The Annals of Statistics
1996, Vol. 24, No. 1, 69–100

ON THE CONVERGENCE OF THE MARKOV CHAIN SIMULATION METHOD

BY KRISHNA B. ATHREYA,[1] HANI DOSS[2] AND JAYARAM SETHURAMAN[3]

Iowa State University, Ohio State University and Florida State University

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.

1. Introduction. Let π be a probability distribution on a measurable space $(\mathscr{X}, \mathscr{B})$, and suppose that we are interested in estimating characteristics of it, such as $\pi(E)$ or $\int f d\pi$, where $E \in \mathscr{B}$ and f is a bounded measurable function. Even when π is fully specified one may have to resort to methods like Monte Carlo simulation, 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 $\mathscr{X}$ 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, discussed further below.

Let P be a transition probability function on a measurable space $(\mathscr{X}, \mathscr{B})$, that is, P is a function on $\mathscr{X} \times \mathscr{B}$ such that, for each $x \in \mathscr{X}$, $P(x, \cdot)$ is a probability measure on $(\mathscr{X}, \mathscr{B})$ and, for each $C \in \mathscr{B}$, $P(\cdot, C)$ is a measurable function on $(\mathscr{X}, \mathscr{B})$. Suppose that π is a probability measure on $(\mathscr{X}, \mathscr{B})$ which is invariant for the Markov chain, that is,

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

We fix a starting point x_0, generate an observation X_1 from $P(x_0, \cdot)$, generate an observation X_2 from $P(X_1, \cdot)$ and so on. This generates the Markov chain $x_0 = X_0, X_1, X_2, \ldots$. In order to make use of the Markov chain

Received June 1992; revised March 1995.
[1]Research supported by NSF Grant DMS-90-07182.
[2]Research supported by Air Force Office of Scientific Research Grant F49620-94-1-0028.
[3]Research supported by Army Research Office Grant DAAL03-90-G-0103.
AMS 1991 *subject classifications*. Primary 60J05; secondary 65U05, 60B10.
Key words and phrases. Calculation of posterior distributions, ergodic theorem, successive substitution sampling.

$\{X_n\}_{n=0}^{\infty}$ to get some information about π, one needs results of the following form.

(A) (Ergodicity) For all or for "most" starting values x, the distribution of X_n converges to π in a suitable sense, such as the following, for example:

(A1) *variation norm ergodicity*: $\sup_{C \in \mathscr{B}} |P^n(x, C) - \pi(C)| \to 0$; or

(A2) *variation norm mean ergodicity*:

$$\sup_{C \in \mathscr{B}} \left| \frac{1}{n} \sum_{j=0}^{n} P^j(x, C) - \pi(C) \right| \to 0.$$

(B) (Law of large numbers) For all or for most starting values x, for each $C \in \mathscr{B}$,

$$\frac{1}{n} \sum_{j=0}^{n} I_C(X_j) \to \pi(C) \quad \text{for a.e. realization of the chain,}$$

and, for each f with $\int |f| \, d\pi < \infty$,

$$\frac{1}{n} \sum_{j=0}^{n} f(X_j) \to \int f d\pi \quad \text{for a.e. realization of the chain.}$$

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.

Thus our goal is to find conditions on a given Markov chain or rather on its transition function $P(\cdot, \cdot)$ so that some or all of the conditions (A) and (B) above hold, assuming that P admits an invariant probability measure π. In applications of Markov chain simulation, the probability measure π of interest is *by construction* the invariant probability measure of the Markov chain.

When $\{X_n\}$ is a Markov chain with a countable state space, say, $\{1, 2, \ldots\}$, and transition probability matrix $P = (p_{i,j})$, the existence of an invariant probability distribution π and the *irreducibility condition* that there exists a state i_0 such that, from any initial state i, there is positive probability that the chain eventually hits i_0 are enough to guarantee that (i) the chain $\{X_n\}$ is recurrent in an appropriate sense, (ii) conditions (B) and (A2) hold and (iii) when an additional aperiodicity condition also holds, then (A1) also holds. These facts are well known [see, e.g., Hoel, Port and Stone (1972)].

A natural question is whether this is true for general state space Markov chains. In particular, when (1.1) holds, is there a form of the irreducibility condition under which some or all of (A) and (B) hold?

The Markov chain literature has a number of results in this direction; see Orey (1971), Athreya and Ney (1978) and Nummelin (1984). Under a condition known as Harris recurrence (see below), the existence of an invariant distribution π implies mean ergodicity [condition (A2)] and the laws of large

numbers [condition (B)]. Unfortunately, Harris recurrence is not an easy condition to verify in general, and it is much stronger than irreducibility.

The main point of this paper is to show that when (1.1) holds, a simple irreducibility condition [(1.4) below] is enough to yield (A2) and (B). An additional aperiodicity condition yields (A1) as well. This provides a complete generalization of the results for the countable case. *It is worth noting that recurrence emerges as a consequence of* (1.1) *and the irreducibility condition* (1.4), *and it is not imposed as a hypothesis*.

Before stating our main theorems, we will need a few definitions. For any set $C \in \mathscr{B}$, 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 $A \in \mathscr{B}$ is said to be *accessible from* x if $P_x(T(A) < \infty) > 0$. Let ρ be a probability measure on $(\mathscr{X}, \mathscr{B})$. The Markov chain is said to be *ρ-recurrent* (or Harris recurrent with respect to ρ) if, for every A with $\rho(A) > 0$, $P_x(T(A) < \infty) = 1$ for all $x \in \mathscr{X}$. The chain is said to be *ρ-irreducible* if every set A with $\rho(A) > 0$ is accessible from all $x \in \mathscr{X}$. The set A is said to be *recurrent* if $P_x(T(A) < \infty) = 1$ for all $x \in \mathscr{X}$.

For the case where the σ-field $\mathscr{B}$ 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 $A \in \mathscr{B}$ with $\rho(A) > 0$, an integer n_0 and a number $\varepsilon > 0$ satisfying

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

and

$$x \in A, C \in \mathscr{B} \quad \text{imply} \quad P^{n_0}(x, C) \geq \varepsilon \rho(C \cap A). \tag{1.3}$$

Let $\rho_A(C) = \rho(C \cap A)/\rho(A)$. 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 accessible from all $x \in \mathscr{X}$ 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, usually will be 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).

We denote the greatest common divisor of any subset $\mathscr{M}$ of integers by g.c.d.($\mathscr{M}$).

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, *that is*, (1.1) *holds. Suppose that there is a set* $A \in \mathscr{B}$, *a probability measure* ρ *with* $\rho(A) = 1$, *a constant* $\varepsilon > 0$ *and an integer* $n_0 \geq 1$ *such that*

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

and

$$P^{n_0}(x, \cdot) \geq \varepsilon\rho(\cdot) \quad \text{for each } x \in A. \tag{1.5}$$

Suppose further that

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

Then there is a set D *such that*

$$\pi(D) = 1 \quad \text{and} \quad \sup_{C \in \mathscr{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \text{for each } x \in D. \tag{1.7}$$

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*

$$\sup_{C \in \mathscr{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r}(x, C) - \pi(C) \right| \to 0 \tag{1.8}$$

as $m \to \infty$ *for* $[\pi]$*-almost all* x,

and hence

$$\sup_{C \in \mathscr{B}} \left| \frac{1}{n} \sum_{j=1}^{n} P^j(x, C) - \pi(C) \right| \to 0 \tag{1.9}$$

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

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

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

and

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

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 an invariant probability measure before establishing (1.7) and (1.8). Theorems 1 and 2 exploit the existence of an invariant probability measure (which is given to us "for free" in the Markov chain simulation context) 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 $n_0 = 1$. Condition (1.5) holds if, for each $x \in A$, $P^{n_0}(x, \cdot)$ has a nontrivial absolutely continuous component with respect to some measure ρ and the associated density $p^{n_0}(x, y)$ satisfies $\inf_{x, y \in A} p^{n_0}(x, y) > 0$ for some A with $\rho(A) > 0$. In a remark following the proof of Lemma 3 we indicate the critical points where one can use additional information to obtain results on the rate of the convergence.

In many interesting problems, including those that arise in Bayesian statistics, described later, the state space $\mathscr{X}$ 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′) of Doob [(1953), page 197], which can be stated in an equivalent way as follows. There is a probability measure ϕ on $(\mathscr{X}, \mathscr{B})$, an integer k and an $\varepsilon > 0$ such that

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

This is a very strong condition. It implies that there exists an invariant 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,*

$$\sup_{C \in \mathscr{B}} |P^n(x, C) - \pi(C)| \leq (1 - \varepsilon)^{(n/k)-1} \quad \textit{for all } x \in \mathscr{X}.$$

A proof of this theorem may be found in Doob [(1953), page 197]. The Doeblin condition, although 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 $\mathscr{X}$. In its absence, one has to impose the obvious conditions of irreducibility and aperiodicity and some other extra conditions, such as 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 $A \in \mathscr{B}$, a probability measure ρ concentrated on A and an ε with $0 < \varepsilon < 1$ such that*

$$P_x(T(A) < \infty) = 1 \quad \textit{for all } x \in \mathscr{X}$$

and

$$P(x, C) \geq \varepsilon\rho(C) \quad \textit{for all } x \in A \textit{ and all } C \in \mathscr{B}.$$

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

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

This theorem establishes ergodicity under the assumption of the existence of an invariant probability measure but also makes the strong assumption of the existence of a recurrent set A. It is often 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. [We remark that our Theorems 1 and 2 state only that convergence occurs for $[\pi]$-almost all starting points. Examples can be given to show that this is the strongest assertion that can be made, even if instead of (1.4) the set A is assumed to be accessible from *all* x.]

Based on the work of Nummelin (1984), Tierney (1994) gives sufficient conditions for convergence of Markov chains to their invariant 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 invariant measure. This distinction is more than cosmetic. To check π-irreducibility, one has to show that $P_x(T_A < \infty) > 0$ for all $x \in \mathscr{X}$ and all A for which $\pi(A) > 0$. For certain Markov chain simulation problems in which the state space is very complicated, it is difficult or impossible even to 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 (1994), where the state space was the set of all distribution functions. In that paper, the Markov chain simulation method was proposed as a way to determine π, but the unknown π was sufficiently complicated that one could not determine the sets to which it gives positive measure. 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). See the discussion in Doss [(1994), Section 4].

We point out that Tierney (1994) does not give a detailed definition of aperiodicity, but refers the reader to Nummelin [(1984), Chapter 2.4], 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 usually easy to check: if the n_0 appearing in (1.5) is 1, then (1.6) is automatic.

The statistical applications of the above include the Metropolis algorithm and its variants which 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 (1994) for a review.

However, in the usual problems of Bayesian statistics, currently 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 for 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 (which 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 at the end of Section 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 from every starting point 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 $\mathscr{L}(Y|\theta)$ and that we wish to obtain $\mathscr{L}(\theta|Y)$, 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 $\pi_{\theta, Z} = \mathscr{L}(\theta, Z|Y)$ has the property that $\pi_{\theta|Z}$ $[= \mathscr{L}(\theta|Y, Z)]$ and $\pi_{Z|\theta}$ $[= \mathscr{L}(Z|Y, \theta)]$ 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) $\mathscr{L}(\theta, Z|Y)$, and retaining the first coordinate gives an observation with distribution (approximately) equal to $\mathscr{L}(\theta|Y)$.

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 Gilks and Wild (1992)], and the use of these algorithms is made routine by the computer language BUGS [Thomas, Spiegelhalter and Gilks (1992)].

Results which give not only convergence of the Markov chain to its invariant distribution but also convergence at a geometric rate are obviously extremely desirable. General results establishing that the convergence rate is geometric are given in Schervish and Carlin [(1992), Theorem 1] and in Chan [(1993), Theorem 2.1]. For certain models it is possible to give actual bounds for the geometric rate of convergence; see Goodman and Sokal (1989) and Amit (1991) for examples involving continuous state spaces. It is, however, important to keep in mind that for most problems arising in Bayesian statistics, checking conditions that ensure convergence at a geometric rate is 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 (1994) 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,

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

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. 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. Under these simplified assumptions we establish ergodicity and some laws of large numbers for 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 relaxed 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 an invariant probability distribution π_0 which is the restriction of π to the set A_0 defined by (2.29). 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.* Let $(\mathscr{X}, \mathscr{B})$ be a measurable space and let $\{X_n\}_0^\infty$ be a Markov chain with a probability transition function $P(\cdot, \cdot)$. Fix a point α in $\mathscr{X}$. 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

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

and

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

be the set of all states 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 $E_\alpha(T) < \infty$.

All the results of this section are part of the applied probability literature [see, e.g., Asmussen (1987)]. We present the results below, with proofs, to make the paper self-contained.

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

(a) *π is an invariant probability measure for P.*
(b) *$\pi\{x: P_x(T < \infty) > 0\} = 1$.*
Then

$$\pi(C_0) = 1 \quad \textit{and} \quad \sup_{C \in \mathscr{B}} \left| \frac{1}{n} \sum_{j=0}^{n} P^j(x, C) - \pi(C) \right| \to 0 \tag{2.3}$$

for each $x \in C_0$.

Suppose in addition that:
(c) g.c.d.$\{n: P^n(\alpha, \alpha) > 0\} = 1$.
Then

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

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 $x \in \mathscr{X}$. 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 $H(n) = P_\alpha(T \le n)$. These facts, the strong law of large numbers and the inequality

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

imply that

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

$[P_x]$-a.e., for each $x \in \mathscr{X}$. From the bounded convergence theorem, it follows that

$$\frac{1}{n}G_n(x, \alpha) = E_x\left(\frac{1}{n}N_n\right) \to \frac{1}{E_\alpha(T)}P_x(N = \infty) \quad \text{for each } x \in \mathscr{X}. \tag{2.9}$$

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

$$\pi(\alpha) = \frac{1}{E_\alpha(T)}\int \pi(dx)P_x(N = \infty). \tag{2.10}$$

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), did not use the full force of condition (b). The following corollary records that fact and will be used later in the proof of Lemma 6.

COROLLARY 1. *Let π satisfy condition* (a) *of Proposition* 1, *and let $E \in \mathscr{B}$ be such that*

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

Then $\pi(E) > 0$.

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

LEMMA 2. *Let α be recurrent. Let*

$$\nu(C) = E_\alpha\left(\sum_{j=0}^{T-1} I(X_j \in C)\right) = \sum_{n=0}^{\infty} P_\alpha(X_n \in C, T > n) \tag{2.11}$$

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

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

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

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

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 an invariant probability measure for $P(\cdot,\cdot)$ with $\pi(\mathscr{X}_0^c) = 0$. Then*

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

and

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

is the unique invariant 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,

$$\begin{aligned} 0 = P_\alpha(T = \infty) &= E_\alpha\big(P_\alpha(T = \infty | X_1, X_2, \ldots, X_n)\big) \\ &= E_\alpha\big(P_{X_n}(T = \infty) I(T > n)\big). \end{aligned}$$

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 (2.11) it now follows that $\nu(C_0^c) = 0$. We now show that ν is an invariant measure. Let $f(x)$ be a bounded measurable function on $(\mathscr{X}, \mathscr{B})$. Then

$$\begin{aligned} \int \nu(dx) f(x) &= \sum_{n=0}^{\infty} E_\alpha(f(X_n) I(T > n)) \\ &= f(\alpha) + \sum_{n=1}^{\infty} \big(E_\alpha(f(X_n) I(T > n-1)) - E_\alpha(f(X_n) I(T = n))\big) \\ &= f(\alpha) + \sum_{n=1}^{\infty} E_\alpha\big(E_\alpha(f(X_n) I(T > n-1)) | X_0, X_1, \ldots, X_{n-1}\big) \\ &\quad - \sum_{n=1}^{\infty} E_\alpha(f(X_n) I(T = n)) \end{aligned}$$

$$= f(\alpha) - f(\alpha) P_\alpha(T < \infty) + \sum_{n=1}^{\infty} E_\alpha\big(E_{X_{n-1}}(f(X_n)) I(T > n-1)\big)$$

$$= \sum_{n=1}^{\infty} E_\alpha\left(\int P(X_{n-1}, dy) f(y) I(T > n-1)\right)$$

$$= \sum_{n=0}^{\infty} E_\alpha\left(\int P(X_n, dy) f(y) I(T > n)\right)$$

$$= \int_{y \in \mathscr{X}} \left(\int_{x \in \mathscr{X}} \nu(dx) P(x, dy)\right) f(y),$$

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

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

$$\nu'(C) = \int \nu'(dx) P(x, C)$$

$$= \nu'(\alpha) P_\alpha(X_1 \in C) + \int_{x \neq \alpha} \nu'(dx) P_x(X_1 \in C)$$

$$= \nu'(\alpha) P_\alpha(X_1 \in C) + \int_{y \in \mathscr{X}} \int_{x \neq \alpha} \nu'(dy) P(y, dx) P_x(X_1 \in C)$$

$$= \nu'(\alpha) P_\alpha(X_1 \in C) + \int \nu'(dy) P_y(X_2 \in C, T > 1)$$

$$\vdots$$

$$= \nu'(\alpha) \sum_{m=1}^{n} P_\alpha(X_m \in C, T > m-1) + \int \nu'(dy) P_y(X_{n+1} \in C, T > n)$$

$$\geq \nu'(\alpha) \sum_{m=1}^{n} P_\alpha(X_m \in C, T > m-1)$$

$$\geq \nu'(\alpha) \sum_{m=1}^{n} P_\alpha(X_m \in C, T > m),$$

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 an invariant 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. Since $\mathscr{X}_0 = \{x\colon G(x, \alpha) > 0\}$, this implies that $\lambda(\mathscr{X}_0) = 0$. This proves that

$$\nu'(C) = \nu'(\alpha)\nu(C), \tag{2.12}$$

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

We now assume that α is positive recurrent. Since $\sum_{n=0}^{T-1} I(X_n \in \mathscr{X}) = T$, we have $\nu(\mathscr{X}) = E_\alpha(T) < \infty$. Let π be an invariant probability measure satisfying $\pi(\mathscr{X}_0^c) = 0$. From (2.12), 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 invariant probability measure. □

In the following, we 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.

COROLLARY 2. *Let conditions* (a) *and* (b) *of Proposition* 1 *hold. Let* $f(x)$ *be a measurable function on* $(\mathscr{X}, \mathscr{B})$ *with* $\int |f(y)|\pi(dy) < \infty$. *Then* $\pi(A_f) = 1$, *where*

$$A_f = \left\{ x : P_x\left\{ \frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \int f(x)\, d\pi(x) \right\} = 1; \right.$$
$$\left. E_x\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j) \right) \to \int f(x)\, d\pi(x) \right\}.$$

PROOF. By considering positive and negative parts, we can assume that $f(\cdot) \geq 0$. Let $\{T_k\}_{k=1}^{\infty}$ and N_n be as in Lemma 1. Define

$$U_n = \sum_{j=1}^{\min(n, T_1 - 1)} f(X_j) \quad \text{and} \quad V_r = \sum_{j=T_r}^{T_{r+1}-1} f(X_j).$$

Note that $V_1, V_2, \ldots$ are i.i.d. and, from Lemma 2, $E(V_1) = \int f(x)\pi(dx) E_\alpha(T_1)$. Since $f(x) \geq 0$, we have the inequality

$$\sum_{r=1}^{N_{n-1}} V_r \leq \sum_{j=1}^{n} f(X_j) \leq U_n + \sum_{r=1}^{N_n} V_r.$$

For $x \in C_0$ [defined in (2.1)], $P_x(T_k < \infty \text{ for all } k \geq 1) = 1$ and hence, from (2.9) and the law of large numbers,

$$\frac{N_n}{n} \to \frac{1}{E_\alpha(T_1)} \quad \text{and} \quad \frac{1}{n}\sum_{r=1}^{N_n} V_r \to \frac{E_\alpha(V_1)}{E_\alpha(T_1)} = \int f(x)\pi(dx),$$

$[P_x]$-a.e. From Fatou's lemma, this yields

$$\liminf E_x\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j) \right) \geq \int f(x)\pi(dx),$$

$[\pi]$-a.e. To obtain the reverse inequality

$$\limsup E_x\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right) \le \int f(x)\pi(dx), \tag{2.13}$$

we begin by using Wald's identity $E(\sum_{r=1}^{N_n} V_r) = E_\alpha(V_1)E_\alpha(N_n)$, to obtain

$$E_x\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j)\right) \le E_x\left(\frac{1}{n}\sum_{j=1}^{T_1-1} f(X_j)\right) + E_\alpha(V_1)E_\alpha\left(\frac{N_n}{n}\right).$$

We have already seen that the second term on the right-hand side of the above converges to $\int f(x)\,d\pi(x)$. We now prove that $E_x(\sum_{j=1}^{T_1-1} f(X_j)) < \infty$ with $[\pi]$-probability 1, and this will establish the inequality (2.13). Note that, for $r = 0, 1, \ldots,$

$$\begin{aligned}\infty &> E_\alpha\left(\sum_{j=0}^{T_1-1} f(X_j)\right)\\ &\ge E_\alpha\left(E_x\left(\sum_{j=0}^{T_1-1} f(X_j) I(T_1 > r)\right)\right) = E_\alpha\left(E_{X_r}\left(\sum_{j=0}^{T_1-1} f(X_j)\right) I(T_1 > r)\right),\end{aligned}$$

since $I(T_1 > r)$ is measurable with respect to $\{X_1, \ldots, X_r\}$. Thus, $E_x(\sum_{j=0}^{T_1-1} f(X_j)) < \infty$ for almost all x with respect to $P_\alpha(X_r \in \cdot;\ T_1 > r)$. From Lemma 2 we conclude that $\sum_{r=0}^{\infty} P_\alpha(X_r \in C;\ T > r) = E_\alpha(T_1)\pi(C)$. Thus $E_x(\sum_{r=0}^{T_1-1} f(X_j)) < \infty$ with $[\pi]$-probability 1. This completes the proof of Corollary 2. □

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

$$r_n = \sum_{k=0}^{\infty} p_n^{(k)}. \tag{2.14}$$

Then, the following holds:

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

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

Furthermore, we have the following:

(b) $$\frac{1}{n}\sum_{j=0}^{n} r_j \to \frac{1}{\mu} \quad as\ n \to \infty.$$

If the additional condition g.c.d$\{n\colon\ p_n > 0\} = 1$ *holds, then we have the following*:

(c) $$r_n \to \frac{1}{\mu} \quad as\ n \to \infty.$$

PROOF. It is easy to establish (a) by direct verification. To prove part (b), we note that $\sum_{j=0}^{n} r_j = \sum_{k=0}^{\infty} P(S_k \le n) = E(N(n))$, where $N(n) = \sup\{k\colon S_k \le n\}$. By the strong law of large numbers and the inequalities

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

it follows that

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

Since $N_n/n \le 1$, it follows that $E(N_n/n) = (1/n)\sum_{j=0}^{n} r_j \to 1/\mu$, which establishes (b).

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., Feller (1950), Chapter XIII.10] and others are probabilistic [see, e.g., Hoel, Port and Stone (1972), Chapter 2]. □

REMARK. The proofs given in this paper require only the convergence of r_n to $1/\mu$ asserted in part (c) of Lemma 3. In fact, geometric bounds on the tail behavior of the probability distribution $\{p_n\}$ can be used to obtain geometric bounds on the rate of convergence of r_n to $1/\mu$ [Stone (1965)]. These can in turn can be used to obtain results on geometric convergence in the ergodic theorem for Markov chains [see Athreya, Doss and Sethuraman (1992), Section 2.4].

PROOF OF PROPOSITION 1. Let D be the collection of all measurable functions f on $(\mathscr{X}, \mathscr{B})$ with $\sup_y |f(y)| \le 1$. Let $f \in D$. Then, for any $x \in \mathscr{X}$,

$$\begin{aligned} E_x(f(X_n)) &= E_x(f(X_n)I(T > n)) \\ &\quad + \sum_{k=0}^{n} P_x(T = k)E_\alpha(f(X_{n-k})), \qquad n = 0, 1, \ldots . \end{aligned} \tag{2.15}$$

Let $v_n = E_\alpha(f(X_n))$, $a_n = E_\alpha(f(X_n)I(T > n))$ and $p_n = P_\alpha(T = n)$, $n = 0, 1, \ldots$. Note that v_n and a_n depend on the function f, while p_n does not. Setting $x = \alpha$ in (2.15), we get the important identity

$$v_n = a_n + \sum_{k=0}^{n} p_k v_{n-k}. \tag{2.16}$$

It is not difficult to check that $v_n = \sum_{k=0}^{n} a_k r_{n-k}$ is the unique solution to (2.16), where r_n is as defined in (2.14). 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 $f \in D$,

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

for any positive integer m. Note that, for fixed m,

$$\sup_{n-m \le k \le n} \left|\frac{R_k}{n} - \frac{1}{\mu}\right| \to 0 \quad \text{as } n \to \infty$$

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

$$\left|\frac{1}{n}\sum_{j=0}^{n} v_j - \int f\,d\pi\right| \to 0 \quad \text{uniformly in } f \text{ as } n \to \infty. \tag{2.17}$$

Let $x \in C_0$. Let $w_n = E_x(f(X_n))$, $b_n = E_x(f(X_n)I(T > n))$ 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 (2.15) once again, we see that w_n satisfies the equation

$$w_n = b_n + \sum_{k=0}^{n} g_k v_{n-k}, \qquad n = 0, 1, \ldots. \tag{2.18}$$

Using (2.17), 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.$\{n\colon P^n(\alpha, \alpha) > 0\} = 1$, and thus g.c.d.$\{n\colon p_n > 0\} = 1$ [see, e.g., the lemma in Chung (1967), page

29]. Thus, from part (c) of Lemma 3 we have $r_n \to 1/\mu$. Repeating the arguments leading to (2.17) and (2.18) with this stronger result on r_n, we see that $v_n \to \int f d\pi$ and $w_n \to \int f d\pi$ uniformly in $f \in D$. This proves (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 and although it is technically weaker, its proof contains the heart of the arguments needed to establish Theorem 1.

PROPOSITION 2. *Suppose that* $A \in \mathscr{B}$ *and let* ρ *be a probability measure on* $(\mathscr{X}, \mathscr{B})$ *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*

$$\pi(D_0) = 1 \quad \text{and} \quad \sup_{C \in \mathscr{B}} |P^n(x, C) - \pi(C)| \to 0 \quad \text{for each } x \in D_0. \tag{2.19}$$

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

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

$$\bar{P}(x, C) = \begin{cases} P(x, C), & \text{if } x \in \mathscr{X} \setminus A,\ C \in \mathscr{B}, \\ P(x, C) - \varepsilon^* \rho(C), & \text{if } x \in A,\ C \in \mathscr{B}, \\ \varepsilon^*, & \text{if } x \in A,\ C = \{\Delta\}, \\ \int_A \rho(dz) \bar{P}(z, C), & \text{if } x = \Delta,\ C \in \bar{\mathscr{B}}. \end{cases} \tag{2.20}$$

Also, define the set function $\bar{\pi}$ on $(\bar{\mathscr{X}}, \bar{\mathscr{B}})$ by

$$\bar{\pi}(C) = \begin{cases} \pi(C) - \varepsilon^* \rho(C) \pi(A), & \text{if } C \in \mathscr{B}, \\ \varepsilon^* \pi(A), & \text{if } C = \{\Delta\}. \end{cases} \tag{2.21}$$

It is easy to verify that $\bar{P}(x, C)$ and $\bar{\pi}(C)$ extend to $\bar{\mathscr{B}}$ as a transition probability function and probability measure, respectively.

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.

Clearly $\bar{P}(\Delta, \Delta) = \varepsilon^* > 0$, so that condition (c) of Proposition 1 is satisfied. Recall that $\mathscr{X}_0 = \{x \in \mathscr{X}\colon\ G(x, A) > 0\}$. Let $\bar{\mathscr{X}}_0 = \{x \in \bar{\mathscr{X}}\colon\ \bar{G}(x, \Delta) > 0\}$. If $x \in \mathscr{X}_0 \setminus A$, then $\bar{G}(x, \Delta) \geq \int_A G(x, dy) \bar{P}(y, \Delta) \geq \varepsilon^* G(x, A) > 0$. This shows that $\mathscr{X}_0 \subset \bar{\mathscr{X}}_0$. Since $\bar{G}(\Delta, \Delta) \geq \bar{P}(\Delta, \Delta) = \varepsilon^* > 0$, we also have $\Delta \in \bar{\mathscr{X}}_0$, that is, $\bar{\mathscr{X}}_0 \supset \mathscr{X}_0 \cup \{\Delta\}$. Since $\pi(\mathscr{X}_0) = 1$ by (1.4), it follows that $\bar{\pi}(\bar{\mathscr{X}}_0) = 1$ by (2.21). Thus, condition (b) of Proposition 1 is satisfied.

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

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

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

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

This verifies condition (a) of Proposition 1.

Let

$$\bar{D}_0 = \left\{x\colon x \in \bar{\mathscr{X}},\ \bar{P}_x\left(\bar{T}_\Delta < \infty\right) = 1\right\}.$$

From Proposition 1 it follows that

$$\bar{\pi}\left(\bar{D}_0\right) = 1 \quad \text{and} \quad \sup_{C \in \bar{\mathscr{B}}} \left|\bar{P}^n(x,C) - \bar{\pi}(C)\right| \to 0 \quad \text{for each } x \in \bar{D}_0. \tag{2.22}$$

To translate (2.22) into a result for $P^n(x,C)$, we define a function $v(x,C)$ on $\bar{\mathscr{X}} \times \mathscr{B}$ by

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

We may view $v(x,C)$ as a transition function from $\bar{\mathscr{X}}$ into $\mathscr{X}$. 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*:

$$P(x,C) = \int_{\bar{\mathscr{X}}} \bar{P}(x,dy)v(y,C) = \bar{P}(x,C) + \bar{P}(x,\Delta)\rho(C) \tag{2.23}$$

for $x \in \mathscr{X}$, $C \in \mathscr{B}$;

$$\bar{P}(x,C) = \int_{\mathscr{X}} v(x,dy)\bar{P}(y,C) \qquad \text{for } x \in \bar{\mathscr{X}},\ C \in \bar{\mathscr{B}}; \tag{2.24}$$

$$P^n(x,C) = \int_{\bar{\mathscr{X}}} \bar{P}^n(x,dy)\nu(y,C) = \bar{P}^n(x,C) + \bar{P}^n(x,\Delta)\rho(C) \quad \text{for } x \in \mathscr{X}, C \in \mathscr{B}; \tag{2.25}$$

and

$$\pi(C) = \int_{\bar{\mathscr{X}}} \bar{\pi}(dx)\nu(x,C) = \bar{\pi}(C) + \bar{\pi}(\Delta)\rho(C) \quad \text{for } C \in \mathscr{B}. \tag{2.26}$$

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

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

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

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

We prove (2.25) by induction on n. For $n = 1$, this is just (2.23). Assume that (2.25) has been proved for $n - 1$.

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

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

where the second inequality follows from (2.24), the third follows from (2.23) and the fourth from the induction step.

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

$$\begin{aligned}&\int_{\bar{\mathscr{X}}} \bar{\pi}(dx)\nu(x,C) \\ &\quad = \int_{\mathscr{X}} (\pi(dx) - \varepsilon^*\pi(A)\rho(dx))\nu(x,C) + \varepsilon^*\pi(A)\rho(C) = \pi(C).\end{aligned}$$

This completes the proof of the lemma. □

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

LEMMA 5. *Let $C \in \mathscr{B}$. Then*

$$\overline{\pi}(C) = 0 \quad \textit{implies} \quad \rho(C) = 0. \tag{2.27}$$

PROOF. From the careful choice of $\varepsilon^* = \varepsilon/2$ used to define $\overline{P}(x, C)$ in definition (2.20), we have

$$\overline{P}(x, C) = P(x, C) - \varepsilon^* \rho(C) > \varepsilon^* \rho(C) \quad \text{whenever } x \in A \text{ and } C \in \mathscr{B}. \tag{2.28}$$

Since $\overline{\pi}$ is an invariant probability measure for $\overline{P}(\cdot, \cdot)$,

$$\begin{aligned} \overline{\pi}(C) &= \int_{\overline{\mathscr{X}}} \overline{P}(x, C)\overline{\pi}(dx) \geq \int_A \overline{P}(x, C)\overline{\pi}(dx) \\ &\geq \varepsilon^* \rho(C)\overline{\pi}(A) = \varepsilon^* \rho(C)\pi(A)(1 - \varepsilon^*) \\ &= \overline{\pi}(\Delta)\rho(C)(1 - \varepsilon^*), \end{aligned}$$

by using the identity $\overline{\pi}(\Delta) = \varepsilon^* \pi(A)$. From Lemma 1 applied to the Markov chain with transition probability function $\overline{P}(\cdot, \cdot)$ on $(\overline{\mathscr{X}}, \overline{\mathscr{B}})$ and distinguished point Δ, we find that $\overline{\pi}(\Delta) > 0$. This establishes (2.27). □

COMPLETION OF THE PROOF OF PROPOSITION 2. Let $D_0 = \overline{D}_0 - \Delta$. From (2.22), (2.25) and (2.26), we have $\overline{\pi}(\overline{D}_0) = 1$, and

$$\sup_{C \in \mathscr{B}} |P^n(x, C) - \pi(C)| = \sup_{C \in \mathscr{B}} \left| \int_{\overline{\mathscr{X}}} \overline{P}^n(x, dy) v(y, C) - \int_{\overline{\mathscr{X}}} \overline{\pi}(dy) v(y, C) \right| \to 0$$

for each $x \in D_0$.

This means that

$$\overline{\pi}(\mathscr{X} - D_0) = \overline{\pi}\left(\overline{\mathscr{X}} - \overline{D}_0\right) = 0.$$

From Lemma 5, it follows that

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

Now, from the definition of $\overline{\pi}(\cdot)$ in (2.21),

$$\pi(\mathscr{X} - D_0) = \overline{\pi}(\mathscr{X} - D_0) + \varepsilon^* \rho(\mathscr{X} - D_0)\pi(A) = 0.$$

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

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

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

PROOF OF THEOREM 1. Let

$$\mathscr{M} = \left\{ m: \text{there is an } \varepsilon_m > 0 \text{ such that } \inf_{x \in A} P^m(x, \cdot) \geq \varepsilon_m \rho(\cdot) \right\}.$$

Then g.c.d.$(\mathscr{M}) = 1$. Let $m_1, m_2 \in \mathscr{M}$. Then, for $x \in A$ and $C \in \mathscr{B}$,

$$P^{m_1 + m_2}(x, C) \geq \int_A P^{m_2}(y, C) P^{m_1}(x, dy) \geq \varepsilon_{m_2} \rho(C) \varepsilon_{m_1} \rho(A) = \varepsilon_{m_1} \varepsilon_{m_2} \rho(C).$$

Thus $m_1 + m_2 \in \mathscr{M}$. Since g.c.d.$(\mathscr{M}) = 1$, there is an integer L such that $m \geq L$ implies that $m \in \mathscr{M}$. Now, from (1.4), for $[\pi]$-a.e. x, there is an s (which may depend on x) such that $P^s(x, A) > 0$. Fix an $m \in \mathscr{M}$. For any integer k, such that $km - s \geq L$, we have $P^{km}(x, A) \geq \int_A P^{km-s}(y, A)P^s(x, dy) \geq \varepsilon_{km-s}\rho(A)P^s(x, A) > 0$. Thus, for $[\pi]$-a.e. x, $P^{km}(x, A) > 0$ for all large k. This means that the Markov chain $\{X_{nm}, m = 0, 1, \ldots\}$ satisfies (1.1), (1.4) and (1.5) with $n_0 = 1$. From Proposition 2, there is a D_0 such that $\pi(D_0) = 1$ and, for each $x \in D_0$, $\Delta_{km}(x) \overset{\text{def}}{=} \sup_{C \in \mathscr{B}} |P^{km}(x, C) - \pi(C)| \to 0$ as $k \to \infty$. We also have $P^r(y, D_0^c) = 0$ for $0 \leq r \leq m - 1$ for $[\pi]$-a.e. y, since $0 = \pi(D_0^c) = \int P^r(y, D_0^c)\pi(dy)$ for $0 \leq r \leq m - 1$. For any n, write $n = km + r$ for $0 \leq r \leq m - 1$. Let $D_1 = \{x: P^r(x, D_0^c) = 0, 0 \leq r \leq m - 1\}$. Then $\pi(D_1) = 1$ and, for $x \in D_1$,

$$\begin{aligned}\Delta_n(x) &\leq \int_{D_0} \sup_{C \in \mathscr{B}} |P^{km}(y, C) - \pi(C)| P^r(x, dy) \\ &\leq \int_{D_0} \sup_{C \in \mathscr{B}} |P^{km}(y, C) - \pi(C)| \sum_{r=0}^{m-1} P^r(x, dy),\end{aligned}$$

which goes to zero as $k \to \infty$ by the bounded convergence theorem. Since $n \to \infty$ implies $k \to \infty$, this proves that $\Delta_n(x) \to 0$ for $x \in D_1$, where $\pi(D_1) = 0$. □

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 set $Q(x, C) = P^{n_0}(x, C)$ for $x \in \mathscr{X}$ and $C \in \mathscr{B}$. 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

$$A_r = \left\{x: \sum_{m=1}^{\infty} P^{mn_0 - r}(x, A) > 0\right\}, \qquad r = 0, 1, \ldots, n_0 - 1. \tag{2.29}$$

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

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

that is, 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 shows that the embedded Markov chain satisfies the conditions of Theorem 1 with the normalized restriction of π to A_0 as its invariant 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 conditions* (1.4) *and* (1.5) *of Theorem* 1 *with* π_0 *as an invariant probability measure and with the* n_0 *appearing in* (1.5) *equal to* 1.

PROOF. Condition (1.4) states that $\pi(\{x\colon P_x(T(A) < \infty) > 0\}) = 1$. Just the fact that this probability is positive and condition (1.1) 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,

$$\pi(C) = \int \pi(dx) Q(x, C) \quad \text{for all } C \in \mathscr{B}, \tag{2.30}$$

$$\pi_0(A_0) = 1 \tag{2.31}$$

and

$$Q(x, \cdot) \geq \varepsilon \rho(\cdot) \quad \text{for all } x \in A. \tag{2.32}$$

Notice that

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

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

$$x \notin A_0 \quad \text{implies} \quad Q(x, A_0) = 0. \tag{2.33}$$

From (2.30) and (2.33) we have the equality

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

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

$$\begin{aligned} \int_{\mathscr{X}} \pi_0(dx) Q(x, C) &= \frac{1}{\pi(A_0)} \int_{A_0} \pi(dx) Q(x, C) \\ &= \frac{1}{\pi(A_0)} \int_{\mathscr{X}} \pi(dx) Q(x, C \cap A_0) = \pi_0(C). \end{aligned} \tag{2.34}$$

Equations (2.34), (2.31) and (2.32) 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 n_0 successive transition functions of the chain converge to $\tilde{\pi}(C)$ for $[\pi_0]$-almost all x.

LEMMA 7. *Define*

$$B_0 = \left\{ x : x \in A_0,\ \sup_{C \in \mathscr{B}} |P^{mn_0}(x, C) - \pi_0(C)| \to 0 \text{ as } m \to \infty \right\}. \tag{2.35}$$

Under the conditions of Theorem 2,

$$\pi_0(B_0) = 1. \tag{2.36}$$

Moreover, for each $x \in B_0$,

$$\sup_{C \in \mathscr{B}} |P^{mn_0+r}(x, C) - \pi_r(C)| \to 0 \quad \text{as } m \to \infty \text{ for } r = 0, 1, \ldots, n_0 - 1, \tag{2.37}$$

and hence

$$\sup_{C \in \mathscr{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r}(x, C) - \tilde{\pi}(C) \right| \to 0 \quad \text{as } m \to \infty. \tag{2.38}$$

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

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

for $[\pi_0]$-almost all x. This establishes (2.36). For $r = 0, 1, \ldots, n_0 - 1$ and $x \in B_0$,

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

as $m \to \infty$, establishing (2.37). □

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

LEMMA 8. *Under the conditions of Theorem* 2,

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

and (2.38) *holds for* $[\tilde{\pi}]$*-almost all* x.

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

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

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

$$B_r = \{x\colon P^{n_0-r}(x, B_0) = 1\}, \qquad r = 1, 2, \ldots, n_0 - 1.$$

Let $x \in B_1$. Then

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

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 $x \in B_1$,

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

From (2.37) it follows that

$$\sup_{C\in\mathscr{B}} \left| P^{mn_0+r+n_0-1}(x, C) - \pi_r(C) \right| \to 0 \quad \text{as } m \to \infty \text{ for } [\pi_1]\text{-almost all } x. \tag{2.39}$$

As a consequence, as $m \to \infty$

$$\begin{aligned}&\sup_{C\in\mathscr{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} P^{mn_0+r}(x, C) - \tilde{\pi}(C) \right| \\ &\quad = \sup_{C\in\mathscr{B}} \left| \frac{1}{n_0} \sum_{r=0}^{n_0-1} \left(P^{mn_0+r}(x, C) - \pi_{r+1}(C) \right) \right| \to 0 \quad \text{for } [\pi_1]\text{-a.e. } x.\end{aligned} \tag{2.40}$$

A similar argument shows that (2.38) holds for $[\pi_r]$-almost all x and all r and hence for $[\tilde{\pi}]$-almost all x. □

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

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

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 $C \in \mathscr{B}$,

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

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 $\pi(\bigcup_{r=0}^{n_0-1} A_r) = 1$. 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 $C \in \mathscr{B}$,

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

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

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

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

$$B_f = \left\{ x\colon P_x\left\{ \frac{1}{n}\sum_{j=1}^{n} f(X_j) \to \int f(x)\, d\tilde{\pi}(x) \right\} = 1; \right.$$
$$\left. E_x\left(\frac{1}{n}\sum_{j=1}^{n} f(X_j) \right) \to \int f(x)\, d\tilde{\pi}(x) \right\}.$$

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) and (1.11). □

2.4. *Remarks on successive substitution sampling*. Theorems 1 and 2 pertain to arbitrary Markov chains. We now give a result that facilitates the use of our theorems when the Markov chain used is the one obtained from the successive substitution sampling algorithm, which is the most commonly used Markov chain in Bayesian statistics. We assume that, for each i, the conditional distributions $\pi_{X^{(i)}|\{X^{(j)},\, j\neq i\}}$ have densities, say $p_{X^{(i)}|\{X^{(j)},\, j\neq i\}}$, with respect to some dominating measure ρ_i.

THEOREM 6. *Consider the successive substitution sampling algorithm for generating observations from the joint distribution* π *of* $(X^{(1)}, \ldots, X^{(p)})$ *as described in Section* 1. *Suppose that, for each* $i = 1, \ldots, p$, *there is a set* A_i, *with* $\rho_i(A_i) > 0$, *and a* $\delta > 0$ *such that, for each* $i = 1, \ldots, p$,

$$p_{X^{(i)}|\{X^{(j)},\, j\neq i\}}(x^{(1)}, \ldots, x^{(p)}) > 0 \tag{2.41}$$

whenever

$$x^{(1)} \in A_1, \ldots, x^{(i)} \in A_i, \quad \textit{and} \quad x^{(i+1)}, \ldots, x^{(p)} \textit{ are arbitrary},$$

and

$$p_{X^{(i)}|\{X^{(j)},\, j\neq i\}}(x^{(1)}, \ldots, x^{(p)}) > \delta \quad \textit{whenever } x^{(j)} \in A_j,\ j = 1, \ldots, p. \tag{2.42}$$

Then conditions (1.4) *and* (1.5) *are satisfied with* $n_0 = 1$. *Thus*, (1.6) *is also satisfied*, *and the conclusions of Theorems* 1 *and* 2 *hold*.

PROOF. Let $p_i = p_{X^{(i)}|\{X^{(j)}, j \neq i\}}$ and $A = A_1 \times \cdots \times A_p$. The transition function used in successive substitution sampling is given by

$$P(x, A) = \int_A p_1(y^{(1)}, x^{(2)}, \ldots, x^{(p)}) \times p_2(y^{(1)}, y^{(2)}, x^{(3)}, \ldots, x^{(p)}) \cdots p_p(y^{(1)}, \ldots, y^{(p)})\, d\rho_1(y^{(1)}) \cdots d\rho_p(y^{(p)}).$$

It is now easy to see that condition (2.41) verifies (1.4) for all starting points x and that (2.42) verifies (1.5) with $n_0 = 1$. □

We note that condition (2.41) is often checked for all $(x^{(1)}, \ldots, x^{(p)})$.

Conditional distributions need not determine the joint distribution. 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 give positive probability to a rectangle set.

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

$$\begin{aligned} &\tfrac{1}{2} \quad \text{at } x_1 + 1, \\ &\tfrac{1}{2} \quad \text{at } x_1 - 1. \end{aligned} \tag{2.43}$$

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

$$\begin{aligned} &\frac{p(x_2 + 1)}{p(x_2 + 1) + p(x_2 - 1)} \quad \text{at } x_2 + 1 \quad \text{and} \\ &\frac{p(x_2 - 1)}{p(x_2 + 1) + p(x_2 - 1)} \quad \text{at } x_2 - 1. \end{aligned} \tag{2.44}$$

It can be seen that the two conditional distributions (2.44) and (2.43) 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 $m + r$, $m = \ldots, -1, 0, 1, \ldots$, defined by $p_r(m + r) = (1/c_r)p(m + r)$. 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 (2.43). It is easy to see that the distribution of $Y_1(r)$ given

$Y_2(r)$ is that given in (2.44), 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 (2.44) and (2.43). Let $f(r)$ be any probability density on $[0, 1)$. Let R have density function $f(r)$ and set $(Z_1, Z_2) = (Y_1(R), Y_2(R))$. Clearly the conditional distributions of (Z_1, Z_2) are as in (2.44) and (2.43). The marginal distribution function of Z_1 is given by

$$P(Z_1 \le x) = \int_{[0,1)} \left(\sum_{m;\, m+r \le x} p_r(m + r) \right) f(r)\, dr = \int_{-\infty}^{x} \frac{p(y) f(y - [y])}{c_{y-[y]}}\, dy.$$

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 $X_n^{[1]}, X_n^{[2]}, \ldots, X_n^{[G]}$ of independent variables. At another extreme, one can generate a very long sample $X_0, X_1, X_2, \ldots, X_{nG}$ and use $X_n, X_{2n}, \ldots, X_{Gn}$, 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 use of (3.1) is preferable, 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 often 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 minipulations 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, that is, 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

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

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

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

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

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

This is the well-known "ratio estimate" in importance sampling theory. The key here is that its calculation requires only knowledge of the ratio $d\nu_{h,X}/d\nu_{0,X}$ 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 (1994).] The fact that this likelihood cancels means that the estimation of the expectation under the prior ν_h requires only the recomputation of r weights, and this can be done very quickly.

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. For some applications we have been able to do the computations quickly enough to display dynamically 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 (1990), for r as large as 500 on a workstation doing about 1.5 million floating point operations per second. [See Doss and Narasimhan (1994).]

Acknowledgments. We are very grateful to the reviewers for their helpful remarks, and especially to an Associate Editor for a particularly careful reading of the paper and for pointing out some simplifications in our proofs.

REFERENCES

AMIT, Y. (1991). On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Anal.* **38** 82–99.

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

ATHREYA, K. B., DOSS, H. and SETHURAMAN, J. (1992). A proof of convergence of the Markov chain simulation method. Technical Report 868, Dept. Statistics, Florida State Univ.

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. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

CHAN, K. S. (1993). Asymptotic behavior of the Gibbs sampler. *J. Amer. Statist. Assoc.* **88** 320–326.

CHUNG, K. L. (1967). *Markov Chains*, 2nd ed. Springer, New York.

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

DOSS, H. (1994). Bayesian nonparametric estimation for incomplete data via successive substitution sampling. *Ann. Statist.* **22** 1763–1786.

DOSS, H. and NARASIMHAN, B. (1994). Bayesian Poisson regression using the Gibbs sampler: sensitivity analysis through dynamic graphics. Technical Report 895, Dept. Statistics, Florida State Univ.

FELLER, W. (1950). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

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

GILKS, W. R. and WILD, P. (1992). Adaptive rejection sampling for Gibbs sampling. *J. Roy. Statist. Soc. Ser. C* **41** 337–348.

GOODMAN, J. and SOKAL, A. D. (1989). Multigrid Monte Carlo method. Conceptual foundations. *Phys. Rev. D* **40** 2035–2071.

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.

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

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. (1992). On the convergence of successive substitution sampling. *Journal of Computational and Graphical Statistics* **1** 111–127.

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. (1990). *Lisp-Stat*. Wiley, New York.

TIERNEY, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **22** 1701–1762.

THOMAS, A., SPIEGELHALTER, D. and GILKS, W. (1992). BUGS: a program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (J. Bernardo, J. Berger, A. Dawid and A. F. M. Smith, eds.) 837–842. Clarendon, Oxford.

KRISHNA B. ATHREYA
DEPARTMENT OF STATISTICS
IOWA STATE UNIVERSITY
AMES, IOWA 50011

HANI DOSS
DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
COLUMBUS, OHIO 43210

JAYARAM SETHURAMAN
DEPARTMENT OF STATISTICS
AND STATISTICAL CONSULTING CENTER
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306-3033