathcal{B}}, \bar{P}(\cdot,\cdot)\}$ which has a stationary distribution $\bar{\pi}(\cdot)$ we get, for any $C \in \mathcal{B}$,

$$\begin{aligned}
\bar{\pi}(C) &= \frac{1}{E_\Delta(\bar{T}_\Delta)} E_\Delta\Big(\sum_{n=0}^{\infty}\Big(I(\bar{X}_n \in C)I(\bar{T}_\Delta > n)\Big)\Big) \\
&= \frac{1}{E_\Delta(\bar{T}_\Delta)} E_\Delta\Big(\sum_{n=0}^{\infty}\Big(I(\bar{X}_n \in C)I(\bar{T}_\Delta > n-1)\Big)\Big) \\
&\geq \frac{1}{E_\Delta(\bar{T}_\Delta)} E_\Delta\Big(\sum_{n=1}^{\infty}\Big(I(\bar{X}_n \in C)I(\bar{T}_\Delta > n-1)I(\bar{X}_{n-1} \in A)\Big)\Big) \\
&= \frac{1}{E_\Delta(\bar{T}_\Delta)} E_\Delta\Big(\sum_{n=1}^{\infty}\Big(I(\bar{T}_\Delta > n-1)I(\bar{X}_{n-1} \in A)\bar{P}(\bar{X}_n \in C \mid X_{n-1})\Big)\Big) \\
&\geq \frac{1}{E_\Delta(\bar{T}_\Delta)} \epsilon^*\rho(C) E_\Delta\Big(\sum_{n=1}^{\infty}\Big(I(\bar{T}_\Delta > n-1)I(\bar{X}_{n-1} \in A)\Big)\Big) \\
&= \frac{1}{E_\Delta(\bar{T}_\Delta)} \epsilon^*\rho(C) E_\Delta\Big(\sum_{n=0}^{\infty}\Big(I(\bar{T}_\Delta > n)I(\bar{X}_n \in A)\Big)\Big) \\
&= \epsilon^*\rho(C)\bar{\pi}(A) \\
&= \epsilon^*\rho(C)\pi(A)(1 - \epsilon^*\rho(A)).
\end{aligned}$$

The equality in the third line follows from the fact that $\{\bar{X}_n \in C,\ \bar{T}_\Delta > n-1\} = \{\bar{X}_n \in C,\ \bar{T}_\Delta > n\}$, since $\Delta \notin C$, and the inequality in the fourth line follows from (2.29). Now since Δ is recurrent for the chain $\{\bar{\mathcal{X}}, \bar{\mathcal{B}}, \bar{P}(\cdot,\cdot)\}$, we have $\bar{\pi}(\Delta) = \epsilon^*\pi(A) > 0$, and this proves (2.28). ◇

Completion of the proof of Proposition 2 Let $D_0 = \bar{D}_0 - \Delta$. From (2.23), (2.26), and (2.27), we have $\bar{\pi}(\bar{D}_0) = 1$, and

$$\sup_{C\in\mathcal{B}} |P^n(x,C) - \pi(C)| = \sup_{C\in\mathcal{B}}\Big|\int_{\bar{\mathcal{X}}} \bar{P}^n(x,dy)v(y,C) - \int_{\bar{\mathcal{X}}} \bar{\pi}(dy)v(y,C)\Big| \to 0 \;\; \text{for each } x \in D_0.$$

This means that

$$\bar{\pi}(\mathcal{X} - D_0) = \bar{\pi}(\bar{\mathcal{X}} - \bar{D}_0) = 0.$$

From Lemma 5, it follows that

$$\rho(\mathcal{X} - D_0) = 0.$$

Now, from the definition of $\bar{\pi}(\cdot)$ in (2.22),

$$\pi(\mathcal{X} - D_0) = \bar{\pi}(\mathcal{X} - D_0) + \epsilon^* \rho(\mathcal{X} - D_0)\pi(A) = 0.$$

This completes the proof that $\pi(D_0) = 1$ and

$$\sup_{C \in \mathcal{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \text{for all } x \in D_0.$$

◇

We now drop the condition $n_0 = 1$ and prove Theorem 1.

Proof of Theorem 1 Let $\mathcal{M} = \left\{m : \text{there is an } \epsilon_m > 0 \text{ such that } \inf_{x \in A} P^m(x, \cdot) \geq \epsilon_m \rho(\cdot)\right\}$. Then g.c.d.$(\mathcal{M}) = 1$. Fix an $m \in \mathcal{M}$. From a standard result on g.c.d.'s of sets of positive integers (see e.g. Problem 2 on p. 77 of Karlin and Taylor (1975)), there is an integer L such that $\mathcal{M}$ will contain all integers larger than L. This together with Condition (1.4) shows that $\sum_{k \geq 1} P^{km}(x, A) > 0$ for each $x \in \mathcal{X}$. This means that the Markov chain viewed only at times which are multiples of m satisfies (1.4) and (1.5) with $n_0 = 1$. Thus from Proposition 2 there is a set D_0 such that $\pi(D_0) = 1$, and for any $m \in \mathcal{M}$,

$$\sup_{C \in \mathcal{B}} |P^{km}(x, C) - \pi(C)| \to 0 \quad \text{for } x \in D_0 \quad \text{as } k \to \infty.$$

Next, there is a finite subcollection $m_1, m_2, \ldots, m_r \in \mathcal{M}$ and integers $a_1, a_2, \ldots, a_r$ such that $\sum_{1 \leq i \leq r} a_i m_i = 1$. This is generally established during the proof of the standard fact on the g.c.d. of sets of integers quoted above (see e.g. Problem 2 on p 77 of Karlin and Taylor (1975)). Permute the indices if necessary and assume that $a_1 > 0, \ldots, a_s > 0$, $-a_{s+1} = b_{s+1} > 0, \ldots, -a_r = b_r > 0$, so that $N - M = 1$ where $N = \sum_{1 \leq i \leq s} a_i m_i$ and $M = \sum_{s < i \leq r} b_i m_i > 0$. Any positive integer K can be written as $K = k(M^2 + N) + r$ where $0 \leq r < M^2 + N$ and $k \geq 0$. Writing $r = r(N - M)$ we have

$$K = N(k + r) + M(kM - r) = (k + r) \sum_{1 \leq i \leq s} a_i m_i + (kM - r) \sum_{s < i \leq r} b_i m_i.$$

Note that $kM - r > 0$ when $k > 4M$. When $K \to \infty$, the integer k defined above tends to ∞, as do the multipliers $k + r$ and $kM - r$. Since for $[\pi]$-almost every x we have $\sup_{C \in \mathcal{B}} |P^{km_i}(x, C) - \pi(C)| \to 0$ as $k \to \infty$ for $i = 1, 2, \ldots r$, it follows that $|\int P^{km_i}(x, dy) f_k(y) - \int f(y)\pi(dy)| \to 0$ for $i = 1, 2, \ldots, r$ if $f_k(y) \to f(y)$ for $[\pi]$-almost every y. Therefore, as $K \to \infty$

$$\sup_{C \in \mathcal{B}} |P^K(x, C) - \pi(C)| =$$

$$\sup_{C \in \mathcal{B}} \left| \int P^{(k+r)a_1 m_1}(x, dy_1) P^{(k+r)a_2 m_2}(y_1, dy_2) \cdots P^{(kM-r)b_r m_r}(y_{r-1}, C) - \pi(C) \right| \to 0.$$

◇

Remark 2 The proof above also establishes the following slight extension of Theorem 1.

Theorem 1′ *Suppose that π is a stationary probability measure for a Markov chain with transition function $P(x,\cdot)$. Let $\mathcal{M}'$ be the set of all integers $m \geq 1$ such that there exist $\epsilon_m > 0$, $A_m \in \mathcal{B}$, and a probability measure ρ_m with $\rho(A_m) = 1$ such that*

$$\pi\Big\{x : P_x(T(A_m) < \infty) > 0\Big\} = 1,$$

$$P^m(x,\cdot) \geq \epsilon_m \rho_m(\cdot) \ \ \textit{for each } x \in A_m$$

and

$$\sum_{k\geq 1} P^{km}(x, A_m) > 0 \ \ \textit{for } [\pi]\textit{-almost every } x.$$

Then the conclusion (1.7) of Theorem 1 holds if g.c.d.$(\mathcal{M}') = 1$.

2.3 Proof of Theorem 2 for General Markov Chains

As mentioned earlier, the key to the proof of Theorem 2 is to recognize an embedded Markov chain which satisfies the conditions of Theorem 1. The proof of Theorem 2 is completed after Lemma 9.

Let $Y_m = X_{mn_0}$, $m = 0, 1, \ldots$ and put $Q(x,C) = P^{n_0}(x,C)$ for $x \in \mathcal{X}$ and $C \in \mathcal{B}$. The subsequence $\{Y_0, Y_1, \ldots\}$ is a Markov chain with transition probability function $Q(x,C)$ and we will call it the embedded Markov chain. Define

$$A_r = \left\{x : \sum_{m=1}^{\infty} P^{mn_0 - r}(x, A) > 0\right\}, \ \ r = 0, 1, \ldots, n_0. \tag{2.30}$$

Since $P^{n_0}(x, A) \geq \epsilon$ for all $x \in A$, one can also define A_r by

$$A_r = \left\{x : \sum_{m=k}^{\infty} P^{mn_0 - r}(x, A) > 0\right\} \ \ \text{for any } \ k \geq 1.$$

i.e. A_r is the set of all points from which A is accessible at time points which are of the form $mn_0 - r$ for all large m and A_0 is the set of all points from which A is accessible in the embedded Markov chain.

Lemma 6 below shows that the embedded Markov chain satisfies the conditions of Theorem 1 with the restriction of π to A_0 as its stationary probability measure.

Lemma 6 *Under the conditions of Theorem 2,*

$$\pi(A_0) > 0.$$

Let

$$\pi_0(C) = \frac{\pi(C \cap A_0)}{\pi(A_0)}.$$

The embedded Markov chain $\{Y_0, Y_1, \ldots\}$ satisfies the conditions (1.4) and (1.5) of Theorem 1 with π_0 as a stationary probability measure and with the n_0 appearing in (1.5) equal to 1.

Proof Condition (1.1) states that $\pi\Big(\{x : P_x(T(A) < \infty) > 0\}\Big) = 1$. Just the fact that this probability is positive and condition (1.4) allow us to use Corollary 1 to conclude that $\pi(A) > 0$. Condition (1.5) implies that $A \subset A_0$. Thus $\pi(A_0) > 0$ and hence π_0 is a well defined probability measure. Clearly,

$$\pi(C) = \int \pi(dx)Q(x,C) \text{ for all } C \in \mathcal{B}, \tag{2.31}$$

$$\pi_0(A_0) = 1 \tag{2.32}$$

and

$$Q(x,\cdot) \geq \epsilon\rho(\cdot) \text{ for all } x \in A. \tag{2.33}$$

Notice that

$$\sum_{2\leq m<\infty} Q^m(x,A) = \int_{\mathcal{X}} Q(x,dy) \sum_{1\leq m<\infty} Q^m(y,A) = \int_{A_0} Q(x,dy) \sum_{1\leq m<\infty} Q^m(y,A).$$

Hence $Q(x,A_0) > 0$ implies that $\sum_{2\leq m<\infty} Q^m(x,A) > 0$, i.e. $x \in A_0$. In other words,

$$x \notin A_0 \text{ implies that } Q(x,A_0) = 0. \tag{2.34}$$

From (2.31) and (2.34) we have the equality

$$\pi(A_0) = \int_{\mathcal{X}} \pi(dx)Q(x,A_0) = \int_{A_0} \pi(dx)Q(x,A_0)$$

which implies that $Q(x,A_0) = 1$ for $[\pi]$-almost all $x \in A_0$. Hence

$$\int_{\mathcal{X}} \pi_0(dx)Q(x,C) = \frac{1}{\pi(A_0)} \int_{A_0} \pi(dx)Q(x,C) = \frac{1}{\pi(A_0)} \int_{\mathcal{X}} \pi(dx)Q(x,C\cap A_0) = \pi_0(C). \tag{2.35}$$

Equations (2.35), (2.32) and (2.33) establish the lemma. ◇

Define

$$\pi_r(C) = \int_{A_0} \pi_0(dx)P^r(x,C)$$

for $r = 1,2,\ldots,n_0 - 1$ and

$$\tilde{\pi}(C) = \frac{1}{n_0} \sum_{r=0}^{n_0-1} \pi_r(C).$$

Note that π_r is the distribution of X_r when $Y_0 = X_0$ has initial distribution π_0. The next lemma shows that averages of the transition functions of the embedded chain converge to $\tilde{\pi}(C)$ for $[\pi_0]$-almost all x.

Lemma 7 *Define*

$$B_0 = \Big\{x : x \in A_0, \sup_{C\in\mathcal{B}} |P^{mn_0}(x,C) - \pi_0(C)| \to 0 \text{ as } m \to \infty\Big\}. \tag{2.36}$$

Under the conditions of Theorem 2

$$\pi_0(B_0) = 1. \tag{2.37}$$

Moreover for each $x \in B_0$,

$$\sup_{C \in \mathcal{B}} |P^{mn_0+r}(x,C) - \pi_r(C)| \to 0 \quad \textit{as } m \to \infty \quad \textit{for } r = 0, 1, \ldots, n_0 - 1, \tag{2.38}$$

and hence

$$\sup_{C \in \mathcal{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r}(x,C) - \tilde{\pi}(C) \right| \to 0 \quad \textit{as } m \to \infty. \tag{2.39}$$

Proof From Lemma 6 the embedded Markov chain satisfies the conditions of Theorem 1 with the n_0 appearing in (1.5) equal to 1. From Proposition 2 it follows that

$$\sup_{C \in \mathcal{B}} |P^{mn_0}(x,C) - \pi_0(C)| \to 0 \text{ as } m \to \infty$$

for $[\pi_0]$-almost all x. This establishes (2.37). For $r = 0, 1, \ldots, n_0 - 1$ and $x \in B_0$,

$$\begin{aligned} \sup_{C \in \mathcal{B}} |P^{mn_0+r}(x,C) - \pi_r(C)| &= \sup_{C \in \mathcal{B}} \left| \int_{\mathcal{X}} (P^{mn_0}(x,dy) - \pi_0(dy)) P^r(y,C) \right| \\ &\leq \sup_{D \in \mathcal{B}} |P^{mn_0}(x,D) - \pi_0(D)| \to 0 \end{aligned}$$

as $m \to \infty$ establishing (2.38). ◇

The next lemma shows that the conclusions of the previous lemma hold $[\tilde{\pi}]$-almost everywhere.

Lemma 8 *Under the conditions of Theorem 2,*

$$\pi_r(A_r) = 1 \quad \textit{for } r = 1, \ldots, n_0 - 1$$

and (2.39) holds for $[\tilde{\pi}]$-almost all x.

Proof Consider the original Markov chain $X_0, X_1, \ldots$. Let $E \in \mathcal{B}$ and let $\pi_0(E) = 1$. Then

$$\begin{aligned} \int_{\mathcal{X}} \pi_1(dx) P^{n_0-1}(x,E) &= \int_{x \in \mathcal{X}} \int_{y \in A_0} \pi_0(dy) P(y,dx) P^{n_0-1}(x,E) \\ &= \int_{A_0} \pi_0(dy) P^{n_0}(y,E) \\ &= \pi_0(E) = 1 \end{aligned}$$

and hence $P^{n_0-1}(x,E) = 1$ for $[\pi_1]$-almost all x. In particular we take $E = B_0$ and rewrite the conclusion as $\pi_1(B_1) = 1$ where the sets B_r are defined by

$$B_r = \left\{ x : P^{n_0-r}(x,B_0) = 1 \right\}, \; r = 1, 2, \ldots, n_0 - 1.$$

Let $x \in B_1$. Then

$$\sum_{m \geq 2} P^{mn_0-1}(x,A) \geq \int_{B_0} P^{n_0-1}(x,dy) \sum_{m \geq 2} P^{(m-1)n_0}(y,A) > 0$$

and hence $x \in A_1$. Thus $B_1 \subset A_1$. Similarly, $\pi_r(B_r) = 1$ and $B_r \subset A_r$ for all r. Notice that for $x \in B_1$

$$\begin{aligned}
\sup_{C\in\mathcal{B}} |P^{mn_0+r+n_0-1}(x,C) - \pi_r(C)| &\leq \int_{y\in\mathcal{X}} \sup_{C\in\mathcal{B}} |P^{mn_0+r}(y,C) - \pi_r(C)| P^{n_0-1}(x,dy) \\
&= \int_{y\in B_0} \sup_{C\in\mathcal{B}} |P^{mn_0+r}(y,C) - \pi_r(C)| P^{n_0-1}(x,dy).
\end{aligned}$$

From (2.38) it follows that

$$\sup_{C\in\mathcal{B}} |P^{mn_0+r+n_0-1}(x,C) - \pi_r(C)| \to 0 \text{ for } [\pi_1]\text{-almost all } x \text{ as } m \to \infty. \tag{2.40}$$

As a consequence,

$$\sup_{C\in\mathcal{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r+n_0-1}(x,C) - \tilde{\pi}(C) \right| \to 0 \text{ for } [\pi_1]\text{-almost all } x \text{ as } m \to \infty. \tag{2.41}$$

Now

$$\begin{aligned}
\sup_{C\in\mathcal{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{(m+1)n_0+r}(x,C) - \tilde{\pi}(C) \right| &\leq \sup_{C\in\mathcal{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r+n_0-1}(x,C) - \tilde{\pi}(C) \right| \\
&\quad + \frac{1}{n_0} \sup_{C\in\mathcal{B}} \left| P^{(m+1)n_0+n_0-1}(x,C) - P^{mn_0+n_0-1}(x,C) \right|
\end{aligned}$$

and as $m \to 0$ this converges to 0 for $[\pi_1]$-almost all x, from (2.40) and (2.41). A similar argument shows that (2.39) holds for $[\pi_r]$-almost all x and all r and hence for $[\tilde{\pi}]$-almost all x. $\diamondsuit$

We now establish that $\pi = \tilde{\pi}$ by using the full force of condition (1.4).

Lemma 9 *Under the conditions of Theorem 2, π_r is the restriction of π to A_r, for $r = 1, 2, \ldots, n_0 - 1$ and*

$$\pi = \tilde{\pi}.$$

Proof We have already shown that $\pi_r(A_r) = 1$, $r = 1, 2, \ldots, n_0 - 1$. We will now show that $A_0, \ldots, A_{n_0-1}$ act like cyclically moving subsets in the sense that

$$x \in A_0^c \text{ implies that } P(x, A_1) = 0.$$

Suppose that $P(x, A_1) > 0$. Then

$$\sum_{m\geq 1} P^{mn_0}(x,A) \geq \int_{A_1} P(x,dy) \sum_{m\geq 1} P^{mn_0-1}(y,A) > 0,$$

which implies that $x \in A_0$. Thus $x \in A_0^c$ implies that $P(x, A_1) = 0$. Now for $C \in \mathcal{B}$,

$$\begin{aligned}
\pi_1(C) &= \pi_1(C \cap A_1) \\
&= \frac{1}{\pi_0(A_0)} \int_{A_0} \pi(dx) P(x, C \cap A_1) \\
&= \frac{1}{\pi_0(A_0)} \int_{\mathcal{X}} \pi(dx) P(x, C \cap A_1) \\
&= \frac{\pi(C \cap A_1)}{\pi(A_0)}.
\end{aligned}$$

Since $\pi_1(A_1) = 1$, this implies that $\pi(A_1) = \pi(A_0)$ and that π_1 is the restriction of π to A_1. A similar conclusion holds for π_r for other values of r.

We now use the full force of Condition (1.4) which can be restated as $\pi(\cup_{r=0}^{n_0-1} A_r) = 1$. This together with the fact that π_r is the restriction of π to A_r, $r = 0, 1, \ldots, n_0 - 1$ implies that the probability measures π and $\tilde{\pi}$ are absolutely continuous with respect to each other. From this observation and Lemma 8, for any $C \in \mathcal{B}$,

$$H_{mn_0}(x, C) = \frac{1}{mn_0} \sum_{j=1}^{mn_0} P^j(x, C) \to \tilde{\pi}(C)$$

for $[\pi]$-almost all x. Now,

$$\pi(C) = \int_{\mathcal{X}} \pi(dx) H_{mn_0}(x, C) \overset{m\to\infty}{\longrightarrow} \int_{\mathcal{X}} \pi(dx)\tilde{\pi}(C) \ = \tilde{\pi}(C).$$

This shows that $\pi = \tilde{\pi}$. ◇

We now complete the proof of Theorem 2

Proof of Theorem 2 It is clear that Lemmas 8 and 9 establish conclusions (1.8) and (1.9) of Theorem 2. Let $f(x)$ be a measurable function satisfying $\int |f(x)|\pi(dx) < \infty$. From a slight extension of Corollary 2 as applied to the embedded Markov chain for the averages of $f(\cdot)$ over the whole chain, we obtain

$$\pi_0(B_f) = 1$$

where

$$B_f = \left\{x : P_x\left\{\frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \int f(x)\tilde{\pi}(dx) \text{ as } n \to \infty\right\} = 1\right\}.$$

From the argument at the beginning of the proof of Lemma 8 we have $P^{n_0-1}(x, B_f) = 1$ for π_1-almost all x. The definition of B_f is such that if $P^{n_0-1}(x, B_f) = 1$ then $x \in B_f$. Hence $\pi_1(B_f) = 1$, and similarly $\pi_r(B_f) = 1$ for $r = 2, 3, \ldots$. This together with the fact that $\tilde{\pi} = \pi$ establishes (1.10). Conclusion (1.11) follows from (1.10) and the uniform integrability of $\frac{1}{n}\sum_{j=1}^{n} f(X_j)$ under P_x for $[\pi]$-almost all x. ◇

2.4 ·Rates of Convergence and Remarks

The proof of the convergence of $P^n(x, \cdot)$ to $\pi(\cdot)$ rested mainly Parts (b) and (c) of Lemma 3. We can translate the equivalence of (2.14) and (2.15) stated in Remark 1 following the proof of Lemma 3 to results on geometric convergence in the ergodic theorem for Markov chains. We will state such a result and give a brief proof.

Theorem 6 *Suppose that the conditions of Proposition 2 hold and there is a $t_0 > 0$ such that*

$$\sum_{n=1}^{\infty} \exp(nt_0) \int |P^n(x, A) - P^{n+1}(x, A)|\rho(dx) < \infty. \tag{2.42}$$

Then there is a set D_0 with $\pi(D_0) = 1$, such that for each $x \in D_0$, there is a β with $\exp(-t_0) < \beta < 1$ and a $K < \infty$ such that

$$\sup_{C\in\mathcal{B}} |P^n(x, C) - \pi(C)| \leq K\beta^n.$$

The constants β and K can depend on x.

We remark that if in Proposition 2 the set A can be taken to be the whole space $\mathcal{X}$, then the series (2.42) converges automatically. In this case, Theorem 6 asserts geometric convergence, which is qualitatively the same result as that given by Theorem 3 (it does not give the explicit constant given in Theorem 3, however).

To prove Theorem 6, we will need the following lemma.

Lemma 10 *For each positive integer n,*

$$\bar{P}^n(\Delta, C) = \int_{x,y\in\mathcal{X}} \rho(dx)P^{n-1}(x,dy)\bar{P}(y,C) \quad \textit{for } C \in \bar{\mathcal{B}} \tag{2.43}$$

and

$$\bar{P}^n(\Delta,\Delta) = \epsilon^* \int_{x\in\mathcal{X}} \rho(dx)P^{n-1}(x,A). \tag{2.44}$$

The lemma is proved by induction on n. For $n = 1$, (2.43) is the same as definition (2.21) of $\bar{P}(x,C)$. The induction step is carried out by direct calculation. Equation (2.44) follows from (2.43) and (2.21).

Proof of Theorem 6 Use the construction of the Markov chain on the enlarged space $\bar{\mathcal{X}}$ as in the proof of Proposition 2. Let $\bar{f}^n(\Delta,\Delta)$ be the probability that the Markov chain $\bar{X}_n$ starting at Δ reaches Δ for the first time at time n. Identify p_n in Lemma 3 and Remark 1 with $\bar{f}^n(\Delta,\Delta)$. This is what was done in the proof of Proposition 2 in an indirect fashion while appealing to Proposition 1. It is easy to see that the r_n appearing in Lemma 3 and Remark 1 is $\bar{P}^n(\Delta,\Delta)$. From (2.44), Condition (2.14) reduces to Condition (2.42). Theorem 6 now follows from Remark 1 on rates of convergence. ◊

Remark 3 In Section 1, we described how to form a transition function from the two conditional distributions $\pi_{X_1|X_2}$ and $\pi_{X_2|X_1}$ obtained from a bivariate distribution π. We mentioned that for a Markov chain with such a transition function to converge in distribution to π it is necessary that $\pi_{X_1|X_2}$ and $\pi_{X_2|X_1}$ determine π. Some researchers have pondered over the question of when do the conditional distributions determine the joint distribution. Besag (1974) noted that uniqueness is guaranteed if the distributions are discrete and the support of π is a permutation invariant set. Theorem 1 gives a sufficient condition for uniqueness in the general case.

One can give a simple nondegenerate example to show that in general, the two conditional distributions do not determine the joint distribution. Let X_1 have a density function $p(x)$ such that

$$\sum_{-\infty<m<\infty} p(m+r) = c_r < \infty \quad \text{for each } r \in [0,1).$$

The density function $p(x) = \frac{1}{2}\exp(-|x|)$, for instance, satisfies this condition. Let $\pi_{X_2|X_1}$ be the distribution that puts masses

$$\begin{array}{ll} 1/2 & \text{at } x_1+1 \\ 1/2 & \text{at } x_1-1. \end{array} \tag{2.45}$$

and

This determines the other conditional distribution $\pi_{X_1|X_2}$. This puts masses

$$\frac{p(x_2+1)}{p(x_2+1)+p(x_2-1)} \quad \text{at } x_2+1 \qquad \text{and} \qquad \frac{p(x_2-1)}{p(x_2+1)+p(x_2-1)} \quad \text{at } x_2-1. \tag{2.46}$$

It can be seen that the two conditional distributions (2.46) and (2.45) do not uniquely determine a joint distribution for (X_1, X_2). Fix $r \in [0,1)$ and consider the discrete distribution p_r on the points $m+r$, $m = \dots, -1, 0, 1, \dots$ defined by $p_r(m+r) = \frac{1}{c_r}p(m+r)$. Let $Y_1(r)$ be distributed according to p_r, and let the conditional distribution of $Y_2(r)$ given $Y_1(r)$ be the distribution defined in (2.45). It is easy to see that distribution of $Y_1(r)$ given $Y_2(r)$ is that given in (2.46), and the joint distribution of $(Y_1(r), Y_2(r))$ has the same conditional distributions as (X_1, X_2).

It is even possible to find joint distributions with continuous marginals for which the conditionals are given by (2.46) and (2.45). Let $f(r)$ be any probability density on $[0,1)$. Let R have density function $f(r)$ and put $(Z_1, Z_2) = (Y_1(R), Y_2(R))$. Clearly the conditional distributions of (Z_1, Z_2) are as in (2.46) and (2.45). The marginal distribution function of Z_1 is given by

$$P(Z_1 \leq x) = \int_{[0,1)} \Big(\sum_{m:m+r \leq x} p_r(m+r) \Big) f(r)dr = \int_{-\infty}^{x} \frac{p(y)f(y-[y])}{c_{y-[y]}} dy$$

A similar expression can be written down for the distribution function of Z_2. Notice that Z_1 and Z_2 have density functions.

3 Remarks on the Sampling Plan

In Section 1 we mentioned that there are a number of ways of using the Markov chain to estimate π or some aspect of π. One can generate G independent chains, each of length n, and retain the last observation from each chain, obtaining a sample $X_n^{[1]}, X_n^{[2]}, \dots, X_n^{[G]}$ of independent variables. At another extreme, one can generate a very long sample $X_0, X_1, X_2, \dots, X_{nG}$, and use $X_n, X_{2n}, \dots, X_{Gn}$, which form a nearly i.i.d. sequence from π. This is at approximately the same cost in CPU time. (Clearly intermediate solutions are possible). If the objective is to estimate an expectation $\int f(x)\pi(dx)$, then there is no reason to discard the intermediate values from a long chain, and one can use

$$\frac{1}{n(G-1)} \sum_{i=n+1}^{nG} f(X_i). \tag{3.1}$$

The almost sure convergence of (3.1) follows from Theorem 2 under the assumption $\int f(x)\pi(dx) < \infty$ (note that we do not need the aperiodicity condition (1.6)). Thus, from the point of view of estimating a particular expectation $\int f(x)\pi(dx)$ or probability it is clear that the optimal way of using the Markov chain is to use (3.1), and so it is natural to ask why one should bother to prove results such as (1.7). In the Bayesian framework, there is another aspect that must be considered, which is that generally, in the exploratory stage, one is interested in calculating posterior distributions and densities for a large number of prior distributions. It will usually not be feasible to run a separate Markov chain for each prior of interest (the time needed is on the order of several minutes for each prior). Instead, one will want to get a sequence of random variables $X_1, \ldots, X_r$ distributed according to the posterior distribution with respect to some fixed prior, and then use that *same* sequence to estimate the posterior with respect to many other priors. (We discuss how this may be done in the next paragraph.) The important point here is that if there are a large number of priors involved, then the manipulations of the sequence $X_1, \ldots, X_r$ to produce the posterior for each prior must be done very quickly. This restricts the size of r, and so one will generally want the sequence $X_1, \ldots, X_r$ to be independent. This precludes running a very long chain and taking sample averages as in (3.1). Instead, one will want to generate independent chains and retain the last random variable in each chain or take a long chain and retain only random variables at equally spaced intervals.

We now discuss in more detail how one might use one sequence $X_1, \ldots, X_r$ to calculate posteriors with respect to many priors. We depart from the notation of the paper and switch to the notation usually used in Bayesian analysis. Suppose that ν_h is a family of priors for the parameter θ. Here, h lies in some interval and we think of it as a hyperparameter for the prior. Suppose that we are in the dominated case, i.e. there is a likelihood function $l_X(\theta)$, where X now represents the data.

Let $\nu_{h,X}$ be the posterior distribution of θ when the prior is ν_h. We know that $\nu_{h,X}$ is dominated by ν_h and

$$\frac{d\nu_{h,X}}{d\nu_h}(\theta) = c_h(X) l_X(\theta),$$

where $c_h(X)$ is a normalizing constant.

Consider the case where we can generate observations $\theta_1, \theta_2, \ldots, \theta_r$ from $\nu_{0,X}$ and therefore estimate $\int f(\theta) d\nu_{0,X}(\theta)$ by $(1/r)\sum_{i=1}^r f(\theta_i)$. We will indicate now how we can obtain estimates of $\int f(\theta) d\nu_{h,X}(\theta)$ for $h \neq 0$.

Suppose that ν_h is dominated by ν_0. Then it is clear that $\nu_{h,X}$ is dominated by $\nu_{0,X}$ and

$$\frac{d\nu_{h,X}}{d\nu_{0,X}}(\theta) = \frac{c_h(X)}{c_0(X)} \frac{d\nu_h}{d\nu_0}(\theta)$$

since the likelihood $l_X(\theta)$ cancels. We may write

$$\int f(\theta) d\nu_{h,X}(\theta) = \int f(\theta) \frac{d\nu_{h,X}}{d\nu_{0,X}}(\theta) d\nu_{0,X}(\theta) = \frac{c_h(X)}{c_0(X)} \int f(\theta) \frac{d\nu_h}{d\nu_0}(\theta) d\nu_{0,X}(\theta).$$

Substituting $f(\theta) \equiv 1$ in the above we can obtain the constant $\frac{c_h(X)}{c_0(X)}$ and write

$$\int f(\theta) d\nu_{h,X}(\theta) = \frac{\int f(\theta) \frac{d\nu_h}{d\nu_0}(\theta) d\nu_{0,X}(\theta)}{\int \frac{d\nu_h}{d\nu_0}(\theta) d\nu_{0,X}(\theta)}.$$

Thus, we may estimate $\int f(\theta) d\nu_{h,X}(\theta)$ by

$$\sum_{i=1}^r f(\theta_i) w_{h,i} \quad \text{where} \quad w_{h,i} = \frac{\frac{d\nu_h}{d\nu_0}(\theta_i)}{\sum_{i=1}^r \frac{d\nu_h}{d\nu_0}(\theta_i)}.$$

This is the well-known "ratio estimate" in importance sampling theory. The key here is its calculation requires only knowledge of the ratio $\frac{d\nu_{h,X}}{d\nu_{0,X}}$ up to a multiplicative constant. See Hastings (1970).

Now in some Bayesian problems, for instance problems with missing or censored data, the likelihood function $l_X(\theta)$ is either extremely difficult or impossible to calculate. (An example of this arises in Doss (1991).) The fact that this likelihood cancels means that the estimation of the expectation under the prior ν_a requires only the recomputation of r weights, and this can be done very fast.

It will often be the case that we wish to consider not just one function, but rather a family of functions. As a simple example, if we wish to estimate the entire posterior distribution of θ, then in effect we wish to consider $f_t(\theta) = I(\theta \leq t)$ for a fine grid of values of t. On a Sparcstation 1, for $r = 50$ we have been able to do the computations fast enough to dynamically display the estimates of the posterior distributions $\int I(\theta \leq t) d\nu_{h,X}(\theta)$ as h varies, using the program Lisp-Stat described in Tierney (1991). For larger values of r it was necessary to precompute these estimates.

References

Asmussen, S. (1987). *Applied Probability and Queues.* Wiley, New York.

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

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Jour. Royal Statist. Soc. Ser. B* **36** 192–236.

Chung, K. L. (1967). *Markov Chains.* Second Edition, Springer Verlag, New York.

Doob, J. L. (1953). *Stochastic Processes.* Wiley, New York.

Doss, H. (1991). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. Technical Report No. M850, Department of Statistics, Florida State University.

Escobar, M. and West, M. (1991). Bayesian density estimation and inference using mixtures. Technical Report No. 533, Department of Statistics, Carnegie Mellon University.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications, Volume I.* Third edition, John Wiley, New York.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **6** 721–741.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Hoel, P., Port, S., and Stone, C. (1972). *Introduction to Stochastic Processes.* Houghton Mifflin, Boston.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes.* Academic Press, New York.

Kendall, D. G. (1960). Geometric ergodicity and the theory of queues. Proceedings of the First Stanford Symposium on Mathematical Methods in the Social Sciences. Stanford University Press, Stanford.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators.* Cambridge University Press, Cambridge.

Orey, S. (1971). *Limit Theorems for Markov Chains Transition Probabilities.* Van Nostrand, New York.

Revuz, D. (1975). *Markov Chains*, North-Holland, Amsterdam.

Schervish, M. and Carlin B. (1990). On the convergence of successive substitution sampling. Technical Report No. 492, Department of Statistics, Carnegie Mellon University.

Stone, C. (1965). On characteristic functions and renewal theory. *Trans. Amer. Math. Soc.* **120** 327–342.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82** 528–550.

Tierney, L. (1991). Markov chains for exploring posterior distributions. Technical Report No. 560, School of Statistics, University of Minnesota.

Tierney, L. (1991). *Lisp-Stat.* Wiley, New York.

Zaman, A. (1992). Generating random numbers from a unimodal density by cutting corners. Technical Report, Department of Statistics, Florida State University.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
July 1992

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
A Proof of Convergence of the Markov Chain Simulation Method

5. FUNDING NUMBERS

6. AUTHOR(S)
Krishna B. Athreya
Hani Doss
Jayaram Sethuraman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Florida State University
Department of Statistics
Tallahassee, FL 32305-3033

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Abstract

The Markov chain simulation method has been successfully used in many problems, including some that arise in Bayesian statistics. We give a self-contained proof of the convergence of this method in general state spaces under conditions that are easy to verify. We also provide a result giving a geometric rate of convergence.

14. SUBJECT TERMS
Successive substitution sampling, calculation of posterior distributions, ergodic theorem.

15. NUMBER OF PAGES
30

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL