Chapter 15

Ergodic theorem

The ergodic theorem ties together the notions of averaging over time and averaging over a distribution. If we consider the example of the shoppers at the shops S and T. Suppose we believe that the process is described by a Markov chain and we wish to determine the equilibrium state (that is the market share). Then there are at least two ways to do this. We sample say 200 shoppers on any given week-end and ask them where they shop. We use the sample data (experiment) to estimate the parameters (theory) for the market share (that is π_S and π_T).

Now an alternative method would be to choose 20 shoppers and follow these individuals over a period of 10 weeks, which again give us 200 data points. In an extreme case we could follow one person over a period of 200 weeks. The thrust of the ergodic theorem is that allowing the shoppers to *explore* the states of the system over time is equivalent to taking a snap-shot of a large group at a specific time. Of course, this assume that the system is governed by a Markov chain.

15.1 The theorem

The law of large numbers for irreducible Markov chains was discussed in the previous chapter. This idea can be extended one step further to arrive at the *ergodic theorem* :

For an irreducible Markov chain:

$$\lim_{n\to\infty} s_j(n) = \pi_j = \frac{1}{\mu_j} \tag{15.1}$$

where we define,

$$\mu_j \equiv \mathbb{E}\left(T_j\right) \tag{15.2}$$

as the *expected return time* for state j: the expected number of steps between two consecutive visits to j. This is also called the *recurrence time.*

The first half of the ergodic theorem was proven in the previous chapter, and was called the law of large numbers in that context. Let us now focus on the second half which deals with recurrence time.

PROOF:

First of all, suppose we start in some state, let's call it j, and then make an n-step Markov chain process. Then there will be N occurrences of j in the n-step process $N > 1$. This means there are $N - 1$ *returns* to state j, on or before the n-step.

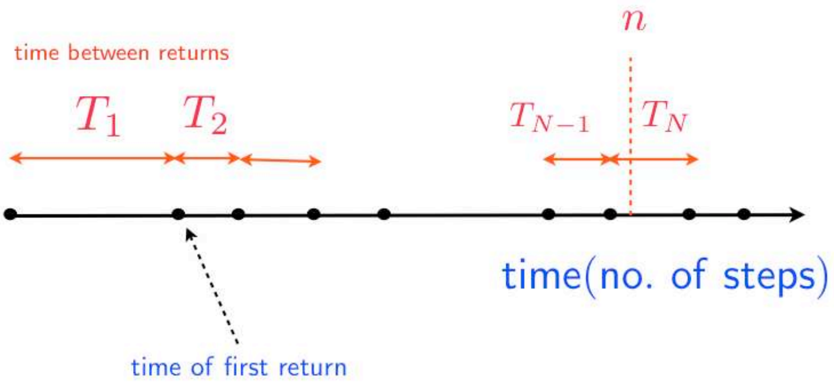

Figure 15.1: Return times, $T_1, T_2 \cdots T_{N-1}$, for an ergodic Markov chain. The chain runs for n time-steps with $N-1$ returns in this time, and a further returns in the future.

Let T be the discrete random variable that defines the number of steps between visits to the starting point j. We call these *return times*. So T_1 is the number of steps taken for the first return. Of course this is a random variable; one cannot predict exactly how many steps this random process will take. Then denoting all the return times as $T_1, T_2, \ldots, T_{N-1}$. Clearly it might take some time, after the start of the sequence to make the first visit to j. Then clearly:

$$\sum_{i}^{N-1} T_i \leq n \qquad . \tag{15.3}$$

The process is illustrated in figure (15.1).

Beyond our sample sequence of n steps, there will be another visit to j at some unknown time in the future, the Nth return:

$$\sum_{i}^{N} T_i > n \qquad . \tag{15.4}$$

This gives us a *squeezing* inequality, since we can write:

$$\frac{1}{N}\left(T_1 + T_2 + T_3 + \cdots + T_{N-1}\right) \leq \frac{n}{N} \leq \frac{1}{N}\left(T_1 + T_2 + T_3 + \cdots + T_N\right) \tag{15.5}$$

The left- and right-hand-sides are in the form of sample means. We can now apply the law of large numbers (in the strong form), taking $n, N \to \infty$ gives us:

$$\lim_{N\to\infty} \frac{1}{N}\left(T_1 + T_2 + T_3 + \cdots + T_N\right) = \mathbb{E}\left(T\right) = \mu \qquad . \tag{15.6}$$

Similarly;

$$\frac{1}{N}\left(T_1 + T_2 + T_3 + \cdots + T_{N-1}\right) = \left(\frac{N-1}{N}\right)\frac{1}{N-1}\left(T_1 + T_2 + T_3 + \cdots + T_{N-1}\right) \tag{15.7}$$

Applying this to the inequality (15.5) gives the squeezing effect desired:

$$\lim_{N\to\infty}\frac{N-1}{N}\lim_{N\to\infty}\frac{1}{N-1}\left(T_1 + T_2 + T_3 + \cdots + T_{N-1}\right) \leq \lim_{N,n\to\infty}\frac{n}{N} \leq \lim_{N\to\infty}\frac{1}{N}\left(T_1 + T_2 + T_3 + \cdots + T_N\right) \tag{15.8}$$

this gives us,

$$\lim_{N\to\infty}\left(\frac{N-1}{N}\right)\mu \leq \lim_{N\to\infty}\frac{n}{N} \leq \mu \qquad . \tag{15.9}$$

So we arrive at the conclusion, with certainty, in the long run:

$$\lim_{N\to\infty}\frac{n}{N} = \mu \tag{15.10}$$

Although we started the discussion for some state j, the choice of state was arbitrary, so the same argument applies to any (and all) states. That is, reinserting the subscripts, and flipping this upside down, we can write

$$\lim_{n\to\infty}\frac{N_j(n)}{n} = \frac{1}{\mu_j} \tag{15.11}$$

for all j, as required.

There is nothing surprising in the ergodic theorem, but it is still a profound result. Like most good theorems it makes perfect sense. It states that if we have a Markov chain that is irreducible and aperiodic (which means that all the states communicate with each other and they do so in a random aperiodic manner) then over time, state j say, will be visited a certain fraction of times f, and in the long run (law of large numbers) this fraction tends to a limit. Then, over a time n we would expect the state to visit (return) fn times to j and then the time between visits (on average) would be the total time n divided by fn. That is the mean return time would be $n/(nf) = 1/f$. Similarly, if we were to choose a random point in the sequence, then the probability that, at this point, the system would be in state j, would be just equal to the fraction of time spent in j , that is the fraction of occurrences of j, namely f.

Just a word about the requirement for aperiodicity. This ensures the states are 'scrambled' and that the signature of the initial conditions is lost. That is, there is no correlations is time between states. And a word of warning, before going on to some applications. Ergodic Markov chains are perfect mathematical entities, but very few *real-world* systems have such ideal properties. So, most applications in which Markov chains are invoked are not Markovian at all. However, non-Markovianity of the real world is much more interesting.

15.2 Applications

Returning to the two-state chain in the previous chapter, the ergodic theorem addresses the expected number of steps until the state repeats. For example we can estimate how many steps between the 1 state repeating.

According to the ergodic theorem:

$$\mu_1 = \frac{1}{\pi_1} = \tfrac{3}{2} \qquad .$$

That is, given that a 1 occurs in a sequence, the expected number of steps until the next 1 occurs is 1.5. For the state 0 we have:

$$\mu_0 = \frac{1}{\pi_0} = 3 \qquad .$$

which is twice as long. Again, these results can be tested (experimentally) using Monte Carlo simulation.

A second example involves tossing a coin. The state of the system (HEADS or TAILS) we will denotes as 1 or 0, respectively. Suppose that we toss the coin many times and record the sequence to give us a Markov chain. If $0 \leq p \leq 1$ is the probability of HEADS (1) and $q = 1 - p$ is the probability of TAILS

(0), then:

$$\mathsf{P} = \begin{pmatrix} q & p \\ q & p \end{pmatrix} \tag{15.12}$$

Let us test the ergodic theorem in this context since we know a great deal about coin sequences.

For example we know the following that, given $X_0 = 1$ the probability of the next HEADS occurring after i tosses is:

$$P(T = i) = q^{i-1}p \qquad i = 1, 2, 3 \ldots \qquad . \tag{15.13}$$

where T is the numbers of tosses, or equivalently the time until the next HEADS occurs.

But this is just the geometric distribution, and it is simple to calculate $\mathbb{E}(T)$, which will just be μ_1, the recurrence time for $X = 1$. We have, the well-known result (see earlier lectures):

$$\mu_1 = \mathbb{E}(T) = \sum_{i=1}^{\infty} i q^{i-1} p = \frac{1}{p} \tag{15.14}$$

as the mean of a geometric distribution.

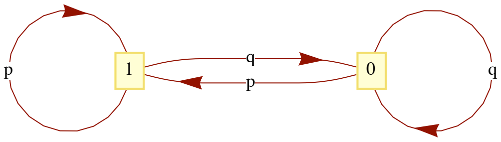

Figure 15.2: Transition graph for a coin toss. The state of the coin is $X \in \{0, 1\}$. The state HEADS corresponds to $X = 1$, and the probability of HEADS on any toss is p.

Let us compare this result with the prediction of the ergodic theorem. For a Markov chain with the given transition matrix, the equilibrium distribution is given by:

$$(\pi_0 \qquad \pi_1) = (\pi_0 \qquad \pi_1) \begin{pmatrix} q & p \\ q & p \end{pmatrix} \tag{15.15}$$

with the solution:

$$\pi_0 = q \qquad \pi_1 = p \qquad . \tag{15.16}$$

Then, according to the ergodic theorem we should have a recurrence time for state 1 given by:

$$\mu_1 = \frac{1}{\pi_1} = \frac{1}{p} \qquad . \tag{15.17}$$

This is in complete agreement with the result derived independently on the basis of the geometric distribution: equation (15.14).

15.3 Monte Carlo simulation

We complete this chapter with a test case in which all the principles of the ergodic theorem are applied. Consider a 4-state system with the transition matrix:

$$\mathsf{P} = \begin{pmatrix} 0.1 & 0.2 & 0.0 & 0.7 \\ 0.0 & 0.2 & 0.4 & 0.4 \\ 0.4 & 0.3 & 0.0 & 0.3 \\ 0.3 & 0.1 & 0.3 & 0.3 \end{pmatrix} \quad . \tag{15.18}$$

The transition graph for this chain is displayed in figure 15.3

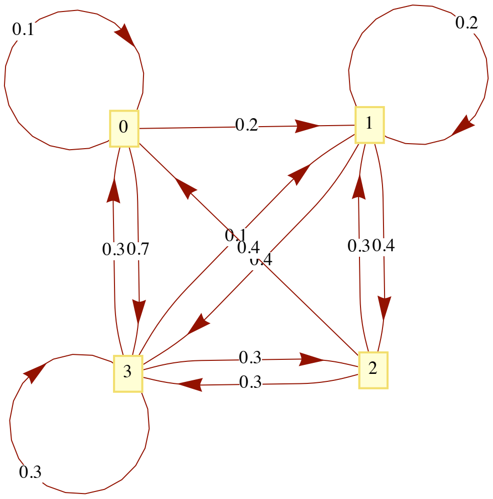

Figure 15.3: Transition graph for the 4-state Markov chain.

Firstly we determine the equilibrium state of the system by numerical methods, using the formula (13.28) that:

$$\lim_{n\to\infty} \mathsf{P}^{\mathsf{n}} = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix} \quad . \tag{15.19}$$

Taking $n = 50$ we find:

$$\mathsf{P}^{50} \approx \begin{pmatrix} 0.2214 & 0.1787 & 0.1934 & 0.4064 \\ 0.2214 & 0.1787 & 0.1934 & 0.4064 \\ 0.2214 & 0.1787 & 0.1934 & 0.4064 \\ 0.2214 & 0.1787 & 0.1934 & 0.4064 \end{pmatrix} \tag{15.20}$$

So that:

$$\pi_0 = 0.2214 \qquad \pi_1 = 0.1787 \qquad \pi_2 = 0.1934 \qquad \pi_3 = 0.4064$$

and consequently the recurrence times are, approximately:

$$\mu_0 = 4.517 \qquad \mu_1 = 5.596 \qquad \mu_2 = 5.171 \qquad \mu_3 = 2.461$$

A Monte Carlo simulation of the process can be implemented to study the time-average of the processes. For a sample of $n = 10^4$ steps we find the fraction of time in each state to be:

$$s_0 = 0.2193 \qquad s_1 = 0.1840 \qquad s_2 = 0.1879 \qquad s_3 = 0.4088 \qquad .$$

These values are good approximations to the true equilibrium distribution. For a larger sample, $n = 10^5$, we find typical values:

$$s_0 = 0.2217 \qquad s_1 = 0.1799 \qquad s_2 = 0.1928 \qquad s_3 = 0.4056 \qquad .$$

which are in the neighbourhood of the equilibrium distribution values.

Let us now conclude with a brief look at the recurrence times. In this case the Monte Carlo simulation runs for $n = 10^5$ steps. The return times are recorded and a sample average is made. An example of the results are the return times:

$$\mu_1 \approx 5.5515 \qquad \text{and} \qquad \mu_3 \approx 5.1854$$

which again, tally with the theoretical return times through the law of large numbers expressed as the ergodic theorem.

15.4 Internet searches

Web pages are like *destinations* that one visits temporarily while surfing the internet. These web pages are connected to each other by *hyperlinks* embedded in the page by the authors. In this way we can view the web pages as being a connected network. This is not the physical network of optical fibres between sites that forms the internet, it is a virtual network created by the authors of the webpages. Suppose one arranges that a person or robot maps out all these connections.

Suppose information (perhaps in the form of *key words*) is distributed across the internet. The internet connects physical sites but also webpages through hyperlinks. Suppose we viewed this network as a kind of Markov chain transition diagram in which the hyperlinks play the role of edges. Then the connections can be stored in an *adjacency matrix*, X. For a network with N nodes A is an $N \times N$ matrix, and we indicate a link from node i to node j ($i \to j$) by a 1 in row i and column j. If there is no connection ($i \nrightarrow j$) then a zero is entered, that is:

$$A_{ij} = \begin{cases} 1 & , \quad i \to j \\ 0 & , \quad i \nrightarrow j \end{cases} \qquad 1 \leq i, j \leq N \qquad . \tag{15.21}$$

The ‘connectedness’ of each node can be measured by the number of links in and out. We define the *out-degree* as the total number of links out from the node. So the out-degree of node j is simply the number of ones in row j, that is:

$$k_j^{\mathsf{out}} \equiv \sum_{i=1}^{N} A_{ji} \qquad . \tag{15.22}$$

Similarly the *in-degree* of node j is the total number of ones in column k of the adjacency matrix, that is:

$$k_j^{\mathsf{in}} \equiv \sum_{i=1}^{N} A_{ij} \tag{15.23}$$

The question now arises, given N nodes with all these connections, which of these nodes is the most useful place to visit first ? That is, is there any method of capturing the collective knowledge of all the various webpage authors in order to rank the pages in order of importance so that they come high in our list ? For our web page network, a high in-degree indicates that a certain node/web page is viewed as important by the other nodes, since it has many referrals. On the other hand, a node with a high out-degree might also be a valuable place to visit since it references many other pages. The following section briefly outlines a technique that uses Markov chains to arrive at the solution of the problem.

15.4.1 Page Ranking algorithm

Suppose we view the network as a recurrent Markov chain, then one could view the most 'important' state as the one where most time is spent. That is, the node corresponding to the maximum in the equilibrium distribution. This is the basic idea of the PageRank ® algorithm that the founders of Google developed [1] to great commercial success.

So the only question is how to define the transition matrix for such a system. The simplest answer would be to close off the internet into a finite set of nodes. Then give non-zero transition probabilities to the nodes that are *linked.* and give these equal weighting, so for $k_i^{\mathsf{out}} \neq 0$ we take:

$$P_{ij}^{(1)} = \frac{A_{ij}}{\sum_k A_{ik}} = \frac{A_{ij}}{k_i^{\mathsf{out}}} \quad , \quad k_i^{\mathsf{out}} \neq 0 \quad . \tag{15.24}$$

Now many webpages are *dead-ends* (absorbing states). So in principle, once we hit those states we spend the rest of the time there. So these could be wrongly indicated as being important. So we artificially introduce transitions away from these states to ensure the chain in recurrent. So for $k_i^{\mathsf{out}} = 0$, that is for an absorbing state i, one with no outgoing connections, we decide to connect to *all* states, and with equal probability so that:

$$P_{ij}^{(1)} = \frac{1}{N} \quad , \quad k_i^{\mathsf{out}} = 0 \quad , \quad i = 1, 2, 3 \ldots, N \quad . \tag{15.25}$$

The denominator (normalising factor) ensures that the matrix is stochastic whatever the state i:

$$\sum_{j=1}^{N} P_{ij}^{(1)} = 1 \quad , \quad 1 \leq i \leq N \quad . \tag{15.26}$$

The choice (15.24) means that any disconnected nodes are ignored. One way of including visits to such pages (say from node i) even though they are not directly hyperlinked from i, would be to allow a small non-zero probability, which we call $d' \geq 0$, of access to these states. This redistributes the possible transitions to unlinked states equally and randomly, rather like (15.27). So a refinement of the model would be:

$$P_{ij}^{(2)} = \begin{cases} d\dfrac{A_{ij}}{k_i^{\text{out}}} + \dfrac{(1-d)}{N} & , \quad k_i^{\text{out}} \neq 0 \\ \dfrac{1}{N} & , \quad k_i^{\text{out}} = 0 \end{cases} \tag{15.27}$$

This ensures that we have an aperiodic irreducible Markov chain.

The value of $0 \leq d \leq 1$ is open to choice, and is called the *damping factor*. The larger the value of d, the more steps are taken away from states with high out degree. So as $d \to 1$, the states with most links are

[1] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine

less important, in terms of their ranking. Conversely, as $d \to 0$, then more importance is given to pages that have many links.

The equilibrium distribution, r, is the N-element row vector that is the solution of the equation:

$$\boxed{r = rP^{(2)}} \quad . \tag{15.28}$$

So the very basic version of the PageRank algorithm for ordering takes the node corresponding to the *largest component* of r and ranks this highest. The next largest component of r is placed second and so on assigning last place in ranking to the note/state with the smallest component of r.

We consider an example of a group of 6 web pages which are connected by hyperlinks as shown below (figure 15.4).

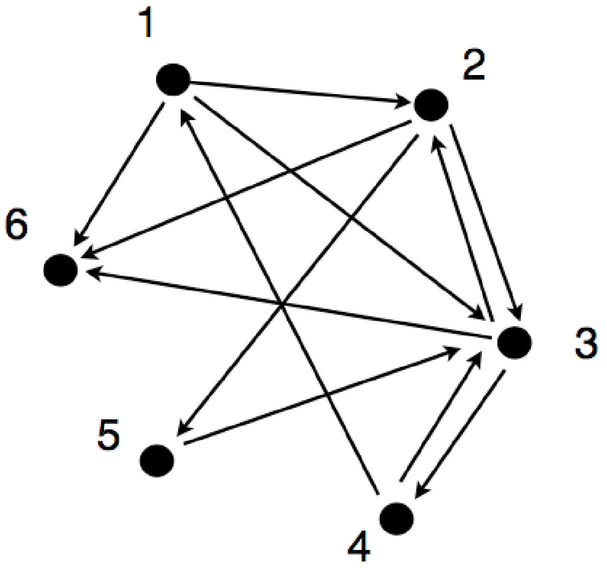

Figure 15.4: Links within a network of $N = 6$ nodes.

The ranking depends on the value of d as shown in the table below and in figure 15.5. Clearly for large d, that is when transitions strongly favour direct links, the highest ranked node is 3. This would be expected given the high in-degree and high out-degree that this node possesses. It is interesting to see that node 6, which appeared as a dead end (absorbing state) moves up to second place in the ranking. Thus a page which has many different links pointing towards it could be viewed as the 'final word' on the subject.

			Ranking	Ranking
Node label i	out − degree k_i^{out}	in − degree k_i^{in}	r_i for $d = 0.1$	r_i for $d = 0.9$
1	3	1	0.205	0.115
2	3	2	0.165	0.164
3	3	4	0.091	0.279
4	2	1	0.219	0.133
5	1	1	0.200	0.102
6	0	3	0.121	0.207

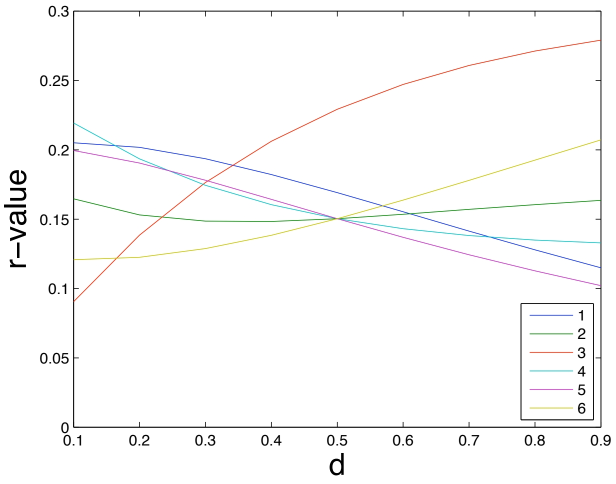

Figure 15.5: Ranking within the network shown in figure 15.4.

If we change the network connections slightly, as shown in figure 15.6, which is slightly sparser, and node 6 has a lower in-degree the ranking changes as shown in figure 15.7.

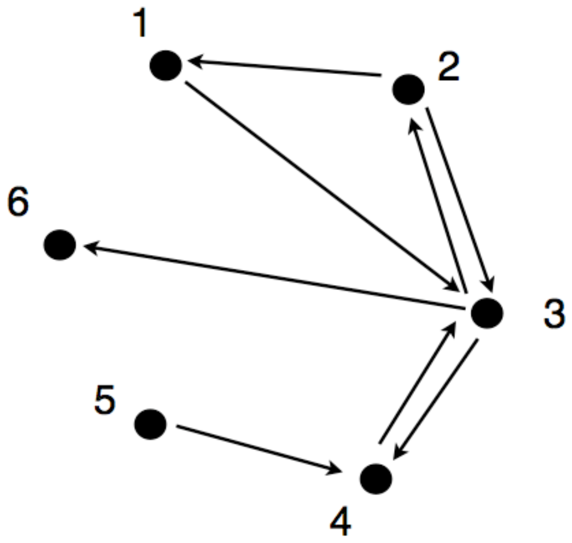

Figure 15.6: Links within a network of $N = 6$ nodes, with a slightly sparser structure compared to figure 15.4.

2 and 6 share the same ranking values for all values of d.

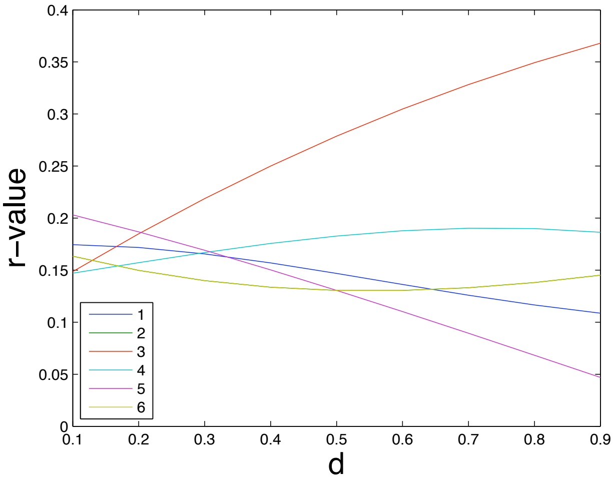

Figure 15.7: Ranking within a network of $N = 6$ nodes as shown in figure 15.6. Only 5 lines appear since nodes 2 and 6 have the *same* rankings for all values of d.