MATH3200: APPLIED LINEAR ALGEBRA
PRACTICE MODULE 15: MORE APPLICATIONS OF MARKOV CHAINS: WALDO'S SURFING PATTERN

WINFRIED JUST, OHIO UNIVERSITY

We will use the terminology and notation of Conversation 10 and some results from Module 6.

Let us consider Waldo's strange surfing pattern of the WWW. Denny had us told the following story:

First Waldo opened his home page and followed a randomly chosen link. Then at each page that he visited:

- If the page had no link to another page, he teleported to a randomly chosen URL.
- If the page had links, he rolled a fair die.
  - If 6 came up, he teleported to a randomly chosen URL.
  - If any other number came up, he followed a randomly chosen link from the current page.

To get a better idea of how Waldo operated, let us revisit the last example from Module 6. Take another look at the mock-up of a class website `http://people.ohio.edu/just/M3200S18Mock/`.
This website has five pages that we represented by nodes that were numbered as follows:
1 = class home, 2 = syllabus, 3 = homework, 4 = supplements, 5 = sample solutions.
We then constructed a digraph by drawing a *separate* arc from node i to node j for each link on page i that points to page j. In this digraph there are multiple arcs with the same source and target. For example, there are many links from the homework page (node 3) to the supplements (node 4).
In Question 6.5 of Module 6 we found the adjacency matrix $\mathbf{A}$ of this directed graph:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 5 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Now let us think of Waldo surfing only the small part that is accessible at this website instead of the entire WWW. Then we can model his surfing pattern by a Markov chain whose states will be the five pages described above, represented by numbers $i = 1, 2, 3, 4, 5$ and a time step would correspond to the time he stays on a page before clicking on a link or "teleporting." The transition probability matrix will then be a 5×5 matrix:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} \end{bmatrix}$$

First assume that Waldo is at the page of sample solutions, page 5. This does not have any links, so Waldo will perform a step of "teleporting" and go to a randomly chosen page. Here

we assume that this would be a randomly chosen page amount the five that comprise our mock website; in the larger story it would be any URL. "Any" would also include the page he is currently at. So in our toy example he has 5 choices, and this transition will take him to any of our pages j with probability $p_{5j} = \frac{1}{5}$. We can now fill in the last row of the transition probability matrix as follows:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

Next suppose Waldo is at the supplements page 4. According to the adjacency matrix $\mathbf{A}$, page 4 has one link each to pages 1 and 5. In particular, Waldo could now move the homework page 3 only by teleporting. He will teleport only when the die comes up 6, with probability $\frac{1}{6}$, and if that happens, then land on any of the five pages with probability $\frac{1}{6}\frac{1}{5} = \frac{1}{30}$. This gives $p_{43} = \frac{1}{30}$, and similarly $p_{42} = p_{44} = \frac{1}{30}$:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ p_{41} & 1/30 & 1/30 & 1/30 & p_{45} \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

In contrast, Waldo could move from page 4 to page 1 in one of two ways: By teleporting, with probability $\frac{1}{30}$, or by clicking on the one of the 2 links at page 4 that points to page 1. The latter will happen when the die does *not* come up 6 *and* Waldo clicks on the relevant link, which will happen together with probability $\frac{5}{6}\frac{1}{2} = \frac{5}{12}$. Thus $p_{41} = \frac{1}{30} + \frac{5}{12}$. The same argument shows that $p_{45} = \frac{1}{30} + \frac{5}{12}$.

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} \\ 1/30 + 5/12 & 1/30 & 1/30 & 1/30 & 1/30 + 5/12 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

If Waldo is at the homework page 2, he can reach pages 2, 3, and 5 only by teleporting, so that $p_{32} = p_{33} = p_{35} = \frac{1}{30}$, and pages 1 and 4 by clicking on links. But now 5 of the 6 links on page 3 point to the supplements page 4, while only 1 of these 6 links points to page 1, the class home page. Thus *if* Waldo clicks on a random link rather than doing a teleporting move, then Waldo will reach page 1 with probability $\frac{1}{6}$ and page 4 with probability $\frac{5}{6}$. When we multiply these probabilities with the probability $\frac{5}{6}$ of clicking on a link and add the probabilities $\frac{1}{30}$ of teleporting to the same page, we obtain $p_{31} = \frac{1}{30} + \frac{5}{6}\frac{1}{6}$ and $p_{34} = \frac{1}{30} + \frac{5}{6}\frac{5}{6}$. Now we can fill in the third row of the matrix of transition probabilities:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} \\ 1/30 + 5/36 & 1/30 & 1/30 & 1/30 + 25/36 & 1/30 \\ 1/30 + 5/12 & 1/30 & 1/30 & 1/30 & 1/30 + 5/12 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \end{bmatrix}$$

Question 15.1: Complete the transition probability matrix $\mathbf{P}$ for this new Markov chain by filling in the first two rows.

Now we ask ourselves: **Where would we expect Waldo to be after 50 time steps?**

In analogy with our previous examples, we could calculate the probability distribution after 50 time steps as $\vec{\mathbf{x}}(50) = \vec{\mathbf{x}}(0)\mathbf{P}^{50}$. But for that we need the initial distribution $\vec{\mathbf{x}}(0)$.

Question 15.2: What is the initial $\vec{\mathbf{x}}(0)$ if we assume that Waldo always starts at the class home page 1?

For the initial distribution $\vec{\mathbf{x}}(0)$ that you found, MatLab will give us:
$\vec{\mathbf{x}}(50) = \vec{\mathbf{x}}(0)\mathbf{P}^{50} = [0.2350, 0.1296, 0.1567, 0.2925, 0.1862]$.

Now let us try to make some sense of this result. One can think of $\vec{\mathbf{x}}(50)$ as representing a kind of "popularity" measure for web pages at our mock website, with higher probabilities indicating that the page tends to be visited more often. This is similar to what we did in our first version of Waldo's story, with the owner of room 4, who is most popular in terms of having the most friends being visited by Waldo most often. Here the most popular page is the supplements page 4, which has the largest indegree in the directed graph that represents this website, that is, the largest total number of links pointing to it. The second most popular page is the class home page 1, which has the largest number of pages on which there are links to it. You can see this by looking at the columns of the adjacency matrix. Interestingly, the third highest popularity rank goes to page 5, which has only a single link on a single page pointing to it.

Question 15.3: Why then is page 5 still more popular than page 3 that has two links pointing to it?

Now think of how one would rank all pages on the WWW by popularity. One could consider a similar Markov chain as we did for our mock class website, but for the entire WWW. Then one could start with some initial page, follow random links as Waldo did, for many time steps, and rank pages in decreasing order by their probabilities in the resulting distribution.
Of course, one would need a kind of directory of the entire WWW, so that one could "teleport." And determining the matrix $\mathbf{P}$ of transition probabilities would be rather more difficult than in our example. However, a rich company could produce both such a directory and transition matrix. If that company then also kept lists of words and phrases that appear on the pages in their directory, it could send you upon request ranked lists of the most popular web pages that contain a given word or phrase.

Question 15.4: Which company are we talking about here? And who is Waldo?