Math-Stat-491-Fall2014-Notes-II

Hariharan Narayanan

October 28, 2014

1 Introduction

We will be closely following the book
"Essentials of Stochastic Processes", 2nd Edition, by Richard Durrett,
for the topic 'Finite Discrete time Markov Chains' (FDTM). These preliminary notes are for giving an overview, and for providing some guidance on what to aim at while studying this topic.

An FDTM has a finite number of states on which it can rest at discrete times $1, 2, \cdots$. Usually the state at time n is denoted by X_n, and the possible set of states could be denoted by lower case symbols $x, y \cdots$ etc.
If the chain is at state i at time n, it moves to state j at time $n+1$, with probability $p(i,j)$. This could be stated alternatively as 'the conditional probability of the chain being at state j at time $n+1$, given that it is at state i at time n is $p(i,j)$.' The key idea is that the probability of the chain being at state j at time $n+1$, does not depend on what happened before time n. Formally, we have the 'Markov property',

$$Pr\{X_{n+1} = j \mid X_n = i\} = Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\}.$$

We have further assumed 'temporal homogeneity', i.e, that $p(i,j)$ does not depend upon n.
In particular this means

$$Pr\{X_{n+1} = j \mid X_n = i\} = Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_k = i_k\}, \tag{1}$$

because this latter expression, when we shift the time origin to k, is the same as

$$Pr\{X_{n+1-k} = j \mid X_{n-k} = i\} = Pr\{X_{n+1-k} = j \mid X_{n-k} = i, X_{n-k-1} = i'_{n-k-1}, \cdots, X_0 = i'_0\},$$

where $i'_r \equiv i_{r+k}$.

Just to be clear on how we should go about proving formal versions of informal statements consider the following. Suppose what is given is not the immediately previous state but say a state k steps before. Surely, what happened before time $n+1-k$ is unimportant? Let us prove this for $k = 2$. Formally, let us prove

$$Pr\{X_{n+1} = j \mid X_{n-1} = i_{n-1}\} = Pr\{X_{n+1} = j \mid X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\}.$$

We will use the law of total probability. The LHS

$$Pr\{X_{n+1} = j \mid X_{n-1} = i_{n-1}\} = \Sigma_{all\ states\ s} Pr\{X_{n+1} = j \mid X_n = s, X_{n-1} = i_{n-1}\} \times Pr\{X_n = s \mid X_{n-1} = i_{n-1}\}.$$

The RHS
$Pr\{X_{n+1} = j \mid X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\}$

$= \Sigma_{all\ states\ s} Pr\{X_{n+1} = j \mid X_n = s, X_{n-1} = i_{n-1}, \cdots, X_0 = i_0\} \times Pr\{X_n = s \mid X_{n-1} = i_{n-1}, \cdots, X_0 = i_0.\}$

Now we use the basic Markov property on each of the two terms being multiplied and using (1) write the above expression as

$$\Sigma_{all\ states\ s} Pr\{X_{n+1} = j \mid X_n = s, X_{n-1} = i_{n-1}\} \times Pr\{X_n = s \mid X_{n-1} = i_{n-1}\}.$$

Since this is the same as the LHS, the proof is complete.

Next let us prove the following variation of the Markov property

$$Pr\{X_{n+1} = j \mid X_n = i\} = Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_2 = i_2, X_0 = i_0\}.$$

We have skipped $X_1 = i_1$ in the right side conditional probability.
Note that the RHS is equal to
$\Sigma_{all\ states\ s} Pr\{X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \cdots, X_2 = i_2, X_1 = s, X_0 = i_0\}$
$\times Pr\{X_1 = s \mid X_n = i, \cdots, X_2 = i_2, X_0 = i_0.\}$
But this is the same as

$$\Sigma_{all\ states\ s} Pr\{X_{n+1} = j \mid X_n = i\} \times Pr\{X_1 = s \mid X_n = i, \cdots, X_2 = i_2, X_0 = i_0.\}$$

This is equal to

$$Pr\{X_{n+1} = j \mid X_n = i\} \times \Sigma_{all\ states\ s} Pr\{X_1 = s \mid X_n = i, \cdots, X_2 = i_2, X_0 = i_0.\}$$

But

$$\Sigma_{all\ states\ s} Pr\{X_1 = s \mid X_n = i, \cdots, X_2 = i_2, X_0 = i_0.\} = 1.$$

So the RHS=LHS.
The same trick can be used even when there are multiple gaps.

Temporal homogeneity implies we can start from $X_r = i_0$ and repeat the above arguments, replacing X_n by X_{n+r}, etc.

(Natural question at this stage: surely we could also use models where the chain state at $n+1$ depends on states at time $n, n-1, \cdots, n-k$?
Answer: They are not needed! This model is sufficiently versatile.)

Note here that everything we know about an FDTM is captured by the set of $p(i, j)s$. So we could represent the FDTM

- pictorially through a graph with vertices as states with, for each ordered pair (i, j), an edge directed from i to j with weight $p(i, j)$. Of course if $p(i, j) = 0$, we omit the edge from i to j.
- in terms of a 'transition matrix' with rows and columns named by the states with the $(i, j)^{th}$ entry being $p(i, j)$.

Since, from a given state i at time n, the chain *must* move to *some* state (including i) at time $n+1$, it follows that that the outgoing edges (including selfloops,if any) must have weights adding up to 1, and, in the transition matrix, the rows should sum up to 1.

STEP 1 **in the study of FDTM:** Convert word problems in the book to FDTMs specifying the transition matrix and drawing the FDTM graph.

Some fundamental notions associated with an FDTM are listed below.

1. Stationary distribution on the states: this is a probability distribution $\pi(\cdot)$, such that if we pick initial states i with probability $\pi(i)$, the next states j will occur with probability $\pi(j)$.

   Note here that if we pick initial states according to some probability distribution $\pi'(\cdot)$, the probability that the next state is j is given by

   $$\Sigma_i Pr(i) \times Pr(j \mid i) = \Sigma_i \pi'(i) \times p(i,j).$$

   Let $(\pi')^T$ denote the row vector whose i^{th} entry is $\pi'(i)$, and let P denote the transition matrix of FDTM (with $p(i,j)$ as the (i,j) entry).
   So if the initial probability distribution is $\pi'(\cdot)$, the next state distribution is $(\pi')^T P$.
   In the case of the stationary distribution $\pi(\cdot)$, we have $(\pi)^T = (\pi)^T P$.

2. Transient and recurrent states: if you start from a 'recurrent' state, you will return to it, with probability 1. If this probability is < 1, then the state is said to be 'transient'. We can prove that the following 'event' has probability 1 : if we start from a transient state we will return to it only a finite number of times.

3. The expected time of return starting from a given recurrent state. This is self explanatory.

4. $\rho_{xy} \equiv Pr\{we\ can\ go\ from\ x\ to\ y\ in\ finite\ time\}$.

5. Absorbing state: A state is said to be absorbing if, when we start from it, we remain at it with probability 1, i.e., i is an absorbing state iff $p(i,i) = 1$.

6. Irreducible Markov Chain: The graph of the FDTM has the property that given any ordered pair of states (i,j), there is a directed path from i to j in the graph of the FDTM.

STEP 2 **in the study of FDTM:** Internalize the above definitions.

Through the lectures we will attempt to answer the following questions rigorously:

1. How to identify recurrent and transient states by looking only at the graph?

   Answer: Check if we can go from the given state i to some state j through a directed path such that there is no return directed path. If this is true the state is transient, otherwise recurrent.

2. Does the FDTM have a stationary distribution? If it exists, is it unique?

   Answer: FDTMs always have a stationary distribution. It is unique if the FDTM is irreducible.

3. Starting from a given recurrent state i, what is the expected time of return to it?

   Answer: If the FDTM is recurrent and has the stationary distribution $\pi(\cdot)$, the expected time of return to the recurrent state i is $1/\pi(i)$.

4. Starting from a transient state what is the expected time to reach a specified absorbing state?

   Answer: We will describe a method of computing this quantity.

5. Suppose we start from any state and keep running the markov chain according to its transition matrix. Will we encounter the different states with frequency in proportion with their $\pi(\cdot)$ value?

   Answer: Yes, if the FDTM is irreducible.

6. Suppose we start the Markov chain according to a probability distribution $\pi^0(\cdot)$ and let it run forever. Let $\pi^n(\cdot)$ denote the probability distribution after n steps. Will $\pi^n(\cdot)$ converge to the stationary distribution in the limit as $n \to \infty$?

   Answer: Yes if the FDTM is irreducible and aperiodic. (Periodic means all states recur after a certain fixed period > 1.)

To answer the above questions we will need to understand the following:

1. What are the characteristic features of a transition matrix (or equivalently the graph) of an FDTM?

   Answer: The matrix should have nonnegative entries and the rows should add up to 1. The outgoing edges from any node in the graph should have the sum of their weights equal to 1.

   Let us call these respectively Markov transition matrix and Markov chain graph.

2. What kind of a matrix is P^k? What is its significance?

   Answer: P^k is also a Markov transition matrix because its entries are nonnegative and the row sum is 1.

   Its significance is the following:

   Its $(i,j)^{th}$ entry gives the probability of reaching j at the time $n+k$ starting with i at time n, i.e., the probability of reaching j from i in k steps.

   The above requires an understanding of the famous ‘Chapman-Kolmogorov equation’. Below, we give a sketch of the proof.

   Let us denote by $p^s(i,j)$ the probability of reaching state j starting from state i in s steps. ‘Chapman-Kolmogorov equation’ states

   $$p^{m+n}(i,j) = \Sigma_k p^m(i,k) \times p^n(k,j).$$

   The proof is straight forward. To reach from i to j in $m+n$ steps we must reach some intermediate state in m steps. So the probability of reaching from i to j in $m+n$ steps through the intermediate state k is the product $p^m(i,k) \times p^n(k,j)$. Now the events ' reaching k' in m steps and then reaching j in n steps are disjoint for different ks and together make up the event 'reaching from i to j in $m+n$ steps. So

$$p^{m+n}(i,j) = \Sigma_k p^m(i,k) \times p^n(k,j).$$

But this is exactly how the matrix P^{m+n} is computed in terms of the matrices P^m, P^n.
(To formalize the sketch we must use conditional probabilities.
Example: $p^s(i,j) = Pr\{X_s = j \mid X_0 = i\}$.

Probability of reaching from i to j in $m+n$ steps passing through k after m steps
$= Pr\{X_{m+n} = j, X_m = k \mid X_0 = i\}$
$= Pr\{X_{m+n} = j \mid X_m = k, X_0 = i\} \times Pr\{X_m = k \mid X_0 = i\}$
$= Pr\{X_{m+n} = j \mid X_m = k\} \times Pr\{X_m = k \mid X_0 = i\} = p^n(k,j) \times p^m(i,k)$.)

3. A random variable T that takes values $0, 1, 2 \cdots$ is called a *stopping time* or *Markov time* if whether $T = k$ or not depends only on the states $X_0, \cdots X_k$, that the markov chain takes at times $0, 1, \cdots, k$. Are the following stopping times?

   (a) $T = n$, if starting at x at time 0 we reach y at time n.

   (b) $T = n$, if starting at x at time 0 we reach y at time n and $n \leq 100$.

   (c) $T = n$, if we are at y for the last time.

   Answer: The first two random variables are stopping times. The last is not, since, to know that we are at y for the last time is impossible knowing only the past history.

   'Stopping time' is one of the most slippery and useful notions that we encounter while studying FDTMs. For us, its power lies in the fact that we can, because of the 'strong Markov property,' treat X_T as X_0, even though T is a random variable.

STEP 3 in the study of FDTM: Know how to prove the above results rigorously. But do not lose the 'common touch'- understand them in the commonsensical intuitive way too.

Special conditions make the computation of stationary probability easy for some Markov chains. Some of these are discussed below.

1. **Doubly stochastic chains** For these chains the transition matrix has its columns also adding up to 1. Now, if we sum all the row vectors, we get the vector $\mathbf{1}^T = (1, 1, \cdots 1)$, i.e., $\mathbf{1}^T P = \mathbf{1}^T$. So scaling $\mathbf{1}^T$ by the reciprocal of the number of states will satisfy the same equation while being a probability distribution. Thus, for such chains, the uniform probability distribution is a stationary distribution.

2. **Duality** Let Markov chain X_n on states S have transition probability $p(i,j)$ and (unique) stationary probability $\pi(i), i \in S$.

   Define a new Markov chain Y_n on S with transition probability $p'(i,j) \equiv \pi(j) \times p(j,i)/\pi(i)$. Let us call this the *dual* Markov chain to X_n. It can be verified that $\Sigma_i p'(i,j) = 1$.

   It can be seen that, whenever there is an edge from i to j in the original Markov chain with probability $p(i,j)$, in the dual there is an edge from j to i with probability $p'(j,i)$. Let $\pi'(\cdot)$ denote the stationary distribution of the dual Markov chain. We can show that $\pi'(\cdot) = \pi(\cdot)$.

3. **Reversibility** When the Markov chain is its own dual, we have the *detailed balance condition*

$$p(i,j) = \pi(j) \times p(j,i)/\pi(i).$$

   We call such a Markov chain reversible.

   The algorithm given below verifies whether a given chain is reversible even where the distribution $\pi(\cdot)$ is not available, and if reversible, computes $\pi(\cdot)$.

   The algorithm is simple. Let us suppose that from any node in the graph we can go to any other node through a directed path (i.e., always going along the direction of the arrow of the edge). We can show then that all entries of $\pi(\cdot)$ are positive even if the Markov chain is not reversible.

   Firstly if $p(i,j) > 0$ we must have $p(j,i) > 0$ for the detailed balance condition $p(i,j) = \pi(j) \times p(j,i)/\pi(i)$, to hold. So we should check if every edge has, in parallel, an edge going in the opposite direction. If this is not true, we can declare the chain to be not reversible.

   Let us start from some node say 0 and assign it the value $\pi(0) = 1$.
   (We will be scaling the $\pi(.)$ values later appropriately, if our algorithm is able to terminate properly.)
   If node 1 has an edge with value $p(0,1)$ coming into it, for the detailed balance condition to be satisfied, we need $\pi(1) = \pi(0) \times p(0,1)/p(1,0)$. So $\pi(1)$ is fixed.

   We repeat this process:
   Starting from the set of nodes V for which the $\pi(\cdot)$ value is currently fixed, check if any out going edges are there to other nodes. Suppose there is an edge from node $j \in V$ to a node k outside. Fix the value $\pi(k)$ as $\pi(k) = \pi(j) \times p(j,k)/p(k,j)$.

   We will stop when there are no outgoing edges from V. This would also mean that we have assigned a $\pi(\cdot)$ value to each node in the graph. Observe that by this time, within a scaling factor, $\pi(\cdot)$ is unique. The scaling factor arises because we could have assigned any positive value for $\pi(0)$.

   However, we do not know whether for *every* edge (i,j), the detailed balance condition is satisfied. So we check this now. If for some edge there is failure, we declare the Markov chain is not reversible.

   If there is no failure, the chain is reversible and we scale the $\pi(\cdot)$ values so that they add up to 1. This is the desired stationary probability distribution.

   An immediate consequence of the above discussion is that if

   - every edge has a parallel edge in the opposite direction

- there are no directed loops other than the parallel edges

the Markov chain is reversible. This is because by the time all nodes have been assigned $\pi(\cdot)$ values by our algorithm, the detailed balance conditions would have also been verified since there are no other edges.

Example: The Ehrenfest chain graph is a simple straight line, if we replace parallel edges with single edges. So there are no loops except the parallel edges and the chain is reversible.

The algorithm is illustrated with an example below.

A test for Reversibility of Markov Chains

Assume that M is an irreducible aperiodic markov chain. Thus it has a unique stationary distribution, with a positive probability at every vertex.

Step 1 Build the R-graph of the Markov chain.

R-graph: Vertices are states of the Markov Chain.If p_{ij} is not equal to zero, put a directed edge $e(i,j)$ from i to j with weight $w_{ij} = \frac{p_{ij}}{p_{ji}}$ in the case of FDTM . If the weight of any edge is infinite STOP. The MC is not reversible since we have

$$\pi_i p_{ij} \neq \pi_j p_{ji} \tag{2}$$

for any stationary distribution π on the states.

Example:

Let the transition matrix in the case of FDTM be the matrix A below

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & a_{14} \\ a_{21} & a_{22} & a_{23} & 0 \\ 0 & a_{32} & a_{33} & a_{34} \\ a_{41} & 0 & a_{43} & a_{44} \end{bmatrix}$$

The R-graph is shown in Figure 1

Step 2 a Start from any node n. Assign it value 1. Suppose a set N_A of nodes have been assigned values. If N_A is not the full set of nodes pick any edge from $i \in N_A$ to $j \notin N_A$. Assign $\pi_j = w_{ij}\pi_i$. If no edges leave these nodes to the (non null) complement declare the Markov Chain to be not irreducible.

b If N_A is the full set of nodes GO TO STEP 3.

Step 3 For every edge $e(i,j)$ verify that $\pi_i w_{ij} = \pi_j$. If this is violated declare chain is not reversible. If not, Scale π_i by k so that

$$\sum k\pi_i = 1 \tag{3}$$

Output π_i as the stationary distribution and STOP.

Justification : If step 3 is satisfied we have the detailed balance equation satisfied for every pair $\{i,j\}$ for which $p_{ij} \neq 0$ and hence the MC is reversible.

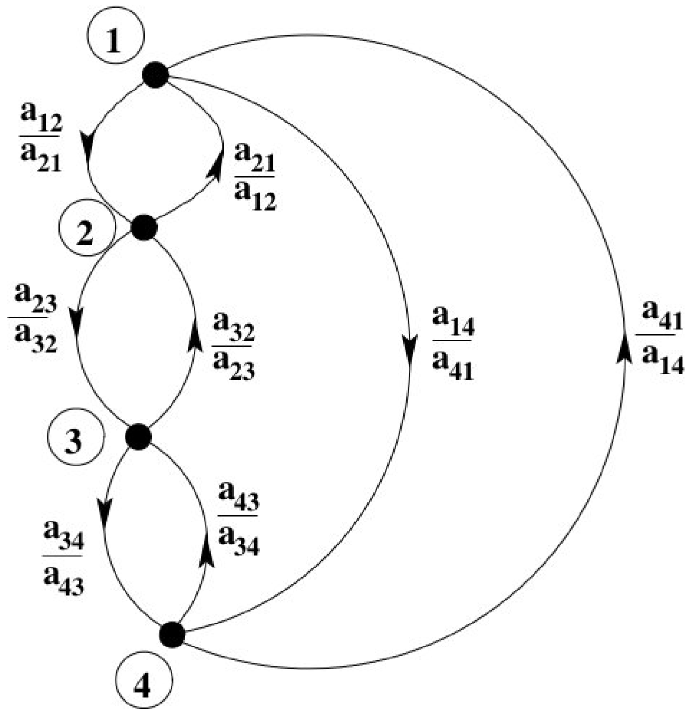

Figure 1:

4. **New Markov Chains from old**

(a) *Merging:* Suppose a subset S_k of k states all have the same probability under the stationary distribution $\pi(\cdot)$. We build a new Markov chain by fusing all these states into a single 'super-state'. All the edges leaving the original states have the probabilities associated multiplied by $1/k$, edges entering have the same probability as before, edges going between nodes in S_k become self loops, with probability multiplied by $1/k$. If parallel edges (of the same direction) result they are replaced by a single edge with the new probability equal to the sum of those of the original parallel edges. If several self loops result at a node, they can be replaced by a single selfloop at the same node with the edge probabilities added. For this Markov chain, we can show that the new stationary distribution $\pi'(\cdot)$ can be obtained by putting $\pi'(S_k) = k \times \pi(i), i \in S_k$, and leaving all other probabilities unchanged.

The proof of these statements is best done by visualizing the above process of merging in terms of the transition matrix.

Let P be the transition matrix. Let S_2 be the set of k states to be merged and S_1 be the complement. The k states in S_2 have the same $\pi(.)$ value. Let $s \in S_2$. For the original Markov chain we have $\pi^T P = \pi^T$. After partitioning the rows and columns according to S_1, S_2,

$$\left(\pi_1^T, \pi_2^T\right) \left(\begin{array}{c|c} P_{11} & P_{12} \\ \hline P_{21} & P_{22} \end{array} \right) = \left(\pi_1^T, \pi_2^T\right). \tag{4}$$

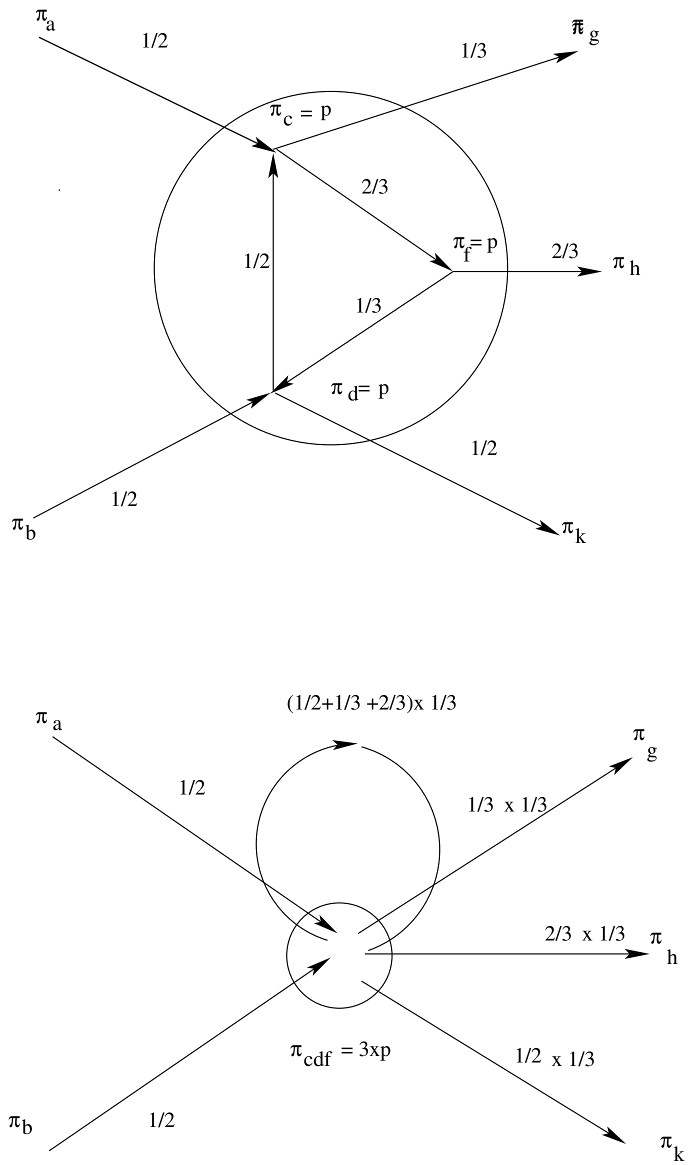

Figure 2: The merging process

Now this is the same as

$$\left(\pi_1^T, \pi(s) \times k\right) \left(\begin{array}{c|c} P_{11} & P_{12} \\ \hline \hat{p}_{21} & \hat{p}_{22} \end{array}\right) = \left(\pi_1^T, \pi_2^T\right), \tag{5}$$

where $(\hat{p}_{21}, \hat{p}_{22})$ is $1/k \times$ *sum of rows of* (P_{21}, P_{22}). Now in the above equation, if we sum the second set of columns of the matrix

$$\left(\begin{array}{c|c} P_{11} & P_{12} \\ \hline \hat{p}_{21} & \hat{p}_{22} \end{array}\right), \tag{6}$$

on the LHS, yielding

$$\left(\begin{array}{c|c} P_{11} & \tilde{P}_{12} \\ \hline \hat{p}_{21} & \tilde{p}_{22} \end{array}\right), \tag{7}$$

the equation would remain correct if we sum the second set of columns of the row vector on the right side too. But summing the second set of columns of $\left(\pi_1^T, {\pi_2}^T\right)$, yields $\left(\pi_1^T, \pi(s) \times k\right)$. We thus have

$$\left(\pi_1^T, \pi(s) \times k\right) \left(\begin{array}{c|c} P_{11} & \tilde{P}_{12} \\ \hline \hat{p}_{21} & \tilde{p}_{22} \end{array} \right) = \left(\pi_1^T, \pi(s) \times k\right). \tag{8}$$

To verify that the matrix in (7), is the transition matrix of the new merged Markov chain, we just note that the edges entering correspond to entries in $\tilde{P}_{12}$, edges leaving correspond to entries in $\hat{p}_{21}$, and self loops correspond to entries in $\tilde{p}_{22}$. It follows therefore that $\pi'(\cdot) = \left(\pi_1^T, \pi(s) \times k\right)$, is the stationary distribution for the merged Markov chain.

(b) *splitting:* This operation could be regarded as one way of reversing the operation of 'merging'. In the old Markov chain graph, we take any node and split it into k nodes. Every incoming edge should be duplicated k times, with the edge probability $1/k$ times the previous one, and feed into each of the split nodes. Every outgoing edge must be duplicated k times and feed out of each of the split nodes, with the same edge probability. Every self loop must be duplicated k times and attached to each of the resulting split nodes with the same probability as before.

Let the old stationary distribution be $\pi(\cdot)$, and the new one be $\pi'(\cdot)$. Let node s in the old Markov chain graph split into the identical k nodes s_i which form set S_k. Then we can show $\pi'(s_i) = 1/k \times \pi(s)$. The probabilities associated with the nodes which have not been split remain as before.

Observe that splitting followed by merging will return to the original Markov chain, but merging followed by splitting might not.

(c) *Metropolis-Hastings method* This is a powerful method of building a reversible Markov chain which has a desired stationary distribution $\pi(\cdot)$, on a given set of states. Let X_n be a Markov chain with edge probability $q(i,j)$, with the additional condition that $q(i,j) \neq 0$ implies $q(j,i) \neq 0$. Let $r(i,j)$ denote $min\ [\pi(j)q(j,i)/\pi(i)q(i,j),\ 1]$.

Let Y_n be the Markov chain on the same graph but with transition probability

$$p(i,j) = q(i,j)r(i,j).$$

Let us verify that Y_n satisfies the detailed balance condition with respect to the distribution $\pi(\cdot)$. Suppose $\pi(j)q(j,i) \leq \pi(i)q(i,j)$. We then have

$$\pi(i)p(i,j) = \pi(i)q(i,j)r(i,j) = \pi(i)q(i,j) \times \pi(j)q(j,i)/\pi(i)q(i,j) = \pi(j)q(j,i).$$

$$\pi(j)p(j,i) = \pi(j)q(j,i)r(j,i) = \pi(j)q(j,i) \times\ 1 = \pi(j)q(j,i),$$

verifying the detailed balance condition for Y_n.

STEP 4 **in the study of FDTM:** DO LOTS OF PROBLEMS.