INTRODUCTION TO MARKOV CHAIN MONTE CARLO

1. Introduction: MCMC

In its simplest incarnation, the Monte Carlo method is nothing more than a computer-based exploitation of the Law of Large Numbers to estimate a certain probability or expectation. It works like this: Suppose you wish to estimate EY, and that you have an algorithm that will produce a stream of i.i.d. random variables $Y_1, Y_2, \dots$, each with the same distribution as Y. Then the sample average

$$\frac{1}{n}\sum_{j=1}^{n} Y_j \tag{1}$$

will, at least for large sample sizes n, be a good estimator of the expectation EY. In simple problems it is often easy to fashion simple and computationally efficient algorithms for producing such streams of random variables from an input stream of i.i.d. uniform-$[0,1]$ random variables; and most mathematical and statistical software libraries are equipped with *pseudo-random* number generators that will provide a stream of uniforms. Often, however, in more complex problems it is not so easy to produce i.i.d. copies of Y, and so the naive approach fails.

Markov chain Monte Carlo is a more sophisticated technique based on a Law of Large Numbers for Markov chains (which we will state and prove below). It works like this: Suppose, once again, that you wish to approximate an expectation EY. Suppose that you have an algorithm that will generate successive states $X_1, X_2, \dots$ of a Markov chain on a state space $\mathcal{X}$ with stationary distribution π, and that there is a real-valued function $f : \mathcal{X} \to \mathbb{R}$ such that

$$\sum_{x\in\mathcal{X}} f(x)\pi(x) = EY. \tag{2}$$

Then the sample averages

$$\frac{1}{n}\sum_{j=1}^{n} f(X_j) \tag{3}$$

may be used as estimators of EY.

Markov chain Monte Carlo is useful because it is often much easier to construct a Markov chain with a specified stationary distribution than it is to work directly with the distribution itself. While this sounds paradoxical, it isn't, as the example in section 2 below will show. However, the use of estimators (1) based on Markov chains introduces certain complications that do not arise for sample averages of i.i.d. random variables: In the i.i.d. case, the variance of the estimator (1) is just the product of the sample size and the variance of Y. If, on the other hand, the random variables are functions of the successive states of a Markov chain, as in (3), then there is no longer any such simple rule governing the variances of the sample averages; in fact, the variability in the distribution of the sample average (1) depends in a rather complicated way on the rate of convergence to equilibrium of the underlying Markov chain.

2. Examples

2.1. **Metropolis-Hastings Algorithm.** It is often the case in simulation problems that the distribution π one wants to simulate has a normalizing constant that is unknown and prohibitively difficult to calculate: thus, one can write $\pi(x) = Cw(x)$ where w is computable but C is not. Fortunately, it is often possible to build Markov chains with stationary distribution π without knowing C. If, for example, one can find transition probabilities $p(x, y)$ such that the detailed balance equations

$$w(x)p(x, y) = w(y)p(y, x) \tag{4}$$

hold, then they must also hold when w is replaced by π, and so π will be stationary.

There is an easy and flexible way to find solutions to the detailed balance equations (4) due to Metropolis. This is the basis for what is now called the *Metropolis-Hastings* algorithm, perhaps the most common method of simulation now used. The idea is this: Suppose that $\pi(x) = Cw(x)$ is a probability distribution on a finite set $\mathcal{X}$ such that $\pi(x) > 0$ for every x, and suppose that $\mathcal{X}$ is the vertex set of a connected graph $\mathcal{G}$ with edge set $\mathcal{E}$. Choose d large enough that no vertex is an endpoint of more than d edges. For each pair $x, y \in \mathcal{X}$ of distinct vertices, define

$$\begin{aligned} p(x, y) &= \left(\frac{w(y)}{w(x)} \wedge 1\right)/d \quad \text{if } x, y \text{ share an edge;} \\ p(x, y) &= 0 \quad \text{if } x, y \text{ do not share an edge; and} \\ p(x, x) &= 1 - \sum_{y \neq x} p(x, y). \end{aligned} \tag{5}$$

Note that $p(x, x) \geq 0$, because $p(x, y) \leq 1/d$ for each pair x, y of neighboring vertices, and no vertex x has more than d neighbors. Thus, equations (5) defines a system of transition probabilities on $\mathcal{X}$. It is easy to check (Exercise: Do it!) that the detailed balance equations hold for this set of transition probabilities. Since the graph $\mathcal{G}$ is connected, the transition probability matrix is irreducible; but it may be periodic (Exercise: Find an example!). However, by choosing d strictly larger than the degree (:= number of incident edges) of some vertex, one may obtain an aperiodic, irreducible transition probability matrix by (5).

Note: You might wonder why we don't just use the complete graph on the vertex set $\mathcal{X}$. (The complete graph on a set $\mathcal{X}$ is the graph that has an edge between every pair $x, y \in \mathcal{X}$.) The reason is this: In many problems, the weight function $w(x)$ is computable, but not easily so. In such problems it is often advantageous to restrict possible transitions $x \to y$ to those pairs for which the *ratio* $w(y)/w(x)$ is easy to compute.

2.2. **Uniform Distribution on Contingency Tables.** Here is an example which is both important and typical of complex simulation problems. Suppose one has an $n \times m$ *contingency table* (that is, an $n \times m$ array [matrix] $\{x_{ij}\}_{i \in [n]; j \in [m]}$ of nonnegative integers). The index pairs (i, j) will be referred to as *cells.* Define the *marginals* of the table to be the row and column totals; these are denoted by

$$x_{i+} = \sum_{j=1}^{m} x_{ij} \qquad \text{and} \qquad x_{+j} = \sum_{i=1}^{n} x_{ij}. \tag{6}$$

Given a particular set r_i and c_j of row and column totals, denote by $\mathcal{T}$ the set of all possible tables with marginals $\{r_i\}$ and $\{c_j\}$, and let π be the uniform distribution π on $\mathcal{T}$. We wish to simulate the distribution π. (See Diaconis and Efron, *Annals of Statistics*, vol. 13, pp. 845 – 913 for an explanation of why this might be of interest to statisticians.)

Even for tables of moderate size (e.g., $n = m = 10$) it may be quite difficult just to list the tables t with given marginals, much less enumerate them. Hence, for large tables, it is virtually impossible to compute the normalizing constant for the uniform distribution on $\mathcal{T}$. (When the row and column totals are all the same, the allowable tables are known as *magic squares*. In this case it is possible [but not easy] to give an explicit representation of the cardinality of $\mathcal{T}$ — see Spencer, *American Math. Monthly* vol. 87, pp. 397-399.)

We will use the Metropolis-Hastings method to find a reversible Markov chain on the set $\mathcal{T}$ whose stationary distribution is uniform. Observe that, since the target stationary distribution π is uniform, the ratio $w(y)/w(x) = 1$ for all pairs x, y of allowable tables, and so its computation is not an issue here. The difficulty in this problem is in finding a suitable graph structure (that is, building a set of edges connecting allowable tables) without actually having to list the vertices.

Neighboring Tables: Given a table $x = \{x_{ij}\} \in \mathcal{T}$, one may obtain another table $y = \{y_{ij}\}$ as follows: choose two rows I, I' and two columns J, J', and set

$$\begin{aligned} y_{IJ} &= x_{IJ} + 1 & \qquad y_{IJ'} &= x_{IJ'} - 1 \\ y_{I'J} &= x_{I'J} - 1 & \qquad y_{I'J'} &= x_{I'J'} + 1; \end{aligned} \tag{7}$$

set $y_{ij} = x_{ij}$ for all other cells. Observe that if $x_{IJ'} = 0$ or $x_{I'J} = 0$ then the corresponding cells of y will have negative entries, and so $y \notin \mathcal{T}$; but otherwise $y \in \mathcal{T}$, because y has the same marginal totals as x. If table y may be obtained from table x in the manner just described, then tables x and y will be called *neighbors*. Note that if this is the case, then x can be obtained from y in the same manner, by reversing the roles of I and I'. Write $x \sim y$ to denote that tables x, y are neighbors.

The Transition Probability Kernel: A simple Markov chain on the set $\mathcal{T}$ may be run as follows: given that the current state is x, choose a pair (I, I') of rows and a pair (J, J') of columns at random; if the neighboring table y defined by (7) is an element of $\mathcal{T}$ (that is, if it has no negative entries) then move from x to y; otherwise, stay at x. The transition probabilities of this Markov chain may be described as follows:

$$\begin{aligned} p(x,y) &= 0 & \qquad &\text{unless } x = y \text{ or } x \sim y; \\ p(x,y) &= 1/mn(m-1)(n-1) & \qquad &\text{if } x \sim y; \\ p(x,x) &= 1 - \sum_{y \in \mathcal{T}: y \sim x} p(x,y) & \qquad &\text{for all } x \in \mathcal{T}. \end{aligned} \tag{8}$$

Note that the transition rule does not require an explicit description or enumeration of the set $\mathcal{T}$. If these were available, then the uniform distribution on $\mathcal{T}$ would, in effect, also be known, and so the use of MCMC would be unnecessary. Note also that the transition rule is computationally simple: it involves only random selection of four integers in the ranges $[m]$ and $[n]$, and access and modification of the entries in only four cells of the table.

Proposition 1. *The transition probability matrix $\mathbb{P}$ is aperiodic and irreducible on $\mathcal{T}$, and its unique stationary distribution is the uniform distribution on $\mathcal{T}$.*

Proof: The transition probabilities clearly satisfy the detailed balance equations $p(x, y) = p(y, x)$, and so the uniform distribution is stationary. That it is the *unique* stationary distribution will follow from the aperiodicity and irreducibility of the transition probability matrix $\mathbb{P}$.

To prove that the transition probability matrix is aperiodic it suffices to show that there is a state x such that $p(x, x) > 0$. Suppose that $x \in \mathcal{T}$ is a table wth at least one vacant cell, that is, a cell (i, j) such that $x_{ij} = 0$. If the random selection of I, I', J, J' results in $I = i$ and $J' = j$ then the attempted move to the neighboring table y defined by (7) fails, because $y_{ij} = -1$. Thus, $p(x, x) > 0$. Consequently, to show that $\mathbb{P}$ is aperiodic it suffices to show that there is a table $x \in \mathcal{T}$ with a vacant cell.

Is there a table $x \in \mathcal{T}$ with a vacant cell? To see that there is, start with any table x, and set $I = J = 1$ and $I' = J' = 2$. Apply the rule (7) to obtain a neighboring table y. Then apply (7) again to obtain a table y^* neighboring y. Continue this process indefinitely until a table with a vacant cell is reached: this must occur in finitely many steps, because at each step the $(2, 1)$ cell is decreased by one.

Finally, to prove that the transition probability matrix is irreducible, it is enough to show that for any two tables $x, y \in \mathcal{T}$ there is a finite chain of neighboring tables starting at x and ending at y. That this is always possible is left as a HOMEWORK PROBLEM. □

Problem 1. Prove that for any two tables $x, y \in \mathcal{T}$ there is a finite chain of neighboring tables starting at x and ending at y. HINT: Use induction on the number n of rows.

Improved Simulation: The Markov chain described above is easy to run, but because the transitions are restricted to neighboring pairs of tables its rate of convergence to equilibrium may be rather slow. There is a simple modification that will, in some circumstances, greatly speed the Markov chain on its way to equilibrium. The modified chain runs according to the following rules: Given that the current state is $x \in \mathcal{T}$, choose a pair (I, I') of rows and a pair (J, J') of columns at random. Consider the 2×2 (sub)table whose entries are the contents of x in the four cells determined by the choices I, I', J, J': for instance, if $I < I'$ and $J < J'$, the table

$$\begin{matrix} x_{IJ} & x_{IJ'} \\ x_{I'J} & x_{I'J'} \end{matrix}$$

From the set of all 2×2 tables with the same row and column totals, choose one at random (uniformly), and replace the four entries of x in the selected cells by those of the new random 2×2 table. This is computationally simple, because it is easy to enumerate all 2×2 tables with given row and column totals.

3. SLLN for an Ergodic Markov Chain

Theorem 2. *Let $\{X_n\}_{n \geq 0}$ be an aperiodic, irreducible, positive recurrent Markov Chain on a finite or countable state space $\mathcal{X}$ with stationary distribution π, and let $f : \mathcal{X} \to \mathbb{R}$ be a* *real-valued function such that*

$$\sum_{x \in \mathcal{X}} |f(x)| \pi(x) < \infty. \tag{9}$$

Then for every initial state x, with P^x−probability one,

$$\lim_{n \to \infty} n^{-1} \sum_{j=0}^{n} f(X_j) = \mu := \sum_{x \in \mathcal{X}} f(x) \pi(x). \tag{10}$$

This is a special case of another — and more important — generalization of the Law of Large Numbers called the *Ergodic Theorem.* To prove the Ergodic Theorem would take us too far afield, so instead I will deduce Theorem 2 from the SLLN for sums of independent, identically distributed random variables. This requires the notion of an *excursion.* Fix a starting state x, and let $0 = \tau(0) < \tau(1) < \tau(2) < \cdots$ be the times of successive visits to state x. Since the Markov chain is recurrent, these random times are all finite, and since the Markov chain is *positive* recurrent, $E^x \tau(1) < \infty$. The *excursions* from state x are the random finite sequences

$$\begin{aligned} W_1 &:= (X_0, X_1, X_2, \dots, X_{\tau(1)-1}), \\ W_2 &:= (X_{\tau(1)}, X_{\tau(1)+1}, X_{\tau(1)+2}, \dots, X_{\tau(2)-1}), \\ &\text{etc.} \end{aligned} \tag{11}$$

Each excursion is a finite sequence of states beginning with state x.

Lemma 3. *Under P^x, the excursions $W_1, W_2, \dots$ are independent and identically distributed.*

Proof: For any finite sequence $w_1, w_2, \dots, w_k$ of possible excursions, with

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

we have

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

Since this is a product of factors identical in form, it follows that the excursions $W_1, W_2, \dots$ are i.i.d. □

Corollary 4. *For any nonnegative function $f : \mathcal{X} \to [0, \infty)$ and any excursion $w = (x_1, x_2, \dots, x_m)$, define $f(w) = \sum_{i=1}^{m} f(x_i)$. Then with P^x−probability one,*

$$\lim_{k \to \infty} k^{-1} \sum_{i=1}^{k} f(W_i) = E^x f(W_1) = E^x \sum_{j=0}^{\tau(1)-1} f(X_j). \tag{12}$$

Therefore (with $f \equiv 1$),

$$\lim_{k \to \infty} \tau(k)/k = E^x \tau(1). \tag{13}$$

Proof of Theorem 2. It suffices to consider only *nonnegative* functions f, because an arbitrary function f may be decomposed into its positive and negative parts $f = f_+ - f_-$, to each of which the SLLN for nonnegative functions will apply. So assume that $f \geq 0$. For each integer $n \geq 0$, let $N_n = N(n)$ be the number of returns to state x by time n, that is,

$$N_n := \max\{k \,:\, \tau(k) \leq n\}.$$

Since $f \geq 0$,

$$\sum_{j=0}^{\tau(N_n)} f(X_j) \leq \sum_{j=0}^{n} f(X_j) \leq \sum_{j=0}^{\tau(N_n+1)} f(X_j). \tag{14}$$

Now each of the bracketing sums in (14) is a sum over excursions, over the first N_n and first $N_n + 1$ excursions, respectively: In fact, for any k,

$$\sum_{j=0}^{\tau(k)} f(X_j) = \sum_{i=1}^{k} f(W_i)$$

where $f(W_i)$ is defined as in Corollary 4. But Corollary 4 implies that the SLLN applies to each of these sums:

$$\lim_{n\to\infty} N_n^{-1} \sum_{j=0}^{\tau(N_n)} f(X_j) = \lim_{n\to\infty} N_n^{-1} \sum_{j=0}^{\tau(N_n+1)} f(X_j) = E^x \sum_{j=0}^{\tau(1)-1} f(X_j).$$

Corollary 4 also implies that $\tau(k)/k \to E^x\tau(1)$, which in turn implies (why?) that

$$\lim_{n\to\infty} N_n/n = 1/E^x\tau(1).$$

Therefore, it follows that with P^x-probability one,

$$\lim_{n\to\infty} n^{-1} \sum_{j=0}^{n} f(X_j) = \frac{E^x \sum_{j=0}^{\tau(1)-1} f(X_j)}{E^x\tau(1)}.$$

We have now proved that the sample averages of f at the successive states visited by the Markov chain converges to a limit. It remains to prove that this limit is μ. The easiest way to do this is to use the fact that the $n-$step transition probabilities converge to the stationary distribution. This implies, if f is bounded, that,

$$\begin{aligned} \lim_{n\to\infty} n^{-1} E^x \sum_{j=0}^{n} f(X_j) &= \lim_{n\to\infty} n^{-1} \sum_{j=0}^{n} \sum_{y\in\mathcal{X}} p_j(x,y) f(y) \\ &= \sum_{y\in\mathcal{X}} \pi(y) f(y) = \mu. \end{aligned}$$

Now if f is bounded, then the almost sure convergence of the sample averages (10) implies the convergence of their expectations to the same limit. Therefore, if f is bounded then the limit of the sample averages must be μ. Finally, to deduce that this is also the case when f is unbounded (and nonnegative), truncate and use the Monotone Convergence theorem. □

4. Exercises

In the problems below, assume that $\{X_n\}_{n\geq 0}$ is an aperiodic, irreducible, positive recurrent Markov chain on a finite or countable state space $\mathcal{X}$ with stationary distribution π. For any state x, let $T_x = \min\{n \geq 1 : X_n = x\}$.

Problem 2. For distinct states x, y, define $U_{x,y}$ to be the number of visits of the Markov chain to state y before the first return to x, that is,

$$U_{x,y} = \sum_{j=0}^{T_x-1} \delta_y(X_j)$$

where δ_y is the Kronecker delta function. Calculate $E^x U_{x,y}$.

Problem 3. Show that, for any integer $k \geq 1$ and any two states x, y,

$$E^x T_x^k < \infty \quad \text{if and only if} \quad E^y T_y^k < \infty.$$

Problem 4. Let $F : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a nonnegative function of *pairs* of states such that

$$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{X}} F(x,y)\pi(x)p(x,y) := \mu < \infty.$$

Show that with P^x-probability one,

$$\lim_{n\to\infty} n^{-1} \sum_{j=0}^{n} F(X_n, X_{n+1}) = \mu.$$

Hint: Try defining a new Markov chain.

5. Coupling from the Past

One of the more obvious problems facing the user of Markov Chain Monte Carlo methods is in deciding how long the Markov chain must be run. This is because the number of steps required by the Markov chain to "reach equilibrium" is usually difficult to gauge. There is a large and growing literature concerning rates of convergence for finite-state Markov chains, especially for those that arise in typical simulation problems; in addition, statisticians have developed nonrigorous "diagnostics" for use in determining how long an MCMC simulation must be run. Unfortunately, the rigorous bounds are often difficult to use, or badly overconservative; and nonrigorous diagnostics are just that — nonrigorous.

In 1996 J. Propp and D. Wilson published an influential paper describing an algorithm that outputs a random variable whose *exact* distribution is the stationary distribution of a given Markov chain. Thus, in principle, the problem of deciding when to stop a Markov Chain Monte Carlo simulation is now solved. In practice, there are still some difficulties: (1) In many applications, the memory requirements of the Propp-Wilson algorithms are prohibitively large; (2) The running time of the algorithm is random, and often difficult to estimate in advance; and (3) The efficiency of the procedure, as measured by the ratio of output bits to CPU cycles, can be quite small. Nevertheless, the Propp-Wilson algorithm is an important development.

The Propp-Wilson algorithm is based on a device called *coupling from the past.* We have already encountered the notion of *coupling* in connection with the convergence of Markov chains to their equilibria; *coupling from the past* involves a new element, to wit, the extension of a Markov chain backwards in time.

5.1. **Random mappings and Markov chains.** When simulating a Markov chain with given transition probabilities, one would ordinarily keep track only of the current state of the Markov chain. In the Propp-Wilson scheme, *all* possible states of the chain must be tracked. Thus, it is useful to think of the randomness driving the Markov chain trajectories as a set of random instructions, one for each time n and each possible state x, that indicates the state visited next when the state at time n is x. These sets of instructions are mappings (functions) from the state space to itself. (Think of each such random mapping as a set of random arrows pointing from states x to states y, with each state x having exactly one arrow pointing out.)

Let $\mathbb{P} = (p(x,y))_{x,y\in\mathcal{X}}$ be an aperiodic, irreducible transition probability matrix on a finite set $\mathcal{X}$, and let $\pi = (\pi(x))_{x\in\mathcal{X}}$ be its unique stationary distribution. Say that a random mapping $F : \mathcal{X} \to \mathcal{X}$ is *compatible* with the transition probability matrix $\mathbb{P}$ if for any two states $x, y \in \mathcal{X}$,

$$P\{F(x) = y\} = p(x,y). \tag{15}$$

It is always possible to construct $\mathbb{P}-$compatible random mappings: For instance, if $\{U_x\}_{x\in\mathcal{X}}$ are independent, identically distributed random variables uniformly distributed on the unit interval $[0,1]$, and if the state space $\mathcal{X}$ is relabelled so that $\mathcal{X} = [M]$, then the random mapping F defined by

$$F(x) = y \quad \text{if and only if} \quad \sum_{j=1}^{y-1} p(x,j) < U_x \leq \sum_{j=1}^{y} p(x,j) \tag{16}$$

is $\mathbb{P}-$compatible, because for each pair x, y of states the event $F(x) = y$ occurs if and only if the uniform random variable U_x falls in an interval of length $p(x,y)$. Observe that the random mapping so constructed requires, along with the transition probabilities $p(x,y)$, only a stream U_x of i.i.d. uniform-$[0,1]$ random variables; the random number generators in most standard software libraries will provide such a stream.

Proposition 5. *Let $\{F_n\}_{n=1,2,\dots}$ be a sequence of independent random mappings compatible with the transition probability matrix $\mathbb{P}$. For any initial state $x \in \mathcal{X}$, define a sequence of $\mathcal{X}-$vaued random variables inductively as follows:*

$$X_0 = X_0^x = x \quad \textit{and} \quad X_{n+1} = X_{n+1}^x = F_{n+1}(X_n). \tag{17}$$

Then the sequence $\{X_n\}_{n\geq 0}$ constitutes a Markov chain with transition probability matrix $\mathbb{P}$ and initial state x.

Thus, the sequence of random mappings determine trajectories of Markov chains with transition probability matrix $\mathbb{P}$ for *all* initial states.

The proof of Proposition 5 is routine: the Markov property follows from the independence of the random mappings F_n, and the hypothesis that these random mappings are $\mathbb{P}-$compatible assures that the transition probability matrix of the Markov chain is $\mathbb{P}$. Observe that it is not necessary that the random mappings F_n be identically distributed; but in almost any application to a simulation problem, they will be. Finally, note that the cost of storing in computer memory the random mappings $F_1, F_2, \dots, F_n$ is considerably greater than that for storing $X_1, X_2, \dots, X_n$, and that the ratio of these storage costs increases (dramatically!) with the size of the state space $\mathcal{X}$. This is the primary obstacle in applying the Propp-Wilson algorithm.

5.2. **Coalescence in random mappings.** A random mapping compatible with the transition probability matrix $\mathbb{P}$ need not be one-to-one or onto; for instance, the random mapping defined by (16) will not in general be one-to-one. In fact, the success of the Propp-Wilson algorithm requires that the random mappings used should eventually collapse the state space to a single point.

Let $F_1, F_2, \dots$ be a sequence of independent, identically distributed random mappings compatible with the transition probability matrix $\mathbb{P}$. Say that the sequence $F_1, F_2, \dots$ has the *coalescence property* if for some $n \geq 1$ there is positive probability that the functional composition $F_n \circ F_{n-1} \circ \cdots \circ F_1$ maps the state space $\mathcal{X}$ to a single point. Not every $\mathbb{P}-$compatible sequence of i.i.d. random mappings has the coalescence property, even if $\mathbb{P}$ is aperiodic and irreducible; however, compatible sequences with the coalescence property exist if the transition probability matrix $\mathbb{P}$ is aperiodic and irreducible.

Problem 5. Give an example of a $\mathbb{P}-$compatible sequence of i.i.d. random mappings that do not have the coalescence property. HINT: You should be able to do this with a two-state transition probability matrix in which all of the transition probabilities are 1/2.

Problem 6. Show that if $\mathbb{P}$ is aperiodic and irreducible then an i.i.d. sequence of random mappings each distributed as (16) must have the coalescence property.

Proposition 6. *Let $F_1, F_2, \dots$ be a $\mathbb{P}-$compatible sequence of independent, identically distributed random mappings with the coalescence property. Define the* coalescence time T *to be the smallest positive integer n such that $F_n \circ F_{n-1} \circ \cdots \circ F_1$ maps the state space $\mathcal{X}$ to a single point. Then*

$$P\{T < \infty\} = 1. \tag{18}$$

Proof: Since the sequence $F_1, F_2, \dots$ has the coalescence property, there exists a positive integer m such that with positive probability ε, the $m-$fold composition $F_m \circ F_{m-1} \circ \cdots \circ F_1$ maps the state space $\mathcal{X}$ to a single point. For each $k = 0, 1, 2 \dots$, denote by A_k the event that the image of the mapping $F_{km+m} \circ F_{km+m-1} \circ \cdots \circ F_{km+1}$ is a single point. Because the random mappings F_j are i.i.d., the events A_k are independent, and all have probability $\varepsilon > 0$; hence, with probability one, at least one of these events must occur. But on A_k, the composition $F_{mk+m} \circ F_{mk+m-1} \circ \cdots \circ F_1$ maps $\mathcal{X}$ to a single point (EXERCISE: Why?), and so $T < \infty$. □

5.3. **The Propp-Wilson Theorem.** The Propp-Wilson algorithm relies on a device called *coupling from the past*, which entails running the Markov chain backwards in time. This is easily accomplished using random mappings.Let $\{F_n\}_{n \leq 0}$ be a doubly infinite sequence of i.i.d. random mappings compatible with the transition probability matrix $\mathbb{P}$. Assume that these random mappings have the coalescence property, that is, that for some $n \geq 1$ there is positive probability that the functional composition

$$\Phi_n := F_0 \circ F_{-1} \circ F_{-2} \circ \cdots \circ F_{-n} \tag{19}$$

maps the state space $\mathcal{X}$ to a single point. Define the *backward coalescence time* τ to be the smallest integer n such that the image of the random mapping Φ_n is a single point. By the same argument as in the proof of Proposition 6, if the sequence $\{F_n\}_{n \leq 0}$ has the coalescence property, then

$$P\{\tau < \infty\} = 1. \tag{20}$$

The random mapping Φ_τ maps the state space $\mathcal{X}$ to a single point in $\mathcal{X}$. This point is, of course, random, as the random mappings F_n used to produce it were random. Denote it by Z, that is, define

$$\{Z\} = \Phi_\tau(\mathcal{X}). \tag{21}$$

Theorem 7. *(Propp-Wilson) Assume that the transition probability matrix $\mathbb{P}$ is aperiodic and irreducible, and that the random mappings $\{F_n\}_{n\leq 0}$ are i.i.d., compatible with $\mathbb{P}$, and have the coalescence property. Then the distribution of the random variable Z is the unique stationary distribution of the transition probability matrix $\mathbb{P}$.*

Proof: It suffices to show that for any $\varepsilon > 0$ the total variation distance between the distribution of Z and the stationary distribution π is less than ε. For this, it suffices to show that there are $\mathcal{X}-$valued random variables Y_n, each with distribution π, such that

$$\lim_{n\to\infty} P\{Z \neq Y_n\} = 0. \tag{22}$$

Let W be a random variable with distribution π, and for each $n \geq 1$ define

$$Y_n = \Phi_n(W).$$

Observe that if $n \geq \tau$ then $Y_n = Z$. Consequently, because $P\{\tau > n\} \to 0$ as $n \to \infty$, by (20), the relation (22) must be true. Thus, it remains only to show that each Y_n has distribution π. But this follows from Proposition 5 and the definition of a stationary distribution: by Proposition 5, the random variables

$$W, F_{-n}(W), F_{-n+1}(F_{-n}(W)), \dots, \Phi_n(W)$$

are the first $n+1$ states of a Markov chain with transition probability matrix $\mathbb{P}$ and initial state W. Since the distribution of W is, by hypothesis, the stationary distribution π, it follows that the distribution of $\Phi_n(W)$ is also π. □

The significance of Theorem 7 is that it specifies an algorithm for simulating a random variable with distribution π from a sequence of random bits (equivalently, from a sequence of i.i.d. uniform-$[0,1]$ random variables). The random bits are used together with the transition probabilities of the Markov chain to construct random mappings compatible with $\mathbb{P}$ (e.g., using the prescription (16)). These are generated one by one, and at each step the algorithm tests for coalescence of the state space. The random point Z on which coalescence occurs is the output.

In practice, the usefulness of the Propp-Wilson algorithm is somewhat limited, as the algorithm requires repeated testing for coalescence, which can in general be extremely expensive computationally, both in memory use and in CPU cycles.