

Markov Chains for Self-Organizing Systems

Quentin F. Stout

Suppose we have a finite collection S of *states* that a system can be in, and suppose that at each time step, the system either stays in the same state or changes to a different state. Further, suppose that there is a *transition probability matrix* P , with rows and columns indexed by S , such that, if the system is in state $s \in S$ at one time step, then the probability that it is in state $t \in S$ in the next time step is $P(s, t)$, where the probability of going from s to t is independent of time and of the path traveled to reach s . Note that $P(s, s)$ is the probability of staying in s for one step, and that the probability of traversing any path is the product of the probabilities of each step. Such a system is a *Markov chain*, denoted $\mathcal{M}(S, P)$.

It is often useful to represent a Markov chain by a graph, where each state is a vertex and each nonzero transition probability is an edge labeled with the probability of that transition. See Figure 1.

In Figure 1, if the system started in state O at time 0, then at time 1 it is in A with probability 1, at time 2 it has 0.5 probability of being in state B and 0.5 probability of being in state D, at time 3 it has 0.5 probability of being in state C and 0.5 probability of being in state E, and at time 4 it will be back in A with probability 1. Thus this system cycles, once it has left O.

Many important Markov chains have a different behavior, which is not cyclic and which almost surely revisits every node infinitely often. For example, in Figure 2, if we use (a, b, c) to represent the probability of being in states A, B, C, respectively, then if the system starts in state A at time 0 (i.e., if the probabilities are $(1, 0, 0)$), then at time 1 the probabilities are $(0.5, 0.5, 0)$, at time 2 they are $(0.25, 0.25, 0.5)$, at time 3 they are $(0.625, 0.125, 0.25)$, at time 4 they are $(0.5625, 0.3125, 0.125)$, and so on. No matter where it starts, the system converges to the distribution $(0.5, 0.25, 0.25)$. Note that if you apply the transition matrix to this vector you get the same vector back, i.e., it is an eigenvector of eigenvalue 1. Further, it is the unique positive eigenvector of eigenvalue 1.

The existence of this eigenvector is guaranteed by the fact that the Markov chain in Figure 2 is strongly mixing.

A finite Markov chain $\mathcal{M}(S, P)$ is *strongly mixing* if there is a constant $d > 0$ such that, for every pair of (perhaps the same) states $s, t \in S$, the probability of starting in s and ending in t in exactly d steps is greater than 0, i.e., $P^d(s, t) > 0$.

Several other terms are used for the same concept.

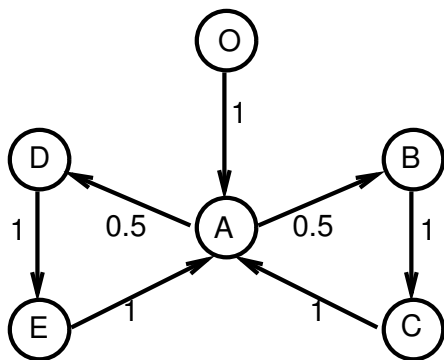


Figure 1: An absorbing, cyclic chain

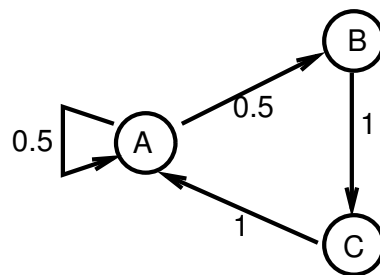


Figure 2: A strongly mixing chain

One important feature of strongly mixing chains is that they converge to a fixed probability distribution.

Theorem: A finite Markov chain $\mathcal{M}(S, P)$ is strongly mixing if and only if there is a probability distribution π on S , such that

- $\pi(s) > 0$ for all $s \in S$, and
- for any starting distribution ρ , for any $s \in S$, the probability that the chain is in state s at time τ converges to $\pi(s)$ as τ tends to infinity, i.e., $\lim_{\tau \rightarrow \infty} P^\tau(\rho) = \pi$.

□

Note that a Markov chain is strongly mixing if and only if the transition graph associated with \mathcal{M} has a simple property, namely, if there is a d such that, for every $s, t \in S$, there is a path from s to t of exactly d steps (where paths may reuse edges and may include edges from a state to itself). One common condition that implies this property is that the graph is strongly connected (i.e., for every $s, t \in S$ there is a path from s to t), and at least one vertex has a self-loop (an edge to itself). This condition holds in most self-organizing systems. Note that while this condition is sufficient to insure strong mixing, it is not necessary.

To see how this condition can be used, consider a self-organizing array of 3 items, and suppose that item i has positive probability $p(i)$ of being requested, $i \in \{1, 2, 3\}$. Assume that all requests are independent, and that the transposition heuristic (Trans) is being used. Then the states are just the 6 possible orderings of the items, and the Markov chain is as in Figure 3. By the above comments, this chain is strongly mixing (in fact, it is easy to see directly that one can go between any two states in exactly 3 steps), and hence has a probability distribution that it converges to.

This distribution is the unique positive eigenvector of eigenvalue 1 (i.e., a fixed point), which can be numerically found by linear algebra for any fixed values of $p(1)$, $p(2)$, and $p(3)$. However, by inspection we can see that, for the general case of n items using Trans, for any permutation ν and probability distribution π , the asymptotic probability $\mathcal{P}_{\text{Trans}}(\nu, \pi)$ that the system is in the state where the items are in order $\nu(1), \nu(2), \dots, \nu(n)$, is

$$\mathcal{P}_{\text{Trans}}(\nu, \pi) = \frac{\prod_{i=1}^n p(\nu(i))^{n-i}}{\sum \{ \prod_{i=1}^n p(\mu(i))^{n-i} : \mu \text{ a permutation} \}}$$

(This was first noted by R. Rivest, in “On self-organizing sequential search heuristics”, *Comm. ACM* 19 (1976), pp. 63–67.) From now on we’ll just drop the π from the notation.

We can verify this is correct by noting first that it is indeed a probability distribution (each entry is positive, and the sum, over all states, is 1), and second that when we plug it into the transition probability matrix, we get the same thing back again. Namely, at any state (permutation) ν , we should have

$$\begin{aligned} \mathcal{P}_{\text{Trans}}(\nu) &= \sum P(\mu, \nu) \cdot \mathcal{P}_{\text{Trans}}(\mu) \\ &= p(\nu(1)) \cdot \mathcal{P}_{\text{Trans}}(\nu) + \sum_{i=1}^{n-1} p(\nu(i)) \cdot \mathcal{P}_{\text{Trans}}(\nu(1), \nu(2), \dots, \nu(i-1), \nu(i+1), \nu(i), \dots, \nu(n)) \\ &\quad \text{i.e., the only way to get to } \nu \text{ in a single step is to be in } \nu \\ &\quad \text{or to be in a state where switching adjacent items yields } \nu \\ &= p(\nu(1)) \cdot \mathcal{P}_{\text{Trans}}(\nu) + \sum_{i=1}^{n-1} p(\nu(i)) \cdot \mathcal{P}_{\text{Trans}}(\nu) \frac{p(\nu(i+1))}{p(\nu(i))} \end{aligned}$$

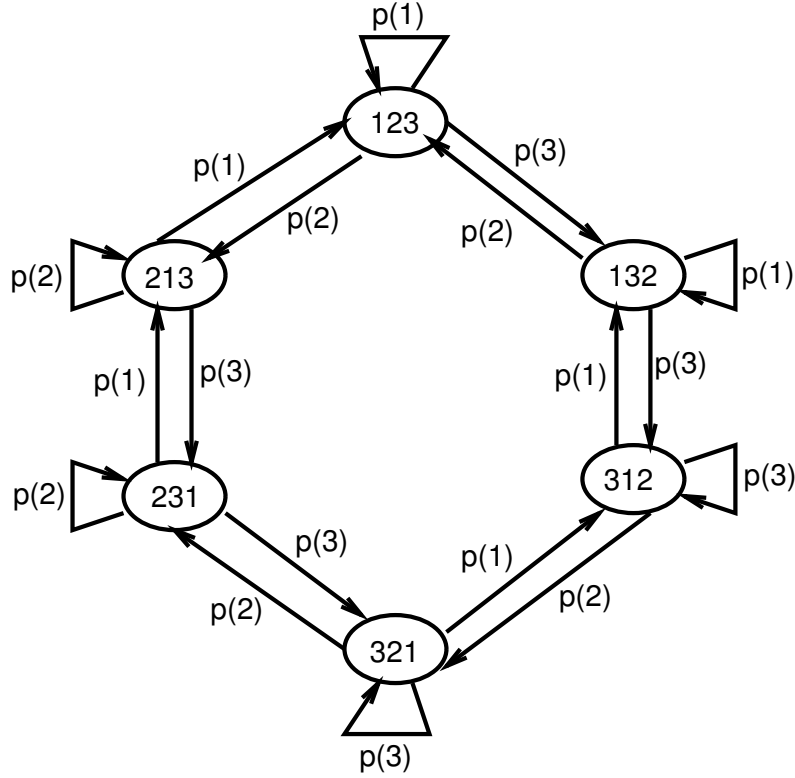


Figure 3: The Markov chain corresponding to Trans

$$\begin{aligned}
 &= \mathcal{P}_{\text{Trans}}(\nu) \cdot \sum_{i=1}^n p(\nu(i)) \\
 &= \mathcal{P}_{\text{Trans}}(\nu)
 \end{aligned}$$

In general, finding the eigenvector in such a symbolic form may be very difficult, but if one can guess the answer then it can be verified as above. Sometimes one can make an educated guess by noting similarities to problems with known solutions, and sometimes one can solve small problems and then extrapolate to make a guess about the solution to larger problems.

Note that one must show that the Markov chain is strongly mixing to insure that the system always converges to the fixed-point distribution. For example, in Figure 1, the probability distribution with mass 0 at O, mass 1/3 at A, and mass 1/6 at each of B, C, D, and E is a fixed point, but it is not true that all starting conditions converge to it.

Using similar techniques as for Trans, one can also see that the move-to-front heuristic (MTF) is strongly mixing, and that $\mathcal{P}_{\text{MTF}}(\nu)$, is given by

$$\mathcal{P}_{\text{MTF}}(\nu) = \prod_{i=1}^n \frac{p(\nu(i))}{1 - \sum_{j=1}^{i-1} p(\nu(j))}.$$

This can also be seen more directly, by noting that the asymptotic probability that the first item is $\nu(1)$ is the probability that the most recent request was for $\nu(1)$, i.e., the probability is $p(\nu(1))$. If the first item is

$\nu(1)$, the asymptotic probability that the second item is $\nu(2)$ is the probability that the most recent request for something other than $\nu(1)$ was for $\nu(2)$, i.e., the probability is $p(\nu(2))/[1 - p(\nu(1))]$. Continuing in this manner gives the formula for $\mathcal{P}_{\text{MTF}}(\nu)$.

Rivest used the formula for Trans to show that it is asymptotically better than MTF. More precisely, he showed:

Theorem: Given n items $1, \dots, n$ with associated request probabilities $p(1), \dots, p(n)$, if $n > 2$ and not all probabilities are the same then the asymptotic expected number of items examined by Trans is strictly less than the asymptotic expected number of items examined by MTF.

Proof: Let i and j be two items where $p(i) > p(j)$. Let ν be any permutation of the items where i precedes j , and let ν' be the permutation where i and j are interchanged but all other items are in the same position as in ν . If i is k positions ahead of j in ν , then

$$\frac{\mathcal{P}_{\text{Trans}}(\nu)}{\mathcal{P}_{\text{Trans}}(\nu')} = \left(\frac{p(i)}{p(j)}\right)^k$$

This can be seen by noting that the formulae for $\mathcal{P}_{\text{Trans}}(\nu)$ and $\mathcal{P}_{\text{Trans}}(\nu')$ differ only in the terms involving $p(i)$ and $p(j)$.

Since all the permutations can be paired up in such a fashion, in transposition, the asymptotic probability that i precedes j , divided by the asymptotic probability that j precedes i , is strictly greater than $p(i)/p(j)$. This is because $[p(i)/p(j)]^k > p(i)/p(j)$ for $k > 1$ and we have more than two items, so for some permutation pairs $k > 1$.

Meanwhile, for MTF, the ratio of the asymptotic probabilities is exactly $p(i)/p(j)$, since i precedes j if and only if the most recent request for either i or j was for i .

Thus whenever $p(i) > p(j)$ and $n > 2$, i is asymptotically more likely to precede j in Trans than in MTF. Since the asymptotic expected number of items examined is

$$1 + \sum_{\text{all pairs } (i,j)} p(i) \cdot (\text{prob. } j \text{ precedes } i) + p(j) \cdot (\text{prob. } i \text{ precedes } j)$$

we have shown that Trans is better. \square

Note that if $n = 2$ or all probabilities are equal, then Trans and MTF examine the same expected number of items.