MARKOV CHAINS: BASIC THEORY

1. Markov Chains and their Transition Probabilities

1.1. Definition and First Examples.

Definition 1. A (discrete-time) Markov chain with (finite or countable) state space $\mathscr{X}$ is a sequence $X_0, X_1, \ldots$ of $\mathscr{X}-$valued random variables such that for all states $i, j, k_0, k_1, \cdots$ and all times $n = 0, 1, 2, \ldots$,

$$P(X_{n+1} = j \parallel X_n = i, X_{n-1} = k_{n-1}, \ldots) = p(i, j) \tag{1}$$

where $p(i, j)$ depends only on the states i, j, and not on the time n or the previous states $k_{n-1}, n-2, \ldots$. The numbers $p(i, j)$ are called the *transition probabilities* of the chain.

Example 1. The *simple random walk* on the integer lattice $\mathbb{Z}^d$ is the Markov chain whose transition probabilities are

$$p(x, x \pm e_i) = 1/(2d) \quad \forall\, x \in \mathbb{Z}^d$$

where $e_{1,} e_2, \ldots e_d$ are the standard unit vectors in $\mathbb{Z}^d$. In other terms, the simple random walk moves, at each step, to a randomly chosen nearest neighbor.

Example 2. The *random transposition* Markov chain on the permutation group $\mathscr{S}_N$ (the set of all permutations of N cards) is a Markov chain whose transition probabilities are

$$p(x, \sigma x) = 1/\binom{N}{2} \quad \text{for all transpositions } \sigma;$$
$$p(x, y) = 0 \quad \text{otherwise.}$$

A *transposition* is a permutation that exchanges two cards. Notice that there are exactly $\binom{N}{2}$ transpositions. Thus, the Markov chain proceeds by the following rule: at each step, choose two different cards at random and switch them.

Example 3. The *Ehrenfest urn model* with N balls is the Markov chain on the state space $\mathscr{X} = \{0, 1\}^N$ that evolves as follows: At each time $n = 1, 2, \ldots$ a random index $j \in [N]$ is chosen, and the jth coordinate of the last state is flipped. Thus, the transition probabilities are

$$p(x, y) = 1/N \quad \text{if the vectors } x, y \text{ differ in exactly 1 coordinate}$$
$$= 0 \quad \text{otherwise.}$$

The Ehrenfest model is a simple model of particle diffusion: Imagine a room with two compartments 0 and 1, and N molecules distributed throughout the two compartments (customarily called *urns*). At each time, one of the molecules is chosen at random and moved from its current compartment to the other.

Example 4. Let $\xi_1, \xi_2, \ldots$ be independent, identically distributed, positive integer-valued random variables with common distribution $\{q_m\}_{m \geq 1}$. Think of these as being the lifetimes of a sequence of AAA batteries that I use in my wireless keyboard. Whenever a battery fails, I immediately replace it by a new one. Consequently, the partial sums $S_n := \sum_{j=1}^n \xi_j$ are the times at

which batteries are replaced. In this context, the sequence of random variables $\{S_n\}_{n\geq 0}$ is called a *renewal process*.

There are several interesting Markov chains associated with a renewal process: (A) The *age process* $A_1, A_2, \ldots$ is the sequence of random variables that record the time elapsed since the last battery failure, in other words, A_n is the age of the battery in use at time n. At each step, the age process either increases by $+1$, or it jumps to 0. It is not difficult to verify that it is a Markov chain with transition probabilities

$$p(m,m+1) = \sum_{j=m+1}^{\infty} q_j / \sum_{j=m}^{\infty} q_j,$$
$$p(m,0) = q_m / \sum_{j=m}^{\infty} q_j.$$

(B) The *residual lifetime* process $R_1, R_2, \ldots$ is the sequence of random variables that record the time until the next battery failure, that is, the remaining lifetime of the battery currently in use. (If a new battery is installed at time n, the residual lifetime is the lifetime of the new battery, not 0.) The sequence R_n is a Markov chain with transition probabilities

$$p(m,m-1) = 1 \quad \text{if } m \geq 2;$$
$$p(1,m) = q_m \quad \text{for all } m \geq 1.$$

Proposition 1. *If X_n is a Markov chain with transition probabilities $p(x,y)$ then for every sequence of states $x_0, x_1, \ldots, x_{n+m}$,*

$$P(X_{m+i} = x_{m+i} \ \forall\ 0 < i \leq n \parallel X_i = x_i \ \forall\ 0 \leq i \leq m) = \prod_{i=1}^{n} p(x_{m+i-1}, x_{m+i}). \tag{2}$$

*Consequently, the $n-$*step transition probabilities

$$p_n(x,y) := P(X_{n+m} = y \parallel X_m = x) \tag{3}$$

depend only on the time lag n and the initial and terminal states x, y, but not on m.

Proof. The first statement can be proved by a completely routine induction argument, using the definition of a Markov chain and elementary properties of conditional probabilities. The second follows from the first, by summing over all possible sequences x_{m+i} of intermediate states: the right side of equation (2) makes it clear that this sum does not depend on m, since the factors in the product depend only on the transitions x_i, x_{i+1} made, and not on the *times* at which they are made. □

1.2. **Chapman-Kolmogorov Equations and the Transition Probability Matrix.** Assume henceforth that $\{X_n\}_{n\geq 0}$ is a discrete-time Markov chain on a state space $\mathscr{X}$ with transition probabilities $p(i,j)$. Define the *transition probability matrix* $\mathbb{P}$ of the chain to be the $\mathscr{X} \times \mathscr{X}$ matrix with entries $p(i,j)$, that is, the matrix whose ith row consists of the transition probabilities $p(i,j)$ for $j \in \mathscr{X}$:

$$\mathbb{P} = (p(i,j))_{i,j\in\mathscr{X}} \tag{4}$$

If $\mathscr{X}$ has N elements, then $\mathbb{P}$ is an $N \times N$ matrix, and if $\mathscr{X}$ is infinite, then $\mathbb{P}$ is an infinite by infinite matrix. Also, the *row sums* of $\mathbb{P}$ must all be 1, by the law of total probabilities. A matrix with this property is called *stochastic*.

Definition 2. A *nonnegative matrix* is a matrix with nonnegative entries. A *stochastic matrix* is a square nonnegative matrix all of whose row sums are 1. A *substochastic matrix* is a square nonnegative matrix all of whose row sums are ≤ 1. A *doubly stochastic* matrix is a stochastic matrix all of whose *column* sums are 1.

Observe that if you start with a stochastic matrix and delete the rows and columns indexed by a set of states i, you get a substochastic matrix. The resulting substochastic matrix contains information about the so-called *first-passage* times to various states, as you will see in one of the homework problems.

Proposition 2. *The $n-$step transition probabilities $p_n(i,j)$ are the entries of the nth power $\mathbb{P}^n$ of the matrix $\mathbb{P}$. Consequently, the $n-$step transition probabilities $p_n(i,j)$ satisfy the* Chapman-Kolmogorov equations

$$p_{n+m}(i,j) = \sum_{k\in\mathscr{X}} p_n(i,k)p_m(k,j). \tag{5}$$

Proof. It is easiest to start by directly proving the Chapman-Kolmogorov equations, by a double induction, first on n, then on m. The case $n = 1, m = 1$ follows directly from the definition of a Markov chain and the law of total probability (to get from i to j in two steps, the Markov chain has to go through *some* intermediate state k). The induction steps are left as an exercise. Once the Chapman-Kolmogorov equation is established, it follows that the $n-$step transition probabilities $p_n(x,y)$ are the entries of $\mathbb{P}^n$, because equation (5) is the rule for matrix multiplication. □

Suppose now that the initial state X_0 is random, with distribution ν, that is,

$$P^\nu\{X_0 = i\} = \nu(i) \quad \text{for all states } i \in \mathscr{X}.$$

(Note: Henceforth when a probability distribution ν is used as a superscript as above, it denotes the initial distribution, that is, the distribution of X_0.) Then by the Chapman-Kolmogorov equations and the law of total probability,

$$P^\nu\{X_n = j\} = \sum_i \nu(i)p_n(i,j),$$

equivalently, if the initial distribution is ν^T (here we are viewing probability distributions on $\mathscr{X}$ as row vectors) then the distribution after n steps is $\nu^T\mathbb{P}^n$. Notice that if there is a probability distribution ν on $\mathscr{X}$ such that $\nu^T = \nu^T\mathbb{P}$, then $\nu^T = \nu^T\mathbb{P}^n$ for all $n \geq 1$. Consequently, if the Markov chain has initial distribution ν then the marginal distribution of X_n will be ν for all $n \geq 1$. For this reason, such a probability distribution is called *stationary*:

Definition 3. A probability distribution π on $\mathscr{X}$ is *stationary* if

$$\pi^T = \pi^T\mathbb{P} \tag{6}$$

1.3. **Accessibility and Communicating Classes.**

Definition 4. A state j is said to be *accessible* from state i if there is a positive-probability path from i to j, that is, if there is a finite sequence of states $k_0, k_1, \ldots, k_m$ such that $k_0 = i$, $k_m = j$, and $p(k_t, k_{t+1}) > 0$ for each $t = 0, 1, \ldots, m-1$. States i and j are said to *communicate* if each is accessible from the other. This relation is denoted by $i \longleftrightarrow j$.

Fact 1. *Communication is an equivalence relation. In particular, it is* transitive*: if i communicates with j and j communicates with k then i communicates with k.*

The proof is an exercise. It follows that the state space $\mathscr{X}$ is uniquely partitioned into *communicating classes* (the equivalence classes of the relation $\longleftrightarrow$). If there is only one communicating class (that is, if every state is accessible from every other) then the Markov chain (or its transition probability matrix) is said to be *irreducible*. In general, if there is more than one communicating class, then states in one communicating class $\mathscr{C}_1$ may be accessible from states in another class $\mathscr{C}_2$; however, in such a case no state of $\mathscr{C}_2$ can be accessible from a state of $\mathscr{C}_1$ (why?).

Definition 5. The *period* of a state i is the greatest common divisor of the set $\{n \in \mathbb{N} : p_n(i,i) > 0\}$. If every state has period 1 then the Markov chain (or its transition probability matrix) is called *aperiodic*.

Note: If i is not accessible from itself, then the period is the g.c.d. of the empty set; by convention, we define the period in this case to be $+\infty$. Example: Consider simple random walk on the integers. If at time zero the walk starts in state $X_0 = 0$ then at any subsequent *even* time the state must be an even integer, and at any *odd* time the state must be an odd integer (why?). Consequently, all states have period 2.

Fact 2. *If states i, j communicate, then they must have the same period. Consequently, if the Markov chain is irreducible, then all states have the same period.*

The proof is another easy exercise. There is a simple test to check whether an irreducible Markov chain is aperiodic: If there is a state i for which the $1-$step transition probability $p(i,i) > 0$, then the chain is aperiodic.

Fact 3. *If the Markov chain has a stationary probability distribution π for which $\pi(i) > 0$, and if states i,j communicate, then $\pi(j) > 0$.*

Proof. It suffices to show (why?) that if $p(i,j) > 0$ then $\pi(j) > 0$. But by definition (6), $\pi(j) = \sum_k \pi(k)p(k,j) \geq \pi(i)p(i,j)$. □

2. Finite State Markov Chains

2.1. **Irreducible Markov chains.** If the state space is finite and all states communicate (that is, the Markov chain is irreducible) then in the long run, regardless of the initial condition, the Markov chain must settle into a steady state. Formally,

Theorem 3. *An irreducible Markov chain X_n on a* finite *state space $\mathscr{X}$ has a* unique *stationary distribution π. Furthermore, if the Markov chain is not only irreducible but also aperiodic, then for any initial distribution ν,*

$$\lim_{n\to\infty} P^{\nu}\{X_n = j\} = \pi(j) \quad \forall\, j \in \mathscr{X} \tag{7}$$

The remainder of this section is devoted to the proof of this theorem. Assume throughout that the hypotheses of Theorem 3 are met, and let $\mathbb{P}$ be the transition probability matrix of the Markov chain. We will prove Theorem 3 by studying the action of the transition probability matrix on the set $\mathscr{P} = \mathscr{P}_{\mathscr{X}}$ of probability distributions on $\mathscr{X}$. Recall from sec. 1.2 above that if ν^T is the initial distribution of the Markov chain then $\nu^T\mathbb{P}^n$ is the distribution after n steps. Thus, the natural action of the transition probability matrix on $\mathscr{P}$ is

$$\nu^T \mapsto \nu^T\mathbb{P}.$$

Notice that if v^T is a probability vector, then so is $v^T\mathbb{P}$, because

$$\begin{aligned}\sum_j (v^T\mathbb{P})_j &= \sum_j \sum_i v(i)p(i,j)\\ &= \sum_i \sum_j v(i)p(i,j)\\ &= \sum_i v(i)\sum_j p(i,j)\\ &= \sum_i v(i),\end{aligned}$$

the last because the row sums of $\mathbb{P}$ are all 1. This implies that the mapping $v^T \mapsto v^T\mathbb{P}$ takes the set $\mathscr{P}$ into itself. It also implies (by induction on n) that $\mathbb{P}^n$ is a stochastic matrix for every $n = 1, 2, \ldots$, because each row of $\mathbb{P}$ is a probability vector.

2.2. **The $N-$Simplex.** The set $\mathscr{P}_{\mathscr{X}}$ is called the *$N-$simplex*, where N is the cardinality of the state space $\mathscr{X}$: it is the subset of $\mathbb{R}^N$ gotten by intersecting the first orthant (the set of all vectors with nonnegative entries) with the hyperplane consisting of all vectors whose entries sum to 1. The crucial geometric fact about $\mathscr{P}$ is this:

Proposition 4. *The $N-$simplex $\mathscr{P}$ is a closed and bounded subset of $\mathbb{R}^N$. Consequently, by the Heine-Borel Theorem, it is* compact.

Proof. Exercise: Show that $\mathscr{P}$ is closed and bounded. (This is an good test of your command of elementary real analysis.) □

2.3. **The Krylov-Bogoliubov Argument.** There is a very simple argument, due to Krylov and Bogoliubov, that a stationary distribution always exists. It is a very useful argument, because it generalizes to many other contexts. Furthermore, it shows that there are stationary distributions even for finite-state Markov chain that are not irreducible.

The argument turns on the fact that the probability simplex $\mathscr{P}$ is compact. This implies that it has the *Bolzano-Weierstrass* property: Any sequence of vectors in $\mathscr{P}$ has a convergent subsequence. Fix a probability vector $v \in \mathscr{P}$ (it doesn't matter what), and consider the so-called *Cesaro averages*

$$v_n^T := n^{-1}\sum_{k=1}^{n} v^T\mathbb{P}^k. \tag{8}$$

Observe that each average v_n^T is a probability vector (because an average of probability vectors is always a probability vector), and so each v_n^T is an element of $\mathscr{P}$. Consequently, the sequence v_n^T has a convergent subsequence:

$$\lim_{k\to\infty} v_{n_k}^T = \pi^T. \tag{9}$$

Claim: The limit of any subsequence of v_n^T is a stationary distribution for $\mathbb{P}$.

Proof. Denote the limit by π, as in (9). Since the mapping $\mu^T \mapsto \mu^T\mathbb{P}$ is continuous (exercise; or see the proof of Theorem 6 below),

$$\begin{aligned}
\pi^T\mathbb{P} &= \lim_{k\to\infty} \nu_{n_k}^T\mathbb{P} \\
&= \lim_{k\to\infty} n_k^{-1}\sum_{j=1}^{n_k} \nu^T\mathbb{P}^j\mathbb{P} \\
&= \lim_{k\to\infty} n_k^{-1}\sum_{j=2}^{n_k+1} \nu^T\mathbb{P}^j \\
&= \lim_{k\to\infty} n_k^{-1}\left(\sum_{j=1}^{n_k} \nu^T\mathbb{P}^j + \nu^T\mathbb{P}^{n_k+1} - \nu^T\mathbb{P}\right) \\
&= \lim_{k\to\infty} n_k^{-1}\sum_{j=1}^{n_k} \nu^T\mathbb{P}^j \\
&= \pi^T.
\end{aligned}$$

(In words: Multiplying the Cesaro average by $\mathbb{P}$ has the effect of changing only the first and last term in the average. When this is divided by n_k, it converges to zero in the limit.) Thus, π^T is a stationary distribution. □

2.4. **Total Variation Metric.** To prove uniqueness of the stationary distribution under the hypotheses of Theorem 3, we will investigate more closely the action of the transition probability matrix on the simplex. The most natural metric (distance function) on the simplex $\mathscr{P}$ is not the usual Pythagorean distance, but rather the *total variation* metric, or taxicab distance. This is defined as follows: For any two probability distributions $\nu, \mu \in \mathscr{P}$,

$$d(\mu,\nu) = \|\mu-\nu\|_{TV} := \frac{1}{2}\sum_{i\in\mathscr{X}} |\nu(i)-\mu(i)|$$

The factor 1/2 is a convention, but a long-established one (it ensures that the distance is never larger than one). It is an *exercise* to show that the following is an equivalent definition:

$$\|\mu-\nu\|_{TV} = \max_{A\subset\mathscr{X}} (\mu(A)-\nu(A))$$

(Hint: Use the fact that both μ and ν are *probability* distributions.) Next term you will learn more about the total variation metric; all we need for now is that the two formulas above give the same value. One other thing (also very easy to check): The total variation metric is equivalent to the Pythagorean metric, in the sense that a sequence of probability vectors converges in the total variation metric if and only if it converges in the Pythagorean metric.

Proposition 5. *Assume that the entries of $\mathbb{P}$ are all strictly positive. Then the mapping $\nu^T \mapsto \nu^T\mathbb{P}$ is a strict contraction of the simplex $\mathscr{P}$ relative to total variation distance, that is, there exists $0<\alpha<1$ such that for any two probability vectors μ,ν*

$$\|\nu^T\mathbb{P} - \mu^T\mathbb{P}\|_{TV} \le \alpha\|\nu^T - \mu^T\|_{TV}$$

Proof. Since every entry of $\mathbb{P}$ is strictly positive, there is a real number $\varepsilon > 0$ such that $p(i,j) \ge \varepsilon$ for every pair of states i, j. Notice that $N\epsilon \le 1$, where N is the total number of states, because the row sums of $\mathbb{P}$ are all 1. We may assume (by choosing a slightly smaller value of $\varepsilon > 0$, if

necessary) that $1-N\epsilon>0$. Define $q(i,j)=(p(i,j)-\varepsilon)/(1-N\varepsilon)$, and let $\mathbb{Q}$ be the matrix with entries $q(i,j)$. Then $\mathbb{Q}$ is a stochastic matrix, because its entries are nonnegative (by the choice of ε), and for every state i,

$$\sum_j q(i,j)=(1-N\varepsilon)^{-1}\sum_j p(i,j)-(1-N\varepsilon^{-1})\sum_j \varepsilon=1,$$

since the row sums of $\mathbb{P}$ are all 1. Observe that $\mathbb{P}=(1-N\varepsilon)\mathbb{Q}+\varepsilon J$, where J is the $N\times N$ matrix with all entries 1.

Now consider the total variation distance between $\nu^T\mathbb{P}$ and $\mu^T\mathbb{P}$. Using the fact that $\sum_i \nu(i)=\sum_i \mu(i)=1$, we have

$$\begin{aligned}2\|\nu^T\mathbb{P}-\mu^T\mathbb{P}\|_{TV}&=\sum_j|(\nu^T\mathbb{P})_j-(\mu^T\mathbb{P})_j|\\&=\sum_j|\sum_i(\nu(i)p(i,j)-\mu(i)p(i,j)|\\&=\sum_j|\sum_i(\nu(i)-\mu(i))q(i,j)(1-N\varepsilon)|.\end{aligned}$$

Factor out $(1-N\varepsilon):=\alpha$. What's left is

$$\begin{aligned}\sum_j|\sum_i(\nu(i)-\mu(i))q(i,j)|&\le\sum_j\sum_i|\nu(i)-\mu(i)|q(i,j)\\&=\sum_i|\nu(i)-\mu(i)|\sum_j q(i,j)\\&=\sum_i|\nu(i)-\mu(i)|\\&=2\|\nu^T-\mu^T\|_{TV}.\end{aligned}$$

□

2.5. **Contraction Mapping Fixed Point Theorem.** What do we gain by knowing that the action of the transition probability matrix on the simplex is a contraction? First, it tells us that if we start the Markov chain in two different initial distributions, then the distributions after one step are closer than they were to start. Consequently, by induction, after n steps they are even closer: in fact, the total variation distance will decrease by a factor of α at each step, and so will approach zero exponentially quickly as $n\to\infty$. This means that the Markov chain will ultimately "forget" its initial distribution.

Variations of this argument occur frequently not only in probability theory but in all parts of mathematics. Following is an important theorem about contractive mappings that formalizes the conclusions of the preceding argument.

Theorem 6. *Let (S,d) be a compact metric space, and $F:S\to S$ a strict contraction, that is, a function such that for some real number $\alpha<1$,*

$$d(F(x),F(y))\le\alpha d(x,y)\quad \textit{for all } x,y\in S. \tag{10}$$

Then F has a unique fixed point $z\in S$ (that is, a point such that $F(z)=z$), and the orbit of every point $x\in S$ converges to z, that is, if F^n is the nth iterate of F, then

$$\lim_{n\to\infty}F^n(x)=z. \tag{11}$$

Proof. First, notice that if F is a contraction, then it must be continuous. (Exercise: check this.) Second, if F is strictly contractive with contraction constant α as in (10), then for any point $x \in S$ and every $n = 1, 2, \ldots,$

$$d(F^n(x), F^{n+1}(x)) \le \alpha^n d(x, F(x)); \tag{12}$$

this follows from the assumption (10), by an easy induction argument. Now because the space S is compact, it has the *Bolzano-Weierstrass* property: every sequence has a convergent subsequence. Hence, for any point $x \in S$ the sequence $\{F^n(x)\}_{n\ge1}$ has a convergent subsequence. The limit z of any such subsequence must be a fixed point of F. Here is why: If

$$z = \lim_{k\to\infty} F^{n_k(x)}$$

exists, then by continuity of F,

$$F(z) = \lim_{k\to\infty} F^{n_k+1}(x);$$

but by (12),

$$d(F^{n_k}(x), F^{n_k+1}(x)) \le \alpha^{n_k} d(x, F(x)),$$

and this converges to 0 as $k \to \infty$, since $\alpha < 1$. Consequently, the two sequences $F^{n_k}(x)$ and $F^{n_k+1}(x)$ cannot converge to different limits, and so it follows that $z = F(z)$.

This proves that the limit of any convergent subsequence of any orbit $F^n(x)$ must be a fixed point of F. To complete the proof, it suffices to show there is only one fixed point. (Exercise: Why does this imply that every orbit $F^n(x)$ must converge?) Suppose there were two fixed points

$$z_1 = F(z_1) \quad \text{and} \quad z_2 = F(z_2).$$

By the assumption (10),

$$d(z_1, z_2) = d(F(z_1), F(z_2)) \le \alpha d(z_1, z_2).$$

Since $\alpha < 1$, it must be that $d(z_1, z_2) = 0$, that is, z_1 and z_2 must be the same point. □

2.6. **Proof of Theorem 3.** We have now shown that (a) if the transition probability matrix $\mathbb{P}$ has strictly positive entries then the mapping $v^T \mapsto v^T\mathbb{P}$ is a strict contraction of the simplex $\mathscr{P}$, (b) a strict contraction of a compact metric space has a unique fixed point, and (c) all orbits approach the fixed point. It follows that if $\mathbb{P}$ has strictly positive entries then the conclusions of Theorem 3 all hold. Thus, it remains to show how to relax the requirement that the entries of $\mathbb{P}$ are strictly positive.

Lemma 7. *Let $\mathbb{P}$ be the transition probability matrix of an irreducible, aperiodic, finite-state Markov chain. Then there is an integer m such that for all $n \ge m$, the matrix $\mathbb{P}^n$ has strictly positive entries.*

This is where the hypothesis of *aperiodicity* is needed. The result is definitely not true if the Markov chain is periodic: for example, consider the two-state Markov chain with transition probability matrix

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

(Exercise: Check what happens when you take powers of this matrix.)

Proof of Theorem 3. The proof of Lemma 7 will be given later. For now, let's take it as true; then if $\mathbb{P}$ is aperiodic and irreducible, as assumed in Theorem 3, there exists an integer $m \ge 1$ such that $\mathbb{Q} := \mathbb{P}^m$ has strictly positive entries. We have already observed that powers of a stochastic matrix are also stochastic matrices, and so $\mathbb{Q}$ satisfies the hypotheses of Proposition 5. Hence,

$\mathbb{Q}$ is strictly contractive on the simplex $\mathscr{P}$. Therefore, by the Contraction Mapping Fixed Point Theorem, there exists a unique probability vector π^T such that

$$\pi^T = \pi^T \mathbb{Q} = \pi^T \mathbb{P}^m, \tag{13}$$

and such that for all $v^T \in \mathscr{P}$,

$$\lim_{n\to\infty} v^T \mathbb{Q}^n = \lim_{n\to\infty} v^T \mathbb{P}^{nm} = \pi^T.$$

This last applies not only to v^T, but also to $v^T\mathbb{P}, v^T\mathbb{P}^2, \ldots$, since these are all probability vectors. Consequently, for every $k = 0, 1, \ldots, m-1$,

$$\lim_{n\to\infty} v^T \mathbb{P}^{nm+k} = \pi^T.$$

Now if a sequence of vectors $\{v_n\}_{n\geq 1}$ has the property that the subsequences $\{v_{nm+k}\}_{n\geq 1}$, for $k = 0, 1, \ldots m-1$, all converge to the same limit w, then the entire sequence must converge to w. (Exercise: Explain why.) Thus,

$$\lim_{n\to\infty} v^T \mathbb{P}^n = \pi^T \tag{14}$$

This is equivalent to the statement (7).

It remains to show that π is a stationary distribution, and that π is the *only* stationary distribution. Set $\mu^T := \pi^T\mathbb{P}$; then by equation (13) (multiply both sides by $\mathbb{P}$ on the right), $\mu^T = \mu^T\mathbb{Q}$, and so μ is a stationary distribution for $\mathbb{Q}$. But π^T is the *unique* stationary distribution of $\mathbb{Q}$, since $\mathbb{Q}$ is strictly contractive on the simplex; thus, $\mu = \pi$, and so π is a stationary distribution for $\mathbb{P}$. That it is the *only* stationary distribution follows from (14). □

The proof of Lemma 7 will make use of the following consequence of the *Euclidean algorithm*. You will find a proof in any respectable book on elementary number theory or abstract algebra.

Lemma 8. *If d is the greatest common divisor of $m, n \in \mathbb{N}$, then there exist integers s, t (not necessarily positive) such that*

$$d = sm + tn.$$

More generally, if d is the greatest common divisor of a finite set of integers $m_1, m_2, \ldots, m_r$, then there exist integers $t_1, t_2, \ldots t_r$ such that

$$d = \sum_{i=1}^{r} t_i m_i.$$

Proof of Lemma 7. First, notice that it is enough to show that the diagonal entries of $\mathbb{P}^n$ are all eventually positive. Here is why: Suppose that $p_n(x,x) > 0$ for all $n \geq m = m(x)$. Let y be any other state. Since the Markov chain is irreducible, there is an integer $k = k(x,y)$ such that $p_k(x,y) > 0$. But it then follows from the Chapman-Kolmogorov equations that

$$p_{k+n}(x,y) \geq p_n(x,x)p_k(x,y) > 0 \quad \forall\ n \geq m.$$

Thus, if $n \geq \max_x m(x) + \max_{x,y} k(x,y)$ then all entries of $\mathbb{P}^n$ will be positive.

Therefore, we need only show that for each state x the return probabilities $p_n(x,x)$ are positive for all large n. Equivalently, we must show that the set

$$A_x := \{n \geq 1 : p_n(x,x) > 0\}$$

contains all but finitely many elements of the natural numbers $\mathbb{N}$. For this, we will use two basic properties of A_x: First, by the Chapman-Kolmogorov equations, A_x is closed under addition, that is, if $m, n \in A_x$ are two elements of A_i then $m + n \in A_x$. This implies also that A_x is closed under

scalar multiplication by positive integers, that is, if $m \in A_x$ then all multiples of m are elements of A_x. Second, the greatest common divisor of the numbers in A_x is 1, because by hypothesis the Markov chain is aperiodic. This implies that there is a finite subset $\{m_i\}_{1\le i\le r}$ of A_x whose greatest common divisor is 1. (Exercise: Explain why.) Hence, by Lemma 8, there are integers t_i such that

$$\sum t_i m_i = 1.$$

Set

$$t_* = \max_{1\le i\le r} |t_i| \quad \text{and} \quad m_+ = \sum_{i=1}^{r} m_i.$$

Any integer n must lie between successive multiples of m_+, so there exist a nonnegative integer k and an integer $0 \le s < m_+$ such that $n = k m_+ + s$. Using the fact that $\sum_i t_i m_i = 1$, we conclude that

$$n = \sum_{i=1}^{r} k m_i + s = \sum_{i=1}^{r} (k + s t_i) m_i.$$

If $k \ge m_+ t_*$ then all of the coefficients $k + s t_i$ in this sum will be nonnegative. Thus, A_x contains all integers larger than $m_+^2 t_*$. □

This argument actually proves the following fact, which we record for later use:

Corollary 9. *If A is a subset of the natural numbers that is closed under addition, and if the greatest common divisor of the elements of A is 1, then A contains all but at most finitely many of the natural numbers.*

3. Stopping Times, Strong Markov Property

Definition 6. Let $\{X_n\}_{n\ge 0}$ be a Markov chain on finite or countable state space $\mathscr{X}$. A *stopping time* is a random variable T with values in the set $\mathbb{Z}_+ \cup \{\infty\}$ such that for every $m \in \mathbb{Z}_+$, the event $\{T = m\}$ is determined by the values $X_0, X_1, \dots X_m$.

Example 5. The *first passage time* to a state x is the random variable $T_x = T_x^1$ whose value is the first time $n \ge 1$ that $X_n = x$, or ∞ if there is no such (finite) n. The kth passage time is the random variable T_x^k that records the time of the kth visit to x, or ∞ if the Markov chain does not visit the state x at least k times. Clearly, T_x^k is a stopping time.

Example 6. Here is a random time that is *not* a stopping time: Fix a state x, and let L be the *last* time $n \le 100$ that $X_n = x$. This isn't (in general) a stopping time, because (for instance) to determine whether $L = 97$, you would need to know not only the first 97 steps, but also the 98th through 100th.

Proposition 10. *[Strong Markov Property.] Let T be a stopping time for the Markov chain $\{X_n\}_{n\ge 0}$. Then the Markov chain "regenerates" at time T, that is, the future $X_{T+1}, X_{T+2}, \dots$ is conditionally independent of the past $X_0, X_1, \dots, X_{T-1}$ given the value of T and the state $X_T = x$ at time T. More precisely, for any $m < \infty$ and all states $x_0, x_1, \dots x_{n+m} \in \mathscr{X}$ such that $T = m$ is possible,*

$$P(X_{T+i} = x_{m+i} \ \forall\, 1 \le i \le n \parallel T = m \text{ and } X_i = x_i \ \forall\, 0 \le i \le m) = \prod_{i=1}^{n} p(x_{m+i-1}, x_{m+i}). \tag{15}$$

Proof. The event $\{X_i = x_i \ \forall\, i \le m\}$ determines whether the event $T = m$ occurs or not. If not, the event $\{T = m\} \cap \{X_i = x_i \ \forall\, i \le m\}$ has probability 0, and therefore is impossible. Otherwise (and this is the whole point of Definition 6), the condition $T = m$ in the conditional probability (15) is redundant, as

$$\{T = m\} \cap \{X_i = x_i \ \forall\, i \le m\} = \{X_i = x_i \ \forall\, i \le m\}.$$

Therefore, the assertion (15) follows from the ordinary Markov property (Proposition 1). □

4. Recurrence and Transience

Definition 7. Let $\{X_n\}_{n\ge 0}$ be a Markov chain on a finite or countable state space $\mathscr{X}$, and for any state x let $T_x = T_x^1$ be the first passage time to x. A state x is

(a) *recurrent* if $P^x\{T_x < \infty\} = 1$;
(b) *transient* if $P^x\{T_x < \infty\} < 1$;
(c) *positive recurrent* if $E^x T_x < \infty$; and
(d) *null recurrent* if it is recurrent but $E^x T_x = \infty$.

Before looking at some examples, let's do some preliminary reasoning that will lead to alternative conditions for transience and recurrence that are often easier to check than the conditions in the definition. First, suppose that state x is recurrent; by definition, if the chain starts at $X_0 = x$ then it is certain to return. But according to the Strong Markov Property, the chain regenerates at the time T_x^1 of first return, that is, the future behaves like a brand new version of the Markov chain started at state x. Thus, it is certain that the state x will be revisited a second time, and similarly, by induction, x will be revisited at least k times, for any k. So: If a state x is recurrent, then it will be visited infinitely often.

Now suppose that x is transient. It is no longer certain that x will be revisited, but on the event that it is, the chain will regenerate at time T_x^1, by the Strong Markov Property. Therefore, for every $k = 1, 2, \ldots$,

$$P^x\{T_x^k < \infty\} = P^x\{T_x < \infty\}^k. \tag{16}$$

Proof of (16). A formal proof goes by induction on k. The case $k = 1$ is obvious, so we need only do the inductive step. Suppose, then, that the formula holds for all positive integers up to k; we'll show that it then holds also for $k+1$. By the Strong Markov Property, for any integers $m, n \ge 1$,

$$P^x(T_x^{k+1} = m + n \parallel T_x^k = m) = P^x\{T_x = n\}. \tag{17}$$

This follows from (15) by summing over all paths $x_{m+1}, \ldots x_{m+n}$ such that $x_{m+i} \ne x$ for $i < n$, but $x_{m+n} = x$. Summing (17) over $n \ge 1$ gives

$$P^x(T_x^{k+1} < \infty \parallel T_x^k = m) = P^x(T_x < \infty).$$

Now sum on m:

$$\begin{aligned} P^x\{T_x^{k+1}<\infty\} &= \sum_{m=1}^{\infty} P^x(T_x^{k+1}<\infty \parallel T_x^k = m)P^x\{T_x^k = m\} \\ &= P^x(T_x<\infty)\sum_{m=1}^{\infty} P^x\{T_x^k = m\} \\ &= P^x(T_x<\infty)P^x\{T_x^k<\infty\} \\ &= P^x(T_x<\infty)^{k+1}, \end{aligned}$$

the last by the induction hypothesis. □

Note: This argument is typical of how the Strong Markov Property is used in doing formal proofs in Markov chain theory. Since such arguments are tedious, and (should be) fairly obvious once you have seen one example, I will omit them from here on.

Corollary 11. *State x is recurrent if and only if the expected number of visits to x is infinite, that is,*

$$E^x N_x = \sum_{n=0}^{\infty} p_n(x,x) = \infty, \quad \textit{where} \tag{18}$$

$$N_x = \sum_{n=0}^{\infty} 1\{X_n = x\}.$$

Proof. Since expectations and sums can always be interchanged (even when the sum has infinitely many terms, provided they are all nonnegative),

$$E^x N_x = E^x \sum_{n=0}^{\infty} 1\{X_n = x\} = \sum_{n=0}^{\infty} P^x\{X_n = x\} = \sum_{n=0}^{\infty} p_n(x,x).$$

But N_x has another representation: it is one plus the number of indices k such that $T_k<\infty$, since each such index counts one visit to x. Hence,

$$\begin{aligned} E^x N_x &= 1 + E^x \sum_{k=1}^{\infty} 1\{T_x^k<\infty\} \\ &= 1 + \sum_{k=1}^{\infty} P^x\{T_x^k<\infty\} \\ &= 1 + \sum_{k=1}^{\infty} P^x\{T_x<\infty\}^k \\ &= 1/(1-P^x\{T_x<\infty\}) = 1/P^x\{T_x=\infty\}. \end{aligned}$$

By definition, x is recurrent if $P^x\{T_x<\infty\}=1$. Our calculation of $E^x N_x$ shows that this will be the case precisely if $E^x N_x = \infty$. □

Corollary 12. *Recurrence and transience are class properties: If x is recurrent and x communicates with y then y is also recurrent.*

Note: Positive and null recurrence are also class properties, as will be shown later. Corollary 12 implies that in an irreducible Markov chain, all states have the same type (recurrent or transient).

We call an irreducible Markov chain *recurrent* or *transient* according as its states are recurrent or transient (and similarly for positive and null recurrence).

Proof. Suppose that x is recurrent, and that y communicates with x. Then y is accessible from x, and x is accessible from y, so there exist integers $k, l \geq 1$ such that $p_k(x,y) > 0$ and $p_l(y,x) > 0$. By Chapman-Kolmogorov,

$$p_{k+n+l}(y,y) \geq p_l(y,x)p_n(x,x)p_k(x,y),$$

so by the recurrence of x and Corollary 11,

$$\sum_{n=0}^{\infty} p_n(y,y) \geq p_l(y,x)p_k(x,y)\sum_{n=0}^{\infty} p_n(x,x) = \infty.$$

It therefore follows from Corollary 11 that y is recurrent. □

Polya's Theorem . *Simple random walk in dimensions $d = 1,2$ is recurrent, and in dimensions $d \geq 3$ is transient.*

Proof. (Sketch) We'll use Corollary 11. This requires approximations for (or bounds on) the return probabilities $P_n(x,x) = P_n(0,0)$. Simple random walk has period 2, so $P_{2n+1}(0,0) = 0$ for all n. Thus, we need only worry about the return probabilities for *even* times $P_{2n}(0,0)$. This probability is the probability that the sum of the $2n$ increments ξ_i is the 0 vector. The increments are i.i.d. random vectors, with mean zero and a covariance matrix that I could calculate if I were any good at that sort of thing. But for the purposes of this calculation, we don't even need to know the value — we only need to know that it is finite, because then the *Local Central Limit Theorem* (which you can look up in your 304 notes, or in Greg Lawler's Random Walk book) implies that

$$P_{2n}(0,0) \sim C/n^{d/2}$$

for some positive constant C that can be calculated from the covariance matrix. The sequence $1/n^{d/2}$ is summable if $d > 2$, but is not summable in $d = 1,2$, and so the theorem follows. □

5. The Excursion Chain

Assume in the remaining sections 5–7 that X_n is an irreducible Markov chain on a finite or countable state space $\mathscr{X}$ with transition probability matrix $\mathbb{P}$. Assume that there is a recurrent state x. (Recall that if there is a recurrent state, then *all* states are recurrent, by Corollary 12.) Suppose that the Markov chain X_n is started in state $X_0 = x$. Keep a random list of states visited, using the following rule: Start the list with just one item x; for each $n = 1,2,\ldots$, add the state X_n to the end of the list if $X_n \neq y$, but if $X_n = x$, erase everything on the list except the item x. The sequence of random lists produced by this algorithm is called the *excursion chain.*

Example: If the sequence of states visited by the Markov chain X_n is $x, y_1, y_2, x, y_3, y_4, \ldots$ then the succcessive states of the excursion chain are

$$x,\ xy_1,\ xy_1y_2,\ x,\ xy_3,\ xy_3y_4,\ \ldots.$$

In general, the lists that can occur as states of the excursion chain are the finite words $xy_1y_2\cdots y_k$ of length ≥ 1 such that (a) the letter x occurs only once in the word, at the beginning; and (b) for every pair y_jy_{j+1} (or xy_1) of adjacent letters, the transition probability $p(y_j, y_{j+1}) > 0$, that is, $y_j \to y_{j+1}$ is an allowable jump of the Markov chain X_n. Denote the set of all such words by $\mathscr{Y}$.

Definition 8. The *excursion chain* (or more properly, the $x-$*excursion chain*) is the Markov chain on the state space $\mathscr{Y}$ with transition probabilities

$$\begin{aligned} q(xy_1y_2\cdots y_k, xy_1y_2\cdots y_ky_{k+1}) &= p(y_k, y_{k+1}) \quad \text{if } y_{k+1}\neq x;\\ q(xy_1y_2\cdots y_k, x) &= p(y_k, x);\\ q(x, xy) &= p(x, y) \quad \text{if } y\neq x;\\ q(w, w')) &= 0 \quad \text{otherwise.} \end{aligned}$$

Let $F:\mathscr{Y}\to\mathscr{X}$ be the projection on the last letter, that is, the mapping that assigns to each word $xy_1\cdots y_k$ its last letter y_k.

Lemma 13. *Let Y_n be a version of the excursion chain, that is, a Markov chain on the state space $\mathscr{Y}$ with transition probability matrix $\mathbb{Q}=(q(u,v))_{u,v\in\mathscr{Y}}$. Then $F(Y_n)$ is a version of the original Markov chain X_n, equivalently, $F(Y_n)$ is a Markov chain on $\mathscr{X}$ with transition probability matrix $\mathbb{P}$.*

Proof. Routine exercise. □

When does the excursion chain have a stationary distribution? Suppose that it does: call it ν. By definition of a stationary distribution, the distribution ν must satisfy the system of equations $\nu^T=\nu^T\mathbb{Q}$. Now if $w\in\mathscr{Y}$ is a word of length 2 or more, then there is only one word w' such that $q(w',w)>0$, namely, the word w' gotten by deleting the last letter of w. Hence, the steady state equation for $\nu(w)$ reads

$$\nu(w)=\nu(w')q(w',w).$$

Applying the same reasoning to $\nu(w')$ and iterating, we find that

$$\nu(xy_1\cdots y_k)=\nu(x)p(x,y_1)\prod_{i=1}^{k-1}p(y_i,y_{i+1}). \tag{19}$$

This shows that there can be at most one stationary distribution for the excursion chain, and that a stationary distribution exists if and only if there is a finite, positive value of $\nu(x)$ such that

$$\sum_{k=0}^{\infty}\sum_{y_1y_2\cdots y_k}\nu(x)p(x,y_1)\prod_{i=1}^{k-1}p(y_i,y_{i+1})=1. \tag{20}$$

Proposition 14. *The excursion chain has a stationary probability distribution ν if and only x is a positive recurrent state of the Markov chain X_n, that is, $E^xT_x<\infty$. In this case, the stationary distribution is given by* (19)*, with*

$$\nu(x)=1/E^xT_x. \tag{21}$$

Proof. Consider the kth term of the outer sum in (20): This is a sum over all paths $y_1y_2\cdots y_k$ of length k that do not contain the state k. The union of all such paths is the event that the Markov chain X_n will not revisit the state x in its first k steps. Thus, for each $k\geq 0$,

$$\sum_{y_1y_2\cdots y_k}p(x,y_1)\prod_{i=1}^{k-1}p(y_i,y_{i+1})=P^x\{T_x>k\}.$$

Hence, the equation (20) reduces to

$$\nu(x)\sum_{k=0}^{\infty}P^x\{T_x>k\}=1.$$

The result (21) now follows, because the expectation of any nonnegative integer-valued random variable N is given by $EN = \sum_{k\geq 0} P\{N > k\}$. □

Corollary 15. *If an irreducible Markov chain has a positive recurrent state x, then it has a stationary distribution π for which*

$$\pi(x) = 1/E^x T_x \tag{22}$$

Proof. We have just seen that the existence of a positive recurrent state x implies that the $x-$excursion chain has a unique stationary distribution ν. We have also seen that the excursion chain Y_n projects (via the mapping F onto the last letter) to a version of the original Markov chain X_n. It follows that the stationary distribution of the chain Y_n projects to a stationary distribution for X_n:

$$\pi(z) = \sum_{y:F(y)=z} \nu(y).$$

Exercise: Verify that stationary distributions project to stationary distributions. □

Later we will show that an irreducible Markov chain can have at most one stationary distribution π, and also that if there is a positive recurrent state then *all* states are positive recurrent. It will then follow that the formula (22) must hold for all states x.

6. Excursions and the SLLN

The excursion chain introduced in the preceding section grows random words one letter at a time. In this section, we will look at complete excursions, that is, the segments of the Markov chain between successive visits to a distinguished state x. Once again, assume that X_n is an irreducible, recurrent Markov chain on a finite or countable state space $\mathcal{X}$ with transition probability matrix $\mathbb{P}$. Fix a state x, and for typographical ease, set

$$\tau(k) = T_x^k \quad \text{for } k = 1, 2, \ldots .$$

Thus, the times $\tau(k)$ mark the successive visits to state x. For convenience, set $\tau(0) = 0$. The *excursions* from state x are the random sequences (words)

$$\begin{aligned} W_1 &:= (X_0, X_1, X_2, \ldots, X_{\tau(1)-1}), \\ W_2 &:= (X_{\tau(1)}, X_{\tau(1)+1}, X_{\tau(1)+2}, \ldots, X_{\tau(2)-1}), \\ &\text{etc.} \end{aligned} \tag{23}$$

Since the Markov chain is recurrent, the stopping times $\tau(k)$ are all finite, so the excursions all terminate, that is, the excursions are finite words with letters in the alphabet $\mathcal{X}$.

Lemma 16. *Under P^x, the excursions $W_1, W_2, \ldots$ are independent and identically distributed. Under P^y (where $y \neq x$), the excursions $W_1, W_2, \ldots$ are independent, and $W_2, W_3, \ldots$ are identically distributed.*

Proof. Consider any finite sequence $w_1, w_2, \ldots, w_k$ of possible excursions, with word representations

$$w_j = (x_{j,1}, x_{j,2}, \ldots, x_{j,m(j)}).$$

In order that these words be allowable as excursions, they may only include the letter x once, at the very beginning. By the Markov property,

$$P^x\{W_j = w_j\ \forall\, j = 1,2,\ldots,k\} = \prod_{j=1}^{k}\left(\prod_{l=1}^{m(j)} p(x_{j,l}, x_{j,l+1})\right) p(x_{j,m(j)-1}, x).$$

(Note that the final factor $p(x_{j,m(j)-1}, x)$ in the inner product occurs because, in order that w_j be the jth excursion, the Markov chain must jump back to the state x at the conclusion of the excursion.) Since this is a product of factors identical in form, it follows that the excursions $W_1, W_2, \ldots$ are i.i.d. A similar calculation applies under the probability measure P^y; the only difference is that the very first excursion must start with the letter y, rather than x, so its distribution differs from the rest. □

This lemma is perhaps the most useful technical tool (along with the Strong Markov Property) in the analysis of discrete Markov chains, because it provides a means for reducing problems about the long-run behavior of the Markov chain to problems about sequences of i.i.d. random variables and vectors.

Corollary 17. *If there is a positive recurrent state x, then all states are positive recurrent.*

Proof. Exercise. Hint: First show that if x is positive recurrent then

$$E^x(T_x \parallel T_y < T_x) < \infty \quad \text{and} \quad E^x(T_x \parallel T_y > T_x) < \infty$$

for all states $y \neq x$. Then show that a $y-$excursion is contained in the conjunction of (i) an $x-$ excursion conditioned to have a visit to y, followed by (ii) a geometric number of $x-$ excursions conditioned *not* to visit to y, followed by (iii) an $x-$ excursion conditioned to have a visit to y. Alternatively, fashion an argument based on the SLLN for excursions formulated in Corollary 18 below. □

Definition 9. Let $f : \mathscr{X} \to \mathbb{R}$ be a real-valued function on the state space $\mathscr{X}$. The *additive extension* of f to the set of finite words with letters in $\mathscr{X}$ is the function f_+ that assigns to a finite word $w = (x_1, x_2, \ldots, x_m)$ the value

$$f_+(w) := \sum_{i=1}^{m} f(x_i). \tag{24}$$

Corollary 18. *Let $f : \mathscr{X} \to \mathbb{R}$ be a nonnegative (or bounded) function, and let f_+ be its additive extension. For any initial state $y \in \mathscr{X}$, with P^y-probability one,*

$$\lim_{k\to\infty} k^{-1} \sum_{i=1}^{k} f_+(W_i) = E^x f_+(W_1) = E^x \sum_{j=0}^{\tau(1)-1} f(X_j). \tag{25}$$

In particular (use $f \equiv 1$), with P^y-probability one,

$$\lim_{k\to\infty} \tau(k)/k = E^x \tau(1). \tag{26}$$

Proof. For $y = x$, this follows directly from the usual SLLN (Strong Law of Large Numbers) for sums of i.i.d. nonnegative random variables, because by Lemma 16, under P^x the random variables $f_+(W_1), f_+(W_2), \ldots$ are independent and identically distributed. On the other hand, if $y \neq x$ then under P^y the distribution of the *first* excursion W_1 may be different from that of the subsequent excursions $W_2, W_3, \ldots$; however, it is still the case that $W_2, W_3, \ldots$ are i.i.d. and have the same

distribution (under P^y) as does W_1 under P^x (because all of the excursions after the first start at x). Hence, even though the distribution of $f_+(W_1)$ may be different, this won't affect the limiting behavior in (25), because for large k the factor k^{-1} will dampen out the effect of $f_+(W_1)$. □

Let's consider the implications of (26). This states that for large k, the time of the kth visit to x will (with probability approaching one as $k \to \infty$) be about $kE^x\tau(1)+o(k)$. But this means that for large n, the number $N_n = N_n^x$ of visits to x by time n will be about $n/E^x\tau(1)$. Recall (Corollary 15) that if the Markov chain is positive recurrent, then there exists a stationary distribution π for which $\pi(x) = 1/E^x\tau(1)$. (We don't yet know that the stationary distribution is unique, but we will prove this shortly.) Therefore, (26) implies that if the Markov chain is positive recurrent, then the limiting fraction of time that the Markov chain spends in state x is $\pi(x)$, and this holds regardless of the initial state y. On the other hand, if the Markov chain is *null* recurrent, then $E^x\tau(1) = \infty$, and so (26) implies that the limiting fraction of time spent in state x is 0. (This is why null recurrent chains are called *null* recurrent.)

Theorem 19. *Fix $x \in \mathcal{X}$, and let $N_n = N_n^x$ be the number of visits to state x up to time n. If the Markov chain is* positive recurrent *and* irreducible*, then there is a unique stationary probability distribution π, and for all states x, y, with P^y-probability* 1*,*

$$\boxed{\lim_{n\to\infty} \frac{N_n^x}{n} = \pi(x)} \tag{27}$$

If, on the other hand, the Markov chain is null *recurrent, then there is no stationary probability distribution, and for all states x, y, with P^y-probability* 1*,*

$$\boxed{\lim_{n\to\infty} \frac{N_n^x}{n} = 0} \tag{28}$$

In either case, the Markov chain visits every state infinitely often.

Proof. The argument outlined in the paragraph preceding the statement of the theorem shows that in both the positive and null recurrent cases,

$$\lim_{n\to\infty} N_n^x/n = 1/E^x T_x \tag{29}$$

with P^y-probability one, for any $y \in \mathcal{X}$. In the null recurrent case, $E^x T_x = \infty$ for every state x (Corollary 17), and so (28) follows. Assume now that the Markov chain has a stationary distribution π (recall that in the positive recurrent case there is always at least one stationary distribution, by Corollary 15). Since $P^\pi = \sum_y \pi(y)P^y$ is a weighted average of the probability measures P^y, the convergence (29) holds with $P^\pi-$probability 1. Since the ratios N_n^x/n are bounded between 0 and 1, the Bounded Convergence Theorem implies that

$$\lim_{n\to\infty} E^\pi N_n^x/n = 1/E^x T_x.$$

But

$$\begin{aligned} E^\pi N_n^x/n &= E^\pi n^{-1} \sum_{j=1}^n 1\{X_j = x\} \\ &= n^{-1} \sum_{j=1}^n P^\pi\{X_j = x\} = \pi(x), \end{aligned}$$

since π is a stationary distribution. Therefore,

$$\pi(x) = 1/E^x T_x \tag{30}$$

for all states x. It follows that there is no stationary distribution in the null recurrent case (because (30) would force it to be identically zero) and in the positive recurrent case there can only be one stationary distribution.

It remains to prove the final assertion of the theorem, that every state is visited infinitely often, with probability one. In the positive recurrent case, this follows directly from (27), because $\pi(x) > 0$. (Recall [Fact 3] that every state in an irreducible, positive recurrent Markov chain must have positive stationary probability.) Now consider the null recurrent case: Fix states x, y, z, and consider the event that an $x-$excursion W_i includes a visit to z. This event has positive P^x-probability, because by irreducibility there exist positive probability paths from x to z and from z back to x, which may be pieced together to give an $x-$excursion with a visit to z. Therefore, the SLLN (25) implies that, under P^y, the limiting proportion of the excursions W_i that visit z is positive. □

7. Coupling and Kolmogorov's Limit Theorem

Theorem 20. *Assume that X_n is an* aperiodic, positive recurrent, irreducible *Markov chain on $\mathscr{X}$, and let π be the unique stationary distribution. Then for all states x, y,*

$$\lim_{n\to\infty} P^x\{X_n = y\} = \pi(y). \tag{31}$$

This is the fundamental limit theorem of discrete Markov chain theory. Several proofs are now known, including one (the one most commonly given in textbooks) based on another deep theorem, the *Feller-Erdös-Pollard* theorem of renewal theory.[1] The proof to be given here avoids the *FEP* Theorem; instead, it relies on a useful technique known as *coupling*, first invented by Doeblin several years after Kolmogorov published his work on countable state Markov chain.

Proof of Theorem 20. The strategy of the coupling argument is this: Suppose that we could construct two versions X_n and X_n^* of the Markov chain simultaneously (on the same probability space $(\Omega, \mathscr{F}, P)$) in such a way that

$$X_0 = x; \tag{32}$$

$$X_0^* \sim \pi; \quad \text{and} \tag{33}$$

$$X_n = X_n^* \quad \text{eventually with probability 1.} \tag{34}$$

(The third condition is the reason for the term "coupling".) Since X_n^* starts in the stationary distribution π, at any subsequent time the distribution of X_n^* will still be π. On the other hand, for large n the chains X_n and X_n^* will be in the same state, by the third requirement, so

$$|P\{X_n = y\} - \pi(y)| = |P\{X_n = y\} - P\{X_n^* = y\}| \le P\{X_n \neq X_n^*\} \to 0$$

as $n \to \infty$. Kolmogorov's theorem then follows.

There are a number of ways to construct the coupling X_n, X_n^*. The one followed here is completely elementary, relying only what we already know about Markov chain theory. The idea is to run the chains X_n and X_n^* independently, starting from initial states $X_0 = x$ and $X_0^* \sim \pi$, until the

[1] This is what is done in Ross. Unfortunately, Ross does not prove the Feller-Erdös-Pollard theorem, so nothing is really proved.

first time τ that they meet (i.e., τ is the first n such that $X_n = X_n^*$). Then, after time τ, we force the two chains to follow the same path, so that $X_n = X_n^*$ for all $n \geq \tau$.

Following is a more precise description of the construction: Let X_n and X_n' be independent versions of the Markov chain, with initial states $X_0 = x$ and $X_0' \sim \pi$. (Observe that independent realizations of a Markov chain can always be constructed – for instance, just use two independent random number generators in conjunction with the transition probabilities to determine the jumps.) Define

$$\tau := \min\{n \geq 0 : X_n = X_n'\}; \tag{35}$$

we will prove below that $\tau < \infty$ with probability one. Finally, define

$$\begin{aligned} X_n^* &= X_n' \quad \text{for } n \leq \tau, \quad \text{and} \\ X_n^* &= X_n \quad \text{for } n \geq \tau. \end{aligned} \tag{36}$$

This definition is valid, because $X_\tau = X_\tau'$.

To prove that $\tau < \infty$ with probability one, and that the process X_n^* just constructed is actually a version of the Markov chain, we shall look more closely at the sequence $V_n := (X_n, X_n')$ that tracks the states of both X and X' together. Since the sequences X_n and X_n' are independent, by hypothesis, the vector process V_n is itself a Markov chain on the state space $\mathscr{X} \times \mathscr{X}$, with transition probabilities

$$q((x,x'),(y,y')) = p(x,y)p(x',y'). \tag{37}$$

(This is easily checked, using the fact that each of the processes X_n and X_n') has the Markov property separately, together with the mutual independence.)

Lemma 21. *The Markov chain V_n is irreducible and positive recurrent, with stationary distribution ν given by*

$$\nu((x,x')) = \pi(x)\pi(x') \tag{38}$$

Proof. It is routine to check that the probability distribution ν is stationary (exercise), and so it follows by Theorem 19 that V_n is positive recurrent. The tricky thing here is to show that V_n is irreducible. This is where we will use (finally!) the assumption that the original Markov chain X_n is aperiodic.

By hypothesis, the chain X_n is aperiodic and irreducible. Fix $x \in \mathscr{X}$, and consider the set $A_x = \{n \geq 1 : p_n(x,x) > 0\}$. By the Chapman-Kolmogorov equations, the set A_x is closed under addition (see the proof of Lemma 7). Furthermore, by irreducibility, the greatest common divisor of A_x is 1. Consequently, by Corollary 9, all but at most finitely many of the natural numbers are included in A_x. Thus, there is an integer n_x such that

$$p_n(x,x) > 0 \quad \forall\, n \geq n_x$$

Now let $x, y \in \mathscr{X}$ be any two states. Since the Markov chain X_n is irreducible, there is a positive-probability path from x to y, of length (say) $k_{x,y}$. Hence, by Chapman-Kolmogorov and the result of the preceding paragraph,

$$p_n(x,y) > 0 \quad \forall\, n \geq n_x + k_{x,y}.$$

Finally, consider any four states x, x', y, y' (not necessarily distinct). For all $n \geq \max(n_x + k_{x,y}, n_{x'} + k_{x',y'})$,

$$q((x,x'),(y,y')) = p(x,y)p(x',y') > 0.$$

This proves that the vector chain V_n is irreducible, and also aperiodic. □

Lemma 21 implies that the Markov chain V_n is irreducible and recurrent. Therefore, by Theorem 19, it must visit every state in $\mathscr{X} \times \mathscr{X}$ infinitely often, and in particular, it must visit the state (x,x) at least once. Thus, $\tau < \infty$ with probability one. Clearly, τ is a stopping time for the Markov chain V_n, and so the Strong Markov Property holds: Conditional on $\tau = m$ and on the history of V_n up to time m, the future depends only on the state $V_m = (X_m, X_m) = (z,z)$, and has the same law as a pair of independent versions of X_n both started at z. It follows (exercise) that the spliced process X_n^* defined by (36) is a version of X_n'.

□