7.3 Stationary distribution

Unit 7: Markov Chains

Adapted from Blitzstein-Hwang Chapter 11.

The concepts of recurrence and transience are important for understanding the long-run behavior of a Markov chain. At first, the chain may spend time in transient states. Eventually though, the chain will spend all its time in recurrent states. But what fraction of the time will it spend in each of the recurrent states? This question is answered by the *stationary distribution* of the chain, also known as the *steady-state distribution*. We will learn in this section that for irreducible and aperiodic Markov chains, the stationary distribution describes the long-run behavior of the chain, regardless of its initial conditions.

> **DEFINITION 7.3.1 (STATIONARY DISTRIBUTION).**
>
> A row vector $\mathbf{s} = (s_1, \ldots, s_M)$ such that $s_i \geq 0$ and $\sum_i s_i = 1$ is a *stationary distribution* for a Markov chain with transition matrix Q if
>
> $$\sum_i s_i q_{ij} = s_j$$
>
> for all j, or equivalently,
>
> $$\mathbf{s}Q = \mathbf{s}.$$

Recall that if $\mathbf{s}$ is the distribution of X_0, then $\mathbf{s}Q$ is the marginal distribution of X_1. Thus the equation $\mathbf{s}Q = \mathbf{s}$ means that if X_0 has distribution $\mathbf{s}$, then X_1 also has distribution $\mathbf{s}$. But then X_2 also has distribution $\mathbf{s}$, as does X_3, etc. That is, a Markov chain whose initial distribution is the stationary distribution $\mathbf{s}$ will stay in the stationary distribution forever.

One way to visualize the stationary distribution of a Markov chain is to imagine a large number of particles, each independently bouncing from state to state according to the transition probabilities. After a while, the system of particles will approach an equilibrium where, at each time period, the number of particles leaving a state will be counterbalanced by the number of particles entering that state, and this will be true for all states. As a result, the system as a whole will appear to be *stationary*, and the proportion of particles in each state will be given by the stationary distribution.

For very small Markov chains, we may solve for the stationary distribution by hand, using the definition. The next example illustrates this for a two-state chain.

Example 7.3.2 (Stationary distribution for a two-state chain).

Let

$$Q = \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix}.$$

The stationary distribution is of the form $\mathbf{s} = (s, 1-s)$, and we must solve for s in the system

$$\begin{pmatrix} s & 1-s \end{pmatrix} \begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix} = \begin{pmatrix} s & 1-s \end{pmatrix},$$

which is equivalent to

$$\frac{1}{3}s + \frac{1}{2}(1-s) = s,$$
$$\frac{2}{3}s + \frac{1}{2}(1-s) = 1-s.$$

The only solution is $s = 3/7$, so $(3/7, 4/7)$ is the unique stationary distribution of the Markov chain.

Existence and uniqueness

Does a stationary distribution always exist, and is it unique? It turns out that for a finite state space, a stationary distribution always exists. Furthermore, in irreducible Markov chains, the stationary distribution is unique.

> THEOREM 7.3.3 (EXISTENCE AND UNIQUENESS OF STATIONARY DISTRIBUTION).
>
> Any irreducible Markov chain has a unique stationary distribution. In this distribution, every state has positive probability.

The theorem is a consequence of a result from linear algebra called the *Perron-Frobenius theorem*.

Convergence

We have already informally stated that the stationary distribution describes the long-run behavior of the chain, in the sense that if we run the chain for a long time, the marginal distribution of X_n converges to the stationary distribution $\mathbf{s}$. The next theorem states that this is true as long as the chain is both irreducible and aperiodic. Then, regardless of the chain's initial conditions, the PMF of X_n will converge to the stationary distribution as $n \to \infty$. This relates the concept of stationarity to the long-run behavior of a Markov chain. The proof is omitted.

Theorem 7.3.4 (Convergence to stationary distribution).

Let $X_0, X_1, \ldots$ be a Markov chain with stationary distribution $\mathbf{s}$ and transition matrix Q, such that some power Q^m is positive in all entries. (These assumptions are equivalent to assuming that the chain is irreducible and aperiodic.) Then $P(X_n = i)$ converges to s_i as $n \to \infty$. In terms of the transition matrix, Q^n converges to a matrix in which each row is $\mathbf{s}$.

Therefore, after a large number of steps, the probability that the chain is in state i is close to the stationary probability s_i, regardless of the chain's initial conditions. Intuitively, the extra condition of aperiodicity is needed in order to rule out chains that just go around in circles, such as the chain in the following example.

Lastly, the stationary distribution tells us the average time between visits to a state.

Theorem 7.3.5 (Expected time to return).

Let $X_0, X_1, \ldots$ be an irreducible Markov chain with stationary distribution $\mathbf{s}$. Let r_i be the expected time it takes the chain to return to i, given that it starts at i. Then $s_i = 1/r_i$.

Here is how the theorems apply to the two-state chain from Example 7.3.3.

Example 7.3.6 (Long-run behavior of a two-state chain).

In the long run, the chain in Example 7.1.3 will spend 3/7 of its time in state 1 and 4/7 of its time in state 2. Starting at state 1, it will take an average of 7/3 steps to return to state 1. The powers of the transition matrix converge to a matrix where each row is the stationary distribution:

$$\begin{pmatrix} 1/3 & 2/3 \\ 1/2 & 1/2 \end{pmatrix}^n \to \begin{pmatrix} 3/7 & 4/7 \\ 3/7 & 4/7 \end{pmatrix} \text{ as } n \to \infty.$$

Google PageRank

We next consider a *vastly* larger example of a stationary distribution, for a Markov chain on a state space with billions of interconnected nodes: the World Wide Web. The next example explains how the founders of Google modeled web-surfing as a Markov chain, and then used its stationary distribution to rank the relevance of webpages. For years Google described the resulting method, known as *PageRank*, as "the heart of our software".

Suppose you are interested in a certain topic, say chess, so you use a search engine to look for useful webpages with information about chess. There are millions of webpages that mention the word "chess", so a key issue a search engine needs to deal with is what order to show the search results in. It would be a disaster to have to wade through thousands of garbage pages that mention "chess" before finding informative content.

In the early days of the web, search engines tended to ignore the *structure* of the web: which pages link to which other pages? Taking the link structure into account led to dramatic improvements in search engines. As a first attempt, one could rank a page based on how many other pages link to it. That is, if Page A links to Page B, we consider it a ''vote" for B, and we rank pages based on how many votes they have.

But this is again very open to abuse: a spam page could boost its ranking by creating thousands of other spam pages linking to it. And though it may seem democratic for each page to have equal voting power, an incoming link from a reliable page is more meaningful than a link from an uninformative page. Google PageRank, which was introduced in 1998 by Sergey Brin and the aptly named Larry Page, ranks the importance of a page not only by how many pages link to it, but also by the importance of those pages.

Consider the web as a directed network---which is what it is. Each page on the web is a node, and links between nodes represent links between pages. For example, suppose for simplicity that the web only has 4 pages, connected as shown in Figure 7.3.7.

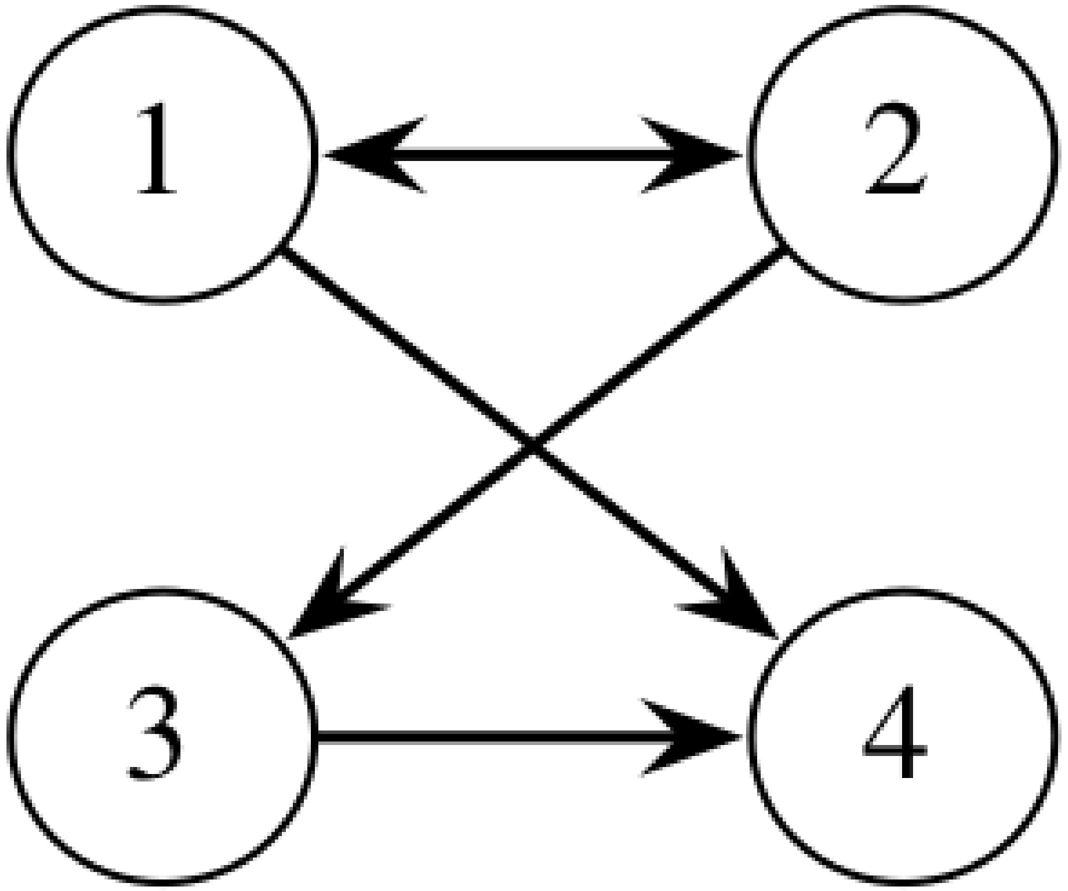

Figure 7.3.7

View Larger Image
Image Description

Imagine someone randomly surfing the web, starting at some page and then randomly clicking links to go from one page to the next (with equal probabilities for all links on the current page). The idea of PageRank is to measure the importance of a page by the long-run fraction of time spent at that page.

Of course, some pages may have no outgoing links at all, such as page 4 above. When the web surfer encounters such a page, rather than despairing he or she opens up a new browser window and visits a uniformly random page. Thus a page with no links is converted to a page that links to *every* page, including itself. For the example above, the resulting transition matrix is

$$Q = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 1 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}.$$

In general, let M be the number of pages on the web, let Q be the M by M transition matrix of the chain described above, and let $\mathbf{s}$ be the stationary distribution (assuming it exists and is unique). Think of s_j as a measure of how important Page j is. Intuitively, the equation

$$s_j = \sum_i s_i q_{ij}$$

says that the score of page j should be based not only on how many other pages link to it, but on their scores. Furthermore, the "voting power" of a page gets diluted if it has a lot of outgoing links: it counts for more if page i's only link is to page j (so that $q_{ij} = 1$) than if page i has thousands of links, one of which happens to be to page j.

It is not clear that a unique stationary distribution exists for this chain, since it may not be irreducible and aperiodic. Even if it is irreducible and aperiodic, convergence to the stationary distribution could be very slow since the web is so immense. To address these issues, suppose that before each move, the web surfer flips a coin with probability α of Heads. If Heads, the web surfer clicks a random link from the current page; if Tails, the web surfer *teleports* to a uniformly random page. The resulting chain has the transition matrix

$$G = \alpha Q + (1-\alpha)\frac{J}{M},$$

where J is the M by M matrix of all 1's. Note that the row sums of G are 1 and that all entries are positive, so G is a valid transition matrix for an irreducible, aperiodic Markov chain. This means there is a unique stationary distribution $\mathbf{s}$, called *PageRank*, and the chain will converge to it! The choice of α is an important consideration; choosing α close to 1 makes sense to respect the structure of the web as much as possible, but there is a tradeoff since it turns out that smaller values of α make the chain converge much faster. As a compromise, the original recommendation of Brin and Page was $\alpha = 0.85$.

PageRank is conceptually nice, but *computing* it sounds extremely difficult, considering that $\mathbf{s}G = \mathbf{s}$ could be a system of 100 billion equations in 100 billion unknowns. Instead of thinking of this as a massive algebra problem, we can use the Markov chain interpretation: for any starting distribution $\mathbf{t}$, $\mathbf{t}G^n \to \mathbf{s}$ as $n \to \infty$. And $\mathbf{t}G$ is easier to compute than it might seem at first:

$$\mathbf{t}G = \alpha(\mathbf{t}Q) + \frac{1-\alpha}{M}(\mathbf{t}J),$$

where computing the first term isn't too hard since Q is very *sparse* (mostly 0's) and computing the second term is easy since $\mathbf{t}J$ is a vector of all 1's. Then $\mathbf{t}G$ becomes the new $\mathbf{t}$, and we can compute $\mathbf{t}G^2 = (\mathbf{t}G)G$, etc., until the sequence appears to have converged (though it is hard to *know* that it has converged). This gives an approximation to PageRank, and has an intuitive interpretation as the distribution of where the web surfer is after a large number of steps.