

## Chapter 5: Markov Chains

### Introduction: looking back, looking ahead.

Here, in outline form, is a quick review to set the stage for the work of this chapter.

#### *Review of the applied problem:*

- Randomization tests let us test scientific hypotheses by comparing an actual data set with data sets drawn at random according to some null model. (We use a test statistic to decide which is more extreme, the actual data set, or the random one.)
- For many situations, choosing at random can't be done in a straightforward way; instead, we start with a known element of the population, and make a long sequence of random "moves." If we choose the moves in the right way, and make enough of them, we get a random data set. We can then create a large collection of data sets to compute p-values, test hypotheses, etc.

#### *Review of the related mathematical questions:*

- We can represent the collection of all possible data sets as the vertices of a graph, with moves between data sets as edges. Choosing moves at random gives us a random walk on the graph.
- Given a graph, what is the limiting distribution of the walk on that graph? How does the limiting distribution depend on the structure of the graph? If the limiting distribution is not uniform, is it possible to alter the graph walk to get a uniform distribution?
- How quickly does the walk reach its limiting distribution, and how does the convergence rate depend on the structure of the graph?

#### *Preview:*

In Chapter 4, you saw how to represent a graph by its adjacency matrix, and how to use the  $n^{\text{th}}$  power of the matrix to count walks of length  $n$ . You also saw how to use tree diagrams to find two-step transition probabilities. In this chapter, you will see how to modify the adjacency matrix so that it shows actual transition probabilities. It will turn out that *powers of the resulting transition matrix give  $n$ -step transition probabilities*. Once we have that simple way to compute  $n$ -step transition probabilities, we can use them to study limiting distributions and convergence rates for random walks on graphs. Our program for the chapter, then will be to continue to study the same three questions from Chapter 2. Here are the versions for this chapter:

### **Question 1: $n$ -step transition probabilities.**

Given a set of 1-step transition probabilities, how can we find the  $n$ -step transition probabilities?

### **Question 2: limiting probabilities**

What happens to the  $n$ -step transition probabilities as  $n$  increases? If the  $n$ -step probabilities converge to a set of limiting values, do those values depend on the starting vertex? What about limiting values of  $\hat{p}$ , i.e., the observed fraction of steps spent at the various vertices?

How are they related to the limiting values for the  $n$ -step transition probabilities?

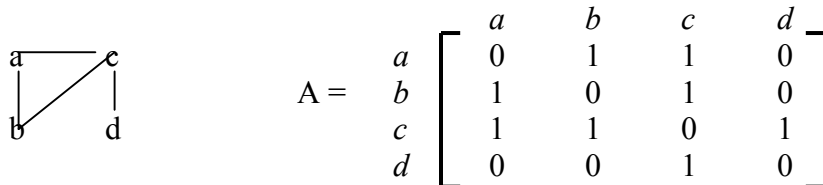
**Question 3: convergence rates.**

If the  $n$ -step transition probabilities do converge to limiting values, how quickly does this happen? How is the convergence rate related to the one-step transition probabilities?

The answers to the first two questions are known. The exercises and investigations of this chapter and the next are designed to guide you to recreate those answers for yourself. The answer to the third question is only partly known. To get ahead of the story, it will turn out that you can compute the rate of convergence whenever you can rewrite the matrix of transition probabilities in terms of its eigenvalues and eigenvectors.<sup>1</sup> This gives a complete answer to Question 3 for small graph walks, up to a few hundred vertices or so. For the finch data, however, the graph walk has somewhere around  $10^{17}$  vertices.<sup>2</sup> For such large problems, there are only partial answers.

**3.1 Question 1:  $n$ -step transition probabilities**

**Transition matrices.** For a graph walk, the adjacency matrix of the graph tells which vertices you can move to.



*Display 5.1 Adjacency matrix for a graph walk with four vertices*

For example, from  $a$  you can move to  $b$  or  $c$ , but not to  $d$ , so Row  $a$  of the adjacency matrix has 1s in the columns for  $b$  and  $c$ , and a 0 in the column for  $d$ . As you have seen, powers of  $A$  count walks: the  $(u,v)$  entry of  $A^n$  tells the number of walks of length  $n$  from  $u$  to  $v$ .

For a random walk on a graph, all paths away from a vertex are equally likely. This makes it easy to convert the adjacency matrix for the graph into a matrix of 1-step transition probabilities. In the example, the two moves away from  $a$  are equally likely, so each has probability  $1/2$ . To indicate this explicitly, we replace the two 1s in Row  $a$  by  $1/2$ s. In the same way, the three moves from  $c$  are equally likely, so we replace each 1 in that row by  $1/3$ . Doing this for each row gives a new matrix  $P$  called the **transition matrix** of the walk.

<sup>1</sup> The convergence rate depends on the behavior of powers of the transition matrix. You may remember from linear algebra that powers of a matrix are easy to find when you can write the matrix in terms of its “eigenstuff.” (If this rings no bells, don’t worry: It will be explained when the time comes.)

<sup>2</sup>Just imagine trying to diagonalize a  $10^{17} \times 10^{17}$  matrix!

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} p_{aa} & p_{ab} & p_{ac} & p_{ad} \\ p_{ba} & p_{bb} & p_{bc} & p_{bd} \\ p_{ca} & p_{cb} & p_{cc} & p_{cd} \\ p_{da} & p_{db} & p_{dc} & p_{dd} \end{bmatrix}$$

Display 5.2 Transition matrix for the graph walk of Display 5.1

Notice that the transition matrix, together with the starting state (or set of starting probabilities) gives all the information you need to simulate the random walk. *At any given step, where you go next depends only on where you are now.* More formally, the probabilities for  $X_{t+1}$  depend *only* on the value of  $X_t$ . This fact is the main defining property of a Markov chain.

### Drill exercises

1. Write the transition matrices for the walks on the two connected graphs of order 3.
2. Write the transition matrices for the walks on the six connected graphs of order 4.
3. Draw graphs corresponding to the following transition matrices.

$$a. \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \end{bmatrix} \quad b. \begin{bmatrix} 0 & 1/3 & 1/3 & 0 & 1/3 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \end{bmatrix}$$

$$c. \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix} \quad d. \begin{bmatrix} 0 & 1/2 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 & 0 \end{bmatrix}$$

### Discussion questions

**Definition.** A **stochastic**<sup>3</sup> **matrix** is a square matrix of non-negative entries, each of whose rows adds to 1. Equivalently, a **probability vector** is a vector of non-negative entries that add to 1; a **stochastic matrix** is a square matrix with probability vectors for rows.

<sup>3</sup> Stochastic means “chance-like.”

4. *True/False and Explain.* The transition matrix of a graph walk is always a stochastic matrix.
5. Give examples of two stochastic matrices that cannot be transition matrices of graph walks.
6. How can you tell whether a stochastic matrix is the transition matrix for some graph walk?

*Informal definition.* The random walk defined by a stochastic matrix is called a **Markov chain**; the stochastic matrix is its **transition matrix**.

7. Write the transition matrix for each of the Markov chains described below.
  - a. *A tiny model of the US economy* consists of three Federal Reserve districts, Atlanta, Boston, and Chicago. Each quarter, one third of the money that starts in Atlanta stays there, one third goes to Boston, and one third goes to Chicago. During the same quarter, half the money that starts the quarter in Boston stays there; of the rest, half goes to Atlanta, half to Chicago. Each quarter,  $3/4$  of the money that starts in Chicago stays there; the other  $1/4$  goes to Atlanta.
  - b. *A simplified model of pharmaco-kinetics* reduces a person to a system of just three organs -- stomach, blood, and liver/kidneys. Suppose you give this abstractly compartmentalized person a dose of cough syrup at time  $t = 0$ . Each half hour after that, 25% of the drug in the stomach has been absorbed into the blood; the other 75% remains in the stomach. Also during each half hour, 10% of the drug in the blood is removed by the liver/kidneys; the rest remains in the blood. None of the drug returns to the blood or stomach from the liver/kidneys.

**Note:** The two scenarios in Problem 12 are deliberately oversimplified, but larger versions can be both realistic and useful for applied work, as in the next two examples.

*Example:* Researchers Freedle and Lewis<sup>4</sup> used a Markov chain as a mathematical model for interactions between a mother and her baby. Their chain had six states:

- 1: neither mother nor infant vocalizes
- 2: infant vocalizes, mother does not
- 3: infant does not vocalize, mother vocalizes to the infant
- 4: infant does not vocalize, mother vocalizes to someone else
- 5: infant vocalizes, mother vocalizes to the infant
- 6: infant vocalizes, mother vocalizes to someone else

---

<sup>4</sup> Freedle, R.O. and M. Lewis (1971). "Application of Markov processes to the concept of state," *Research Bulletin 71-34*, Princeton, NJ: Educational Testing Service. Example taken from Ingram Olkin, Leon J. Gleser, and Cyrus Derman (1980). *Probability Models and Applications*, New York: Macmillan, pp. 440-443.

Freedle and Lewis recorded, every 10 seconds, which of 1-6 matched the behavior of the mother and infant. Their data suggested that the behavior could be described by a Markov chain with the transition probabilities shown in Display 5.3.

### Discussion question

8. (a) Use the transition matrix to decide which behavior was most likely to continue from one 10-second observation to the next. (b) Apparently only one of the six categories of behavior was unlikely to go on for very long. Which one was it? (c) By examining the transition matrix, imagine what a typical sequence of steps would look like.

$$\mathbf{P} = \begin{bmatrix} .42 & .09 & .13 & .22 & .02 & .12 \\ .22 & .46 & .00 & .08 & .02 & .22 \\ .18 & .04 & .51 & .12 & .05 & .10 \\ .05 & .01 & .05 & .71 & .01 & .17 \\ .27 & .13 & .20 & .07 & .07 & .26 \\ .05 & .06 & .01 & .33 & .02 & .53 \end{bmatrix}$$

*Display 5.3 Estimated transition probabilities for mother-infant vocalizations*

9. I simulated 100 steps of the chain, starting in State 1. Here's the sequence, with spaces after every five digits:

```
14444 66622 14445 66666 64446 44444 44466 66444 44444 66664
44444 66444 44656 44466 66411 35133 31111 44644 62244 64116
```

Compare this sequence with the one you imagined in (8c). Does this look like the one you imagined? If not, do you consider this walk atypical in some respects, or are there features of the transition matrix you overlooked?

10. You can think of the data Freedle and Lewis used to estimate transition probabilities as being like the one in (9), only much longer. Based on this sequence, what would be your estimated value for  $p_{14}$ ? Explain your reasoning.

11. This Markov chain is not a graph walk. How can you tell?

*Example.* DNA is built from amino acid bases adenine (A), cytosine (C), guanine (G), and thymine (T), arranged in a linear order from the starting (5') end to the other (3') end. If you let  $X_0$  be the starting base,  $X_1$  the next base, and so on, it is approximately true that the

probabilities for  $X_{t+1}$  depend only on  $X_t$ . That is, the sequence of bases behaves like a Markov chain. Display 5.4 gives the transition matrix.<sup>5</sup>

$$\mathbf{P} = \begin{matrix} & \begin{matrix} A \\ C \\ G \\ T \end{matrix} \\ \begin{matrix} A \\ C \\ G \\ T \end{matrix} & \begin{bmatrix} .32 & .18 & .23 & .27 \\ .37 & .23 & .05 & .35 \\ .30 & .21 & .25 & .24 \\ .23 & .19 & .25 & .33 \end{bmatrix} \end{matrix}$$

*Display 5.4 Transition matrix for the sequence of amino acid bases in strands of DNA*

**Discussion questions:**

12. Of the two bases, two are more frequent than the other two. Which ones are they, and how can you tell?

13. *A new chain.* The enzyme AluI “recognizes” the four-base sequence AGCT: it cuts a DNA strand anywhere that sequence occurs, and only at those places. In order to study the frequency of the sequence AGCT, researchers<sup>6</sup> used the four-state transition matrix from Display 5.4 to construct a new chain with seven states A, C, G, T, AG, AGC, and AGCT. (Here state AG means the current base is G and the previous base is A.)

Assume  $X_t = A$ . Use the transition matrix in Display 5.4 to find

- $P(X_{t+1} = G \mid X_t = A)$
- $P(X_{t+1} = G \text{ and } X_{t+2} = C \mid X_t = A)$
- $P(X_{t+1} = G \text{ and } X_{t+2} = C \text{ and } X_{t+3} = T \mid X_t = A)$

14. Consider transitions from A, referring to the first row of Display 5.4 as needed. Explain why  $P(A \rightarrow G) = 0$  for the new chain whose states are A, C, G, T, AG, AGC, and AGCT. What other transitions from A have probability 0? Write the first row of the new transition matrix.

15. Now consider transitions from G. Explain why transitions from G to A, C, G and T have the same probabilities as in the original chain, even though that was not true of all such transitions from A. Write the rows of the new transition matrix that correspond to states C, G, and T.

<sup>5</sup> Swartz, M.N., T.A. Trautner, and A. Kornberg (1962). “Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest-neighbor base sequences in deoxyribonucleic acids.” *J. Biol. Chem.* **237**: 864-875. Taken from Lange, Kenneth (1997). *Mathematical and Statistical Methods for Genetic Analysis*. New York: Springer-Verlag, pp. 145-6.

<sup>6</sup> Bishop, D.T., J.A. Williamson, and M.H. Skolnick (1983). “A model for restriction fragment length distributions,” *Amer. J. Human Genetics* **35**: 795-815.

16. Next, consider transitions from AG. Which three of the seven states cannot be reached from AG? For which destination states are the transition probabilities from AG the same as from G? What is the transition probability from AG to AGC? Write the row of the transition matrix for AG.

17. Complete the transition matrix.

18. The limiting probabilities for this Markov chain are:

A	C	G	T	AG	AGC	AGCT
0.301	0.185	0.130	0.294	0.070	0.015	0.005

What does this tell you about the average distance between “restriction sites”, that is, between occurrences of AGCT?

### Exercises: Transition probabilities

#### Preliminary drill: recognizing matrix products

To work skillfully with products of matrices, learning to multiply matrices is only the first step. An important second step is to learn to recognize the results of matrix multiplication when it appears in a different context. That’s the goal of these drill exercises.

19. Write out the following products:

$$\text{a. } (a_1 \ a_2 \ a_3) \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad \text{b. } \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$\text{c. } (a_1 \ a_2 \ \dots \ a_n) \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \quad \text{d. } (a_{i,1} \ a_{i,2} \ \dots \ a_{i,n}) \begin{bmatrix} b_{1,j} \\ b_{2,j} \\ \vdots \\ b_{n,j} \end{bmatrix}$$

20. Let  $C = \{c_{ij}\}$  be an  $r \times s$  matrix.

- Write a sum that gives the (1,2) element of  $CC^T$ .
- Write a sum that gives the (3,5) element of  $C^T C$ .
- Suppose  $C$  is a co-occurrence matrix, with  $c_{ij} = 1$  if species  $i$  occurs on island  $j$ .

What does the (1,2) element of  $CC^T$  tell? What does the (3,5) element of  $C^T C$  tell?

21. Matching. Each matrix element in column (a) is the matrix product of a row vector from column (b) and a column vector from column (c). For each element in (a), tell the appropriate row from (b) and column from (c).

(a)	(b)	(c)
$(C^T C)_{ij}$	$i^{\text{th}}$ row of $C$	$i^{\text{th}}$ col of $C$
$(C C^T)_{ij}$	$i^{\text{th}}$ row of $C^T$	$i^{\text{th}}$ col of $C^T$
$(C^T C)_{ji}$	$j^{\text{th}}$ row of $C$	$j^{\text{th}}$ col of $C$
$(C C^T)_{ji}$	$j^{\text{th}}$ row of $C^T$	$j^{\text{th}}$ col of $C^T$

22. Let  $\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}$       Abbreviation:  $\mathbf{P} = \{p_{ij}\}_{3 \times 3}$

- a. Write out a formula for the (1,2) element of  $\mathbf{P}^2$ .
- b. Now let  $\mathbf{Q} = \{q_{ij}\}_{4 \times 4}$ . Based on your answer to (a), write out a formula for the (2,3) element of  $\mathbf{Q}^2$ .
- c. Let  $\mathbf{P}^2 = \{p^{[2]}_{ij}\}_{3 \times 3}$ , that is, let  $p^{[2]}_{ij}$  be the (i,j) element of  $\mathbf{P}^2$ . (This is not standard notation, but the similar notation using parenthesis  $p^{(2)}_{ij}$  has already been used, for two-step transition probabilities, so we can't use it here with a different meaning. We'll only need the non-standard notation temporarily.) Similarly, let  $\mathbf{P}^3 = \{p^{[3]}_{ij}\}$ . Using the fact that  $\mathbf{P}^3 = \mathbf{P}^2 \mathbf{P}$ , and following the pattern of your answer to (a) as a guide, write out a formula for  $p^{[3]}_{ij}$  in terms of the  $p^{[2]}_{ij}$  and  $p_{ij}$

23. Let  $\mathbf{P}$  be as in the previous problem. Define unit vectors  $\mathbf{u}_1 = (1 \ 0 \ 0)$ ,  $\mathbf{u}_2 = (0 \ 1 \ 0)$ , and  $\mathbf{u}_3 = (0 \ 0 \ 1)$ . Find the following products, then tell in words what multiplication by unit vectors like these does.

- a.  $\mathbf{u}_1 \mathbf{P}$
- b.  $\mathbf{u}_2 \mathbf{P}$
- c.  $\mathbf{P} \mathbf{u}_1^T$
- d.  $\mathbf{P} \mathbf{u}_3^T$

**Two-step transition probabilities for Markov chains**

Consider the transition matrix for the Federal Reserve example, but for the sake of this exercise, use subscripted letters to represent the transition probabilities.

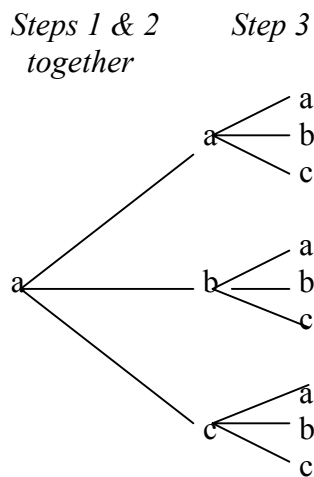
$$P = \begin{bmatrix} p_{aa} & p_{ab} & p_{ac} \\ p_{ba} & p_{bb} & p_{bc} \\ p_{ca} & p_{cb} & p_{cc} \end{bmatrix}$$

## 24. Two-step probabilities.

- Use a tree diagram to find algebraic expressions for the two-step transition probabilities from  $a$  to  $a$ ,  $a$  to  $b$ , and  $a$  to  $c$ .
- Based on the patterns in the previous problem, write algebraic expressions for two-step transition probabilities from  $b$  to  $a$ ,  $b$  to  $b$ , and  $b$  to  $c$ .
- Suppose  $\mathbf{P}$  is a  $5 \times 5$  stochastic matrix with elements  $p_{ij}$  with  $i, j = 1, 2, \dots, 5$ . Write an expression in terms of the  $p_{ij}$ , for  $p^{(2)}[3, 5] = p^{(2)}_{3,5}$ .

**Three-step transition probabilities.**

Now that you can compute two-step transition probabilities like  $p^{(2)}[a, b] = p^{(2)}_{ab}$ , you can use them to find three-step transition probabilities. Display 5.5 shows one possible tree diagram, with a three-step transition shown as a two-step transition followed by a one-step transition:



*Display 5.5 Tree diagram for 3-step transitions for the Federal Reserve example*

## 25. Three-step probabilities.

- Label the branches with appropriate probabilities. Use elements of  $\mathbf{P}$  for 1-step transitions, elements of  $\mathbf{P}^{(2)}$  for 2-step transitions.
- Explain what the notation  $a-b-c$  on this tree diagram tells you about the walk of a dollar bill: Where was it at each of time 0, 1, 2 and 3? (The notation and tree diagram let you determine all but one of these; that one can't be decided without additional information.)

- c. Explain why the probability of the walks corresponding to the notation  $a—b—c$  is  $p^{(2)}_{ab}p_{bc}$ .
- d. Now write formulas for the 3-step transition probabilities from  $a$  to  $a$ , and  $a$  to  $b$ .

26. *A complementary tree diagram.*

- a. Display 5.5 shows each 3-step transition as a 2-step transition followed by a 1-step transition. Draw and label a tree diagram that shows each 3-step transition as a 1-step transition followed by a 2-step transition.
- b. For your new tree, explain what the notation  $a—b—c$  tells about the walk by a dollar bill: Where was it at time 0, 1, 2, 3? (Here, as before, the vertex at one of the four times can't be determined.)
- c. Using your new tree diagram, compute 3-step transition probabilities from  $a$  to  $a$ , and  $a$  to  $b$ .

27. *Comparing the two*

- a. Compare your 3-step transition probabilities in 26(c) with the ones in 25(d) based on Display 3.2. Should the two sets of transition probabilities be equal? (Explain.) Are they in fact equal? (How can you tell?)
- b. State a general rule for finding 3-step transition probabilities.

**Discussion questions:  $n$ -step transition probabilities.**

28. You can use the same ideas to find  $\mathbf{P}^{(4)}$ . Each 4-step transition can be written as a 3-step transition followed by a 1-step transition, or, as a 2-step transition followed by another 2-step transition, or, as a 1-step transition followed by a 3-step transition. Explain why all three representations lead to the same transition probabilities.

29. Compare the two matrix equations  $\mathbf{P}^k\mathbf{P}^{n-k} = \mathbf{P}^n$  and  $\mathbf{P}^{(k)}\mathbf{P}^{(n-k)} = \mathbf{P}^{(n)}$ . One is trivial, the other is deep. One equation has no mathematical substance; it is largely a matter of notation. The other is a very compact statement of a result about Markov chains that is not at all obvious. Which is which? Explain what each one says. (The deep result has a name: the **Chapman-Kolmogorov equation**.)

## 5.2 Question 2: Limiting distributions (continued)

What follows is set in the context of the simple Federal Reserve model whose transition matrix is

$$P = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 0 & 3/4 \end{bmatrix} = \begin{bmatrix} p_{aa} & p_{ab} & p_{ac} \\ p_{ba} & p_{bb} & p_{bc} \\ p_{ca} & p_{cb} & p_{cc} \end{bmatrix}$$

At times it will be easier to think intuitively using the concrete numerical version on the left; at other times it will be easier to see general patterns using the abstract version on the right.

Here are some questions that set the agenda for the next few pages. Imagine that at the Beginning of Economic Time, all the money in our little universe is concentrated in Atlanta. At the instant of the monetary Big Bang, the money starts spreading, step by step, according to the transition probabilities in the matrix  $P$ . How will the money be distributed after one step? after two steps? after  $n$  steps? What happens to the distribution as  $n$  increases without limit? Now imagine a parallel universe. For this one, all the money starts in Boston. How will it be distributed after  $n$  steps? Will the distribution at time  $n$  be the same as for the first universe? Will the limiting distribution be the same as for the first universe?

### Some notation:

Let  $p^{(n)}[a]$  be the fraction of money in  $a$  at time  $n$ . More generally, for any Markov chain, let  $p^{(n)}[a]$  be the probability that the process is in state  $a$  at time  $n$ :

$$p^{(n)}[a] = P(X_n = a).$$

Thus  $p^{(0)}[a] = 1$  means the process starts in  $a$ . Let the vector  $\mathbf{p}^{(n)} = (p^{(n)}[a] \ p^{(n)}[b] \ p^{(n)}[c])$  give the distribution at time  $n$ . In this notation, for example,  $\mathbf{p}^{(0)} = (1, 0, 0)$  means the random walk starts in  $a$ .

### Investigation: Limiting distributions

30.  $n$ -step transition probabilities. What happens to  $\mathbf{P}^n$  as  $n \rightarrow \infty$ ?
- Pick some small graphs, find their transition matrices, and use a computer to compute  $\mathbf{P}^n$  for increasing values of  $n$ .
  - Is the limiting behavior of  $\mathbf{P}^n$  the same for all graphs?

**Definition.** If  $\mathbf{p}^{(n)} = \mathbf{p}^{(0)}\mathbf{P}^n$  converges to a limit vector  $\mathbf{p}^\infty$  as  $n \rightarrow \infty$ , that vector is called the **limiting distribution** of the Markov chain.<sup>7</sup>

<sup>7</sup> Here  $\mathbf{p}^{(n)} \rightarrow \mathbf{p}^\infty$  means  $p^{(n)}[i] \rightarrow p^\infty[i]$  for all states  $i$ .

31. Limiting distributions. Find enough examples to support answers to the following questions:
- Does every graph walk have a limiting distribution?
  - Does any graph walk have more than one limiting distribution?
  - Does the limiting distribution ever depend on the starting vertex? Always depend on the starting vertex?

Finding  $\mathbf{p}^\infty$  is a challenge. You can do it numerically by computing  $\mathbf{p}^{(0)}\mathbf{P}^n$  for increasing values of  $n$ , but this won't help you prove that the limiting probabilities for a graph walk are always proportional to the vertex degrees. For that, you need an algebraic approach.

**Definition.** A probability vector  $\mathbf{x}$  is a **stationary distribution** (or **equilibrium vector** or **steady state vector**, or **fixed point**) for a Markov chain with transition matrix  $\mathbf{P}$  if and only if

$$\mathbf{p}^{(n)} = \mathbf{x} \Rightarrow \mathbf{p}^{(n+1)} = \mathbf{x} \text{ for any and all } n.$$

In words, once the distribution reaches a stationary distribution, it never changes.

32. Use the relationship between  $\mathbf{p}^{(n)}$  and  $\mathbf{p}^{(n+1)}$  to find an equation that the equilibrium vector  $\mathbf{x}$  must satisfy.
33. Use your equation in (32) to prove that for a graph walk, the vector with components  $d_i / \sum d_i$  is a stationary distribution.
34. If the transition matrix  $\mathbf{P}$  of a Markov chain is symmetric ( $\mathbf{P} = \mathbf{P}^T$ ) what can you say about the stationary distribution of the chain?
35. Suppose  $\mathbf{u} = (1/k, 1/k, \dots, 1/k)$  is a stationary distribution for a Markov chain with transition matrix  $\mathbf{P}$ . What must be true of the column sums of  $\mathbf{P}$ ?
36. What is the difference between an equilibrium distribution and the limiting distribution? How are the two related to each other?

**Definition.** The **ergodic average vector**  $\hat{\mathbf{p}}_{(n)}$  gives the observed proportion of steps spent in each state:  $\hat{p}_{(n)}[i] = \#\{X_t = i \mid 1 \leq t \leq n\} / n$ .

37. Observed proportion of time in state  $i$ .
- Does the ergodic average vector always converge to a limit?
  - Give an example of a chain for which  $\hat{\mathbf{p}}_{(n)}$  converges but  $\mathbf{p}^{(n)}$  does not.
  - Suppose that  $\hat{\mathbf{p}}_{(n)}$  converges to a limit vector  $\hat{\mathbf{p}}_{(\infty)}$  and that  $\mathbf{x}$  is a stationary vector for  $\mathbf{P}$ . Must  $\hat{\mathbf{p}}_{(n)}$  necessarily equal  $\mathbf{x}$ ? Give examples to support your answer.
  - Suppose  $\hat{\mathbf{p}}_{(n)}$  and  $\mathbf{p}^{(n)}$  both converge. Will the limit vectors always be equal?

## Basic Matrix Operations on the Computer

### MathCad:

#### Entering a matrix and naming it A:

1. Type  $A$
2. Click the equality/inequality button on the tool bar to open the corresponding palette.
3. Select  $:=$  from the palette. (Use this symbol for  $=$  in a definition.)
4. Click the matrix button on the tool bar to open the corresponding palette.
5. Select the matrix button and enter the dimensions.
6. Then enter the elements of the matrix itself, using the tab key to move between entries.

### Exercises:

M1. Enter the matrix  $A$ , column vector  $x$ , column vector  $y$ , and matrix  $B$  as shown in Display 5.6.

M2. Decide which of the various matrix and vector products  $AB^T$ ,  $y^T A$ ,  $Bx^T$ ,  $Bx$ ,  $A^T A$ ,  $AA^T$  can be computed, and which cannot.

M3. Now use MathCad to carry out the multiplications or confirm that they aren't defined. For matrix multiplication, use the button on the matrix palette marked  $\bar{x} \cdot \bar{y}$ .

M4. Experiment with the system until you figure out how to find matrix inverses and solutions to systems of linear equations.

### S-Plus:

Display 5.7 illustrates basic matrix operations that parallel the ones just described above for MathCad.

### Exercises.

S1 – S4. Read through the commands in the display. Then use similar commands to enter the matrix  $B$  and compute the matrix products involving  $B$  from the MathCad exercises M1 – M4.

$$A := \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad B := \begin{pmatrix} 1 & 1 & 0 & 1 \\ 2 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 \end{pmatrix} \quad x := \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad y := \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$A \cdot B^T = \begin{pmatrix} 2 & 2 & 3 \\ 2 & 2 & 4 \\ 1 & 0 & 2 \end{pmatrix} \quad y^T \cdot A = (2 \ 2 \ 1 \ 1) \quad B \cdot x^T =$$

array size mismatch

$$B \cdot x = \begin{pmatrix} 3 \\ 2 \\ 5 \end{pmatrix} \quad A^T \cdot A = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad A \cdot A^T = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

$$P := A \cdot A^T \quad \text{lsolve}(P, y) = \begin{pmatrix} 0.5 \\ 0 \\ 0.5 \end{pmatrix} \quad P^{-1} = \begin{pmatrix} 0.75 & -0.5 & 0.25 \\ -0.5 & 1 & -0.5 \\ 0.25 & -0.5 & 0.75 \end{pmatrix}$$

*Display 5.6 Basic Matrix Operations in MathCad*

```

> row1 <- c(1, 0, 0, 1)
> row2 <- c(1, 1, 0, 0)
> row3 <- c(0, 1, 1, 0)

> A <- rbind(row1, row2, row3)
# This will also work: A <- matrix(c(row1,row2,row3),3,4,byrow=T)
> A
      [,1] [,2] [,3] [,4]
row1    1    0    0    1
row2    1    1    0    0
row3    0    1    1    0

> t(A)
      row1 row2 row3
[1,]    1    1    0
[2,]    0    1    1
[3,]    0    0    1
[4,]    1    0    0

> P <- A %**% t(A)
> P
      row1 row2 row3
row1    2    1    0
row2    1    2    1
row3    0    1    2

> Q <- t(A) %**% A
> Q
      [,1] [,2] [,3] [,4]
[1,]    2    1    0    1
[2,]    1    2    1    0
[3,]    0    1    1    0
[4,]    1    0    0    1

> x <- t(rep(1, 4))
> x
      [,1] [,2] [,3] [,4]
[1,]    1    1    1    1

> y <- t(rep(1, 3))
> y
      [,1] [,2] [,3]
[1,]    1    1    1

> solve(P, t(y))
      [,1]
row1 5.000000e-001
row2 1.990455e-016
row3 5.000000e-001

> solve(P) # gives the inverse of P if it has one
      row1 row2 row3
row1 0.75 -0.5 0.25
row2 -0.50 1.0 -0.50
row3 0.25 -0.5 0.75

```

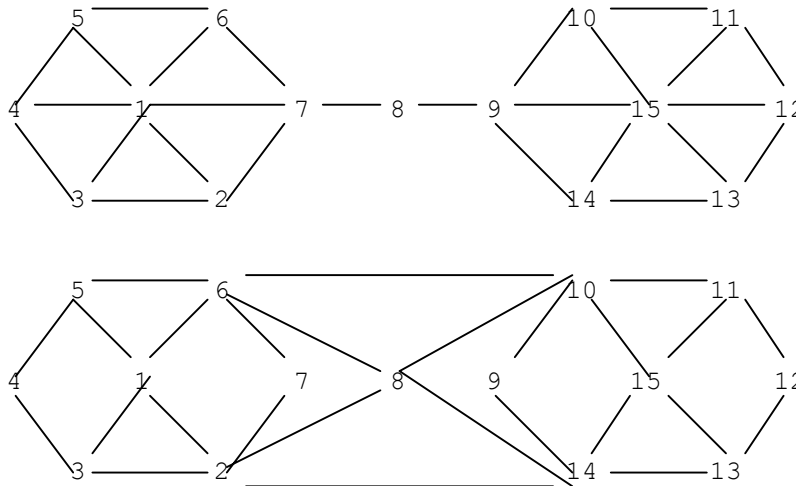
*Display 5.7 Basic Matrix Operations in S-Plus*

### 5.3 Question 3: Convergence rates (continued)

In what follows, you'll begin a transition from thinking about convergence rates intuitively based on the appearance of a graph to looking for numerical patterns based on more formal definitions.

#### Discussion: Making predictions based on intuition

38. Two graphs are shown below in Display 5.8. Each has 15 vertices and 26 edges. Imagine simulating 100 steps of a random walk on each graph, starting from vertex 1. In what ways do you expect the two walks to differ? Explain your reasoning. For which graph would you predict faster convergence of the  $n$ -step transition probabilities to the limiting probabilities? Why? (If you have trouble imagining how the graph walks will behave, use S-Plus to generate a few walks and examine their behavior.)



Display 5.8 Two graphs, each with 15 vertices and 26 edges

#### Activity: Convergence rates

Refer to the graphs in Display 5.9. Each graph has seven vertices, six that form a hexagon, plus a center point. Think of the vertices numbered, starting with 1 for the leftmost vertex, continuing clockwise around the hexagon with 2, 3, ..., 6 and ending with 7 for the center vertex.

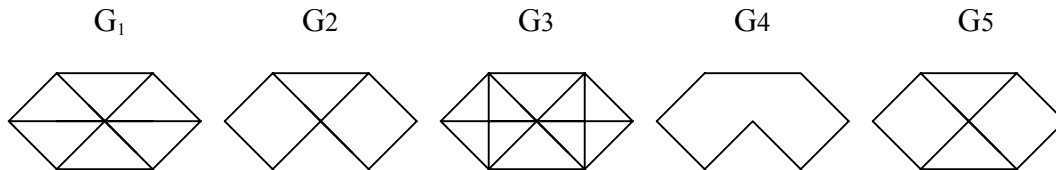
#### Part A. Exploring convergence rates

*Step 1.* Find the limiting probabilities for a random walk on each graph.

*Step 2.* Now think about the behavior of a walk on each graph, and imagine how the first few steps of the walk might go. Next, extend your imagined walks to a much larger number of steps, and think about how quickly or slowly the  $n$ -step transition probabilities for the walk

will converge to their limiting probabilities. Call this the **mixing rate**. (We don't at this point have a way to measure the mixing rate numerically, but you can think comparatively using an intuitive sense of faster and slower.) For each of the ten pairs of graphs you can form from the five graphs below, predict which one of the pair will converge faster, and tell why.

*Step 3.* Use your results from Step 2 to put the graphs in order of predicted mixing rate, from slowest to fastest.



Display 5.9 Five 7-point graphs

### Part B. Assigning numbers to graph features

As part of your work in the last activity, you no doubt developed some ideas about which features of a graph would be associated with faster convergence, and how to tell visually whether a walk on a graph would converge slowly or quickly. Some of these features of graphs can be made quantitative, and indeed graph theorists have developed many ways to assign numbers to graphs. As a quick review before looking at several of these, remember that a **path** is a walk with no repeated vertices, and a **cycle** is a walk of length 3 or more with only one repeated vertex (its beginning and ending point), and the **length** of any walk is its number of edges.

- The **distance** between two vertices is the length of the shortest path joining them. The **diameter** of a graph is the largest distance between any two vertices of the graph.
- The **circumference** of a graph is the length of its longest *cycle*.
- The **girth** of a graph is the length of its *shortest cycle*.
- A **coloring** of a graph is an assignment of colors to vertices for which no adjacent vertices have the same color. The **chromatic number** of a graph is the smallest number of colors needed to color it.
- An **independence set** of vertices is a set for which no two vertices in the set are neighbors. The **independence number** of a graph is the number of elements in the largest independence set.
- The **degree** of a vertex is its number of neighbors. The **average degree** of a graph is the sum of its vertex degrees divided by the number of vertices.

*Step 4.* Find the values of these quantities for the graphs in Display 5.9, and use them to complete the chart in Display 5.10.

**Discussion questions.**

39. Which features seem to be most closely associated with your intuitive sense of how fast a graph walk reaches its limiting distribution?

40. Based on your work in the activity, invent a measure that distinguishes the two graphs in Display 5.8.

	G1	G2	G3	G4	G5
diameter					
circumference					
girth					
chromatic number					
independence number					
average degree					
maximum degree					
minimum degree					

*Display 5.10 Summary table of features of the graphs in Display 5.9*

Take a moment to think again about the difference between observed frequencies  $\hat{p}_{(n)}[i]$  and  $n$ -step transition probabilities  $p^{(n)}[i]$ . It may help to rely here on a concrete example, such as the walk on a triangle  $abc$ , starting in vertex  $a$ . Both the observed frequency  $\hat{p}_{(n)}[a]$  and the  $n$ -step transition probability  $p^{(n)}[i]$  converge to a limiting value of  $1/3$ , but the behavior is different. For one thing, the observed frequencies  $\hat{p}_{(n)}[a]$  are “things you can see” in the sense that they are computed from actual observed outcomes. For example, here is a random path and the set of observed frequencies computed from it:

Time	0	1	2	3	4	5	6	7	8	9	10
State	$a$	$b$	$a$	$b$	$a$	$b$	$a$	$c$	$a$	$b$	$a$
Freq. $\hat{p}_{(n)}[a]$		0	1/2	1/3	2/4	2/5	3/6	3/7	4/8	4/9	5/10

Although these frequencies are observable, they are based on random outcomes, and so the values of  $\hat{p}_{(n)}[a]$  change from one walk to the next. Here are the results from a second walk on the same triangle. The transition probabilities haven’t changed, but the observed frequencies have:

Time	0	1	2	3	4	5	6	7	8	9	10
State	$a$	$b$	$c$	$b$	$c$	$a$	$b$	$a$	$b$	$a$	$b$
Freq. $\hat{p}_{(n)}[a]$		0	0	0	0	1/5	1/6	2/7	2/8	3/9	3/10

The  $n$ -step transition probabilities are theoretical quantities, computed from powers of the transition matrix. For many students, they are harder to think about initially, because you can't compute them directly from observed data. On the other hand, because they don't depend on random outcomes the way the  $\hat{p}$  values do, their behavior tends to be simpler.

### Part C. Computational Investigation

To this point, your work with convergence rates of random walks on graphs has been largely visual and intuitive. One way to be more systematic about convergence rates is to quantify the distance between an  $n$ -step transition probability  $p^{(n)}[i]$  and its limiting value. For the following investigation, you will need to pick one of the graphs from Display 5.9 to work with. Assume the starting state is the leftmost vertex, 1.

*Step 1.* Find the stationary distribution for your graph, and the probability  $p^{(\infty)}[7]$  that it assigns to the center vertex.

*Step 2.* Find the transition matrix  $\mathbf{P}$ , and its (1,7) element, which is  $p^{(1)}[7]$ . Compute the distance  $|p^{(1)}[7] - p^{(\infty)}[7]|$ .

*Step 3.* Use S-plus to find the two-step transition matrix  $\mathbf{P}^2$ , and its (1,7) element,  $p^{(2)}[7]$ . Then compute the distance  $|p^{(2)}[7] - p^{(\infty)}[7]|$ . (In S-plus, the command for matrix multiplication is `%*%`.)

*Step 4.* Compute  $\mathbf{P}^n$ ,  $p^{(n)}[7]$ , and the distance  $|p^{(n)}[7] - p^{(\infty)}[7]|$  for  $n = 2, 3, 4, \dots, 100$ . You may want to use the following S-plus function to compute powers of  $\mathbf{P}$ :

```
Power <- function(P,n){
  A <- diag(rep(1,dim(P)[1]))
  for (i in 1:n){A <- A %*% P}
  return(A)
}
# Computes the nth power of P
# Identity matrix
# Multiply by P n times
# Return the product
```

*Step 5.* Find the functional form of the relationship between the distance and the number of steps. In Chapter 3 you found that the relationship between the error in estimating a  $p$ -value and the sample size  $N_{\text{Rep}}$  followed a power law with power =  $-1/2$ . What do you predict for the relationship between the distance  $|p^{(n)}[7] - p^{(\infty)}[7]|$  and the number of steps? Graph

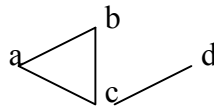
- the distance versus  $n$ ,
- the log distance versus  $n$ ,
- the distance versus  $\log(n)$ , and
- the log distance versus  $\log(n)$ .

Is convergence linear, logarithmic, exponential, power law, or none of these? If one of your four graphs gives a line, find its slope.

**Total variation norm: the distance between two distributions**

So far in our study of convergence rates, we have either worked without a formal definition of closeness, or we have looked at only one vertex at a time, asking how the distance  $|p^{(n)}[i] - p^\infty[i]|$  behaves as the number of steps increases. Why not replace the one-at-a-time approach with a whole-graph approach?<sup>8</sup>

*Example:* Consider the graph



whose limiting probabilities are  $\mathbf{p} = (p_a, p_b, p_c, p_d) = (2/8, 2/8, 3/8, 1/8)$ . I simulated a random walk on this graph four times. The first walk ran for 25 steps, the next ran for 100 steps, the next for 400, and the last for 1600 steps. Here are the fractions of steps spent at each vertex:

Number of steps	Observed proportion of steps in			
	a	b	c	d
25	0.3200	0.2800	0.3600	0.0400
100	0.2200	0.1900	0.4200	0.1700
400	0.2475	0.2300	0.3900	0.1325
2000	0.2500	0.2385	0.3700	0.1325
<i>infinite</i>	<i>0.2500</i>	<i>0.2500</i>	<i>0.3750</i>	<i>0.1250</i>

*Display 5.11. Fraction of steps spent at various vertices, for walks of various lengths*

As the table makes clear, at each vertex the sequence of  $\hat{p}$  values converges to its limiting value. How can we combine the set of values to get a measure of how far the vector of  $\hat{p}$  values is from the vector of limiting probabilities after  $n$  steps? One simple answer is to add the individual distances  $|\hat{p}_{(n)}[i] - p^\infty[i]|$  together. For technical reasons (see the drill problems) we then divide by 2. This “distance” between probability vectors is called the **total variation distance** (or often just the variation distance):

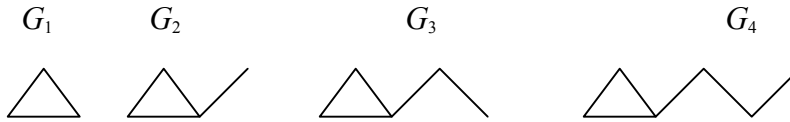
$$\|\hat{p} - p\| = (1/2) \sum_v |\hat{p}(v) - p(v)|$$

**Drill**

41. For each of the rows of Display 5.11, compute the variation distance between the observed frequencies and the limiting probabilities.

<sup>8</sup> As a child Benjamin Franklin (young and irreverent) once suggested to his father (old and pious) that they could save a lot of time by saying a collective blessing for the whole family larder instead of saying a large number of blessings for individual meals. Irreverence and piety aside, the idea of grouping like objects and treating them as a single structure is one of the important recurring themes that characterizes mathematical thinking.

42. Consider the stationary distributions  $\mathbf{p}_k$  for the set of graphs that begins



Compute the following distances:

$$\|p_1 - p_2\| \quad \|p_2 - p_3\| \quad \|p_{n-1} - p_n\|$$

where  $p_i(1) = p_i(2) = p_i(3) = 1/3$ ,  $p_i(i) = 0$  for  $i > 3$ .

$p_2(1) = p_2(2) = 2/8$ ,  $p_2(3) = 3/8$ ,  $p_2(4) = 1/8$ ,  $p_2(i) = 0$  for  $i > 4$ .

$p_3(1) = p_3(2) = p_3(4) = 2/10$ ,  $p_3(3) = 3/10$ ,  $p_3(5) = 1/10$ ,  $p_3(i) = 0$  for  $i > 5$ .

etc.

43. *Why 1/2?* Suppose we had defined an unnormalized distance, without the factor of 1/2:

$$\|p_1 - p_2\|_* = \sum_v |p_1(v) - p_2(v)|$$

a. Consider two probability distributions (vectors),  $\mathbf{p}_1$  and  $\mathbf{p}_2$ , each with only two outcomes, 0 and 1. Suppose the probabilities are given in terms of constants  $a$  and  $b$ :

$$p_1(0) = a, \quad p_1(1) = 1-a, \quad \text{and so } \mathbf{p}_1 = (a, 1-a)$$

$$p_2(0) = b, \quad p_2(1) = 1-b, \quad \text{and so } \mathbf{p}_2 = (b, 1-b)$$

Compute the unnormalized distance between  $\mathbf{p}_1$  and  $\mathbf{p}_2$  (i) if  $a = b = 1/2$ , (ii) if  $a = 1/2$ ,  $b = 1/4$ .

b. Notice that any values of  $a$  and  $b$  between 0 and 1 are possible. What choices for  $a$  and  $b$  make the distance above (without the factor of 1/2) as large as possible? What is this largest possible value? What is the largest possible value of the variation distance? Why do you think the factor 1/2 is included in the definition?

c. Now let  $\mathbf{p}_1$  and  $\mathbf{p}_2$  stand for any two probability distributions with an arbitrary number  $k$  of outcomes. What is the largest possible value of the variation norm?

44. Write S-Plus code to compute the value of the variation distance between two probability vectors  $\mathbf{p}_1$  and  $\mathbf{p}_2$ .

## Investigations

45. Functional form of the relationship between  $n$ -step transition probabilities and the limiting distribution. Based on your work in Part C of the activity, what pattern do you predict for the convergence of  $n$ -step transition probabilities? Work with the same graph you used in Part C of the activity, and start your walk at the left-most vertex. Let  $\mathbf{p}^{(n)}$  be the vector of probabilities after  $n$  steps, assuming the walk starts at a particular vertex  $v_0$ .

$$\mathbf{p}_n(v) = P\{X_n = v | X_0 = v_0\}$$

$$\mathbf{p}_n = (p_n(v_1), \dots, p_n(v_k))$$

Graph  $\|\mathbf{p}^{(n)} - \mathbf{p}^{(\infty)}\|$  versus  $n$ , and use logs to investigate whether convergence is linear, logarithmic, exponential, power law, or none of these. In fact one of your graphs should give a linear relationship. Estimate the slope of the fitted line. Is this number a reasonable measure of convergence rate?

46. Does the relationship between observed frequencies after  $n$  steps and the limiting distribution for a graph walk follow a familiar pattern? Make a guess (i.e., think about what you expect here) before you gather evidence. Now use your graph walk from (45). Let  $\hat{\mathbf{p}}_{(n)}$  be the vector of observed frequencies after  $n$  steps.

$$\hat{p}_{(n)}[v] = \#\{t : X_t = v\} / n$$

$$\hat{\mathbf{p}}_{(n)} = (\hat{p}_{(n)}[v_1], \dots, \hat{p}_{(n)}[v_k])$$

Remind your self that because observed frequencies are computed from random outcomes, patterns may be harder to see than they were in (45). Use S-plus to simulate a walk of 10000 steps, and use the resulting path to compute  $\hat{\mathbf{p}}_{(n)}$  and the distance  $\|\hat{\mathbf{p}}_{(n)} - \mathbf{p}^{(\infty)}\|$  for  $n = 10, 20, 40, 100, 200, 400, 1000, 2000, 4000,$  and 10000. Graph distance versus  $n$ , using logs to investigate the functional form of convergence. What do you conclude?

## Discussion

47. Compare your results in (45) and (46) with your predictions. What is your current best explanation for the fact that the two kinds of convergence have different functional forms?

## Exercises

48. Every graph walk has a transition matrix, but not every probability matrix is the transition matrix for a graph walk. What conditions must the elements of a probability matrix satisfy in order for it to be the transition matrix of a graph walk?

49. Which graph walks have more than one stationary distribution, and which graph walks have just one? If a graph walk has more than one stationary distribution, can it have more than one limiting distribution? If so, what is it that determines which limiting distribution the walk converges to? If a graph has just one stationary distribution, can it have more than one limiting distribution?

50. *The game of Monopoly.*

The moves of each player in the board game Monopoly can be regarded as an instance of a Markov Chain.<sup>9</sup>

- a. Explain how to find transition probabilities.
- b. Explain why the limiting probabilities aren't equal.
- c. Explain why the value of a monopoly is related to the sum of its limiting probabilities.
- d. Explain why the fact in part (c) makes the orange properties especially valuable.
- e. According to the rules, if you go to jail, and choose not to pay, you can roll the two dice, and if you get doubles, you are out. Otherwise you stay in jail for that turn. Explain how to model this by incorporating a self-loop. Are there any other self-loops in the game?
- f. Although the moves in Monopoly define a Markov chain, the chain is not a random walk on a graph. Identify a feature of the game that prevents the moves on the board from being a graph walk.

51. **Investigation.** Create various connected graphs with 6 vertices. Use S-plus to compute powers of the transition matrix, and use the fitted slope, as described in (45), as a measure of convergence rate for  $n$ -step transition probabilities. Draw the connected 6-point graph for which you expect convergence to be fastest. Draw another for which you expect convergence to be as slow as possible. Create other 6-point graphs, trying to get as many different convergence rates as possible. Relate convergence rates to various quantitative features of graphs, as in the activity. Which features are the best predictors of mixing rate?

---

<sup>9</sup> To make this connection easier to think about, for the time being ignore Chance and Community Chest. If you get serious about this, you can add them to the model later on. For more on using a Markov chain model to study Monopoly strategy, see xxx.