Lecture 6: Markov chains

Rajat Mittal

IIT Kanpur

Up till now we have mostly tried to model a randomized experiment, sometimes looking at independent repetition of such an experiment. This chapter will look at *stochastic processes*, which can be thought of as a randomized process or a sequence of randomized experiments (might depend on previous outcomes).

We view a stochastic process as a sequence of random variables dependent on each other. It can also be thought of as a sequence of observations on a randomized process. This is a much more general setting and can model many more situations; amount of traffic on road everyday to movement of stock prices.

Mathematically, a stochastic process is a sequence $\{X_t\}_{t\in I}$ of random variables taking values in some common state space (say S), where t comes from some index set I. Notice a crucial difference, we defined random variables to be $X : \Omega \to \mathbb{R}$ before. In this case, the set of outputs are called the state space of these random variables. It is different from the sample space Ω and X_t can be thought of as a function from Ω to S. We will denote the elements of state space by letters $s_1, s_2, \cdots$.

The state space of random variables, S, could be a discrete or a continuous set. In almost all situations, we can think of t as time and index being some passage of time. Even the time index set I could be discrete (e.g., random experiment every day) or continuous (e.g., movement of stock prices).

We will mostly worry about the case when index set is discrete in this chapter, index coming from the set of natural numbers. The probability, in most general terms, could be written as,

$$P(X_1 = s_1, \cdots, X_k = s_k) = P(X_1 = s_1)P\left(X_2 = s_2 | X_1 = s_1\right) \cdots P\left(X_k = s_k | X_1 = s_1, \cdots, X_{k-1} = s_{k-1}\right).$$

We assumed that value of X_k only depends on the random variables *before* it. As mentioned above, this is the most general setting and very hard to deal with. The easiest version is the one studied before, sequence of independent random variables.

$$P(X_1 = s_1, X_2 = s_2, \cdots, X_k = s_k) = P(X_1 = s_1)P(X_2 = s_2) \cdots P(X_k = s_k).$$

Our focus in this chapter is on processes which are slightly more general than independent random variables.

1 Markov process

An independent random variable does not depend on anything in the past. Roghly, a *Markov process* (also called *Markov chain*) is a sequence where future random variables depend *only* on the present state, as opposed to being dependent on both present and past.

Formally, a sequence of random variables $\{X_t\}_{t\in\mathbb{N}}$ is said to satisfy the Markov property (or called a Markov chain) if,

$$P\left(X_k = s_k | X_1 = s_1, X_2 = s_2 \cdots, X_{k-1} = s_{k-1}\right) = P\left(X_k = s_k | X_{k-1} = s_{k-1}\right).$$

Exercise 1. Convince yourself that if X_i's are independent random variables, then trivially they are a Markov chain too.

Intuitively, the value of the last random variable is sufficient to determine the future (nothing about the past is needed). To be precise, it is not that the value of X_{k+1} does not depend upon $X_1, X_2, \cdots$; just that the the influence of $X_1, X_2, \cdots$ is absorbed by X_k. In other words, conditioned on X_k, random variables X_{k+1} and X_1 are independent. For example, let Y_t denote the maximum rainfall seen till month t. Assuming rainfall every month is independent, $\{Y_t\}$ is a Markov chain.

To take another example, suppose that the probability of India winning against Pakistan in a cricket match is .6. They are going to play a long series of matches and the probability of winning remains the same for every match. Suppose X_i counts the number of wins of India after i matches. You can check that $\{X_i\}_i$ is a Markov chain. Even the number of matches by which India leads Pakistan (wins of India - losses of India) is also a Markov chain.

We simplify the process more, a *homogeneous Markov chain* is one where the conditional dependences above do not depend on time, i.e.,

$$P\left(X_k = i | X_{k-1} = j\right) = P\left(X_1 = i | X_0 = j\right),$$

for all pair of elements i, j in state space S.

Exercise 2. Convince yourself that the number of wins of India in a cricket match (Markov chain defined above) is a homogeneous Markov chain. In general, if $\{X_i\}_{i=1}^n$ are a series of independent random variables, then $\{Y_i = \sum_{j=1}^i X_j\}_{i=1}^n$ is a Markov process (if X_i's are i.i.d.'s then a homogeneous Markov process).

For the rest of this chapter, a Markov chain would mean a homogeneous Markov chain. A Markov chain can be specified by just two set of values,

- Initial probability distribution μ, a $|S|$ length vector. Since it describes the probabilities in the beginning, all entries are positive and sum to 1.
- Transition matrix T ($|S| \times |S|$), describing the probabilities of moving between states. So,

$$T_{ij} = P\left(X_1 = j | X_0 = i\right) \quad \forall i, j \in S.$$

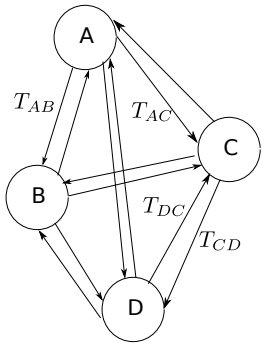

Fig. 1. A general homogeneous Markov chain.

Exercise 3. What can you say about the sum of the entries of a row of T? What about sum of a column?

One of the simplest example of a Markov chain is the drunkard's walk in one dimension. The state space is integers. Assume that the drunkard starts at 0. At any point of time, he can potentially move one step to right or one step to left with equal probability. So, the transition matrix is given by,

$$T_{ij} = \begin{cases} 1/2 & (j = i \pm 1), \\ 0 & \text{otherwise} \end{cases}$$

It is clear that such a walk satisfies the Markov property.

Exercise 4. What is the starting state μ?

The drunkard's walk can be generalized to more dimensions and then the drunkard can take step in more directions. Another generalization is the concept of *random walk* on a graph a G. In this case, the state space is the set of all vertices. At any time step, you can move to any neighbouring vertex with some probability (in many instances, it is uniform on all neighbours). Drunkard's walk can be thought of as a random walk on a line.

Exercise 5. Write the state space S and transition matrix for a random walk on a complete graph (move to any neighbour with equal probability).

The concept of random walk has been very helpful in complexity theory, for example, it has been used in designing randomized algorithms and pseudo-random generators.

In general, Markov chains are used to model many diverse situations; to describe evolution of DNA, motion of particles in physics (Brownian motion), information from a communication source and so on. The application of a Markov chain stems from the fact that its evolution can be studied easily and has strong structural properties.

1.1 Evolution of a Markov chain

We consider a Markov chain with a state space S and transition matrix M. The initial state (or a probability distribution) is specified by a vector μ of dimension $|S|$. The different entries of μ correspond to the probability of being in some state $i \in S$ at $t = 0$. We need to find the probability distribution at some future time t, say μ_t. In other words, $\mu_0 = \mu$.

Exercise 6. What will be the probability distribution at $t = 1$?

First, let us calculate the probability of being in some state $i \in S$ at $t = 1$. We must have moved to i from some $j \in S$ at time $t = 0$. So,

$$P(X_1 = i) = \sum_{j \in S} P\left(X_1 = i | X_0 = j\right) P(X_0 = j) = \sum_{j \in S} M_{ji} \mu_0(j).$$

A keener look at the equation shows that this is the dot product between μ_0 and the i-th column vector of M. We get the equation,

$$\mu_1^T = \mu^T M.$$

What about μ_2? What is the probability of begin in state j in 2 steps starting from state i? Again, we can be any intermediate step $u \in S$ at $t = 1$.

$$P\left(X_2 = j | X_0 = i\right) = \sum_{u \in S} P\left(X_1 = u | X_0 = i\right) P\left(X_2 = j | X_1 = u\right) = \sum_{u \in S} M_{iu} M_{uj}.$$

Notice that the equality follows from Markov property. Again, by the property of matrix multiplication,

$$P\left(X_2 = j | X_0 = i\right) = M_{ij}^2.$$

You will prove in the assignment that M_{ij}^n is nothing but the probability of being in state j after n steps, if we start from state i. So, we get the more general equation for evolution,

$$\mu_n^T = \mu^T M^n.$$

This shows that the state of a Markov chain at any time t is completely determined by its initial state μ and transition matrix M. Also, M^n gives the probabilities of moving between two states in n steps.

We will now see that under some mild conditions on a Markov chain, much more can be said about its evolution.

2 Regular Markov chains and stationary distribution

The exposition in this section is taken from the book by Grinstead and Snell ([2]).

We are going to prove a very surprising result; under a few restrictions, whatever be the starting state, a Markov chain ends up in the same distribution. Such a distribution is called a *stationary distribution*.

Before we formally define stationary distributions and these restrictions, let us first see why some such restrictions are needed. Suppose someone tells you that stationary distribution exists for all Markov chains.

Exercise 7. Can you think of an example where these is no stationary distribution.

Trivially, you can have two kind of states, s.t., going from one kind to another is impossible. So, if we start from a distribution on first kind, we can't be in the same distribution as if we started at a distribution from second kind. Though, it is possible that both tend to different distributions.

What about this simple chain with transition matrix,

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The state will keep fluctuating from state 1 to state 2 and vice versa.

Removing these weird cases, we can define a *regular* Markov chain to be the one, s.t.,

$$\exists m : M_{ij}^m > 0 \quad \forall i, j \in S.$$

Intuitively, there exists some m, such that, the probability of going from state i to state j after m steps is positive for all i, j. Notice the order of quantifiers here, there exists a single m for which all the entries of M^m are positive.

Our goal in this section is to prove that a stationary distribution exists for all regular Markov chains. Let us define the stationary distribution now. A stationary distribution is a vector w, s.t., $w^T = w^T M$. In other words, it is the eigenvalue 1 eigen-vector of M^T. We will also prove, whatever be the starting state μ, the final state $\lim_{n\to\infty} \mu^T M^n$ tends to w (the stationary distribution).

Note 1. The formal definition of stationary distribution seems to be different from the intuition given at the beginning of this section. These two definitions will merge after the proof below.

Let M be the transition matrix of a regular Markov chain. Let us first see how M^n behaves as n tends to infinity. Let $Q = M^m$ be such that all entries of Q are positive (there exists such an m, because M is regular).

Exercise 8. Show that all rows of Q are probability vectors themselves.

We can instead look at the behavior of Q^n as n tends to infinity. That means, we can think of Q as one step of the Markov chain.

We already noticed that each row of Q is a *probability vector*, i.e., each entry is positive and they sum up to 1. In other words, multiplying any vector y by a row of Q, averages the entries of y. Doing it for all rows, the entries of y become mixed. The following theorem proves it formally.

Theorem 1. *Let M be a transition matrix of a regular Markov chain. Define $Q := M^m$ be the matrix, such that, all of its entries are positive. Then, for any y, $Q^n y$ tends to a constant vector (all entries are same).*

Proof. To start with, assume that the maximum entry and minimum entry of y are M_0 and m_0 respectively. After n applications of Q, say the maximum and minimum entries become M_n and m_n respectively. We want to show that the gap $M_t - m_t$ keeps decreasing, eventually moving to zero.

We know that all entries of Q are positive, assume that all of them are more than d (some constant bigger than zero).

Exercise 9. Prove that d is smaller than half, if there are more than 2 states in the state space.

Denote $y_1 = Qy$, then every entry of y_1 is smaller than,

$$(1-d)M_0 + dm_0.$$

This follows because the coefficient of m_0 is at least d and all other entries are at most as big as M_0.

Similarly every entry is bigger than $(1-d)m_0 + dM_0$. In other words, $M_1 - m_1$ is smaller than,

$$(1-d)M_0 + dm_0 - ((1-d)m_0 + dM_0) = (1-2d)(M_0 - m_0).$$

Since $0 < d \leq 1/2$, this shows that the gap between the maximum and minimum entry keeps decreasing. In other words, $Q^n y$ tends to a constant vector as $n \to \infty$.

□

Our assertions before the theorem follows from it easily. We will only give an informal argument and not worry about limits. We know that $Q^n = M^{nm}$ and $Q^n y = c_y \mathbf{1}$, where $\mathbf{1}$ is the all 1's vector and c_y is some constant (might depend upon y).

To simplify the notation, replace nm by n, we get that $M^n y = c_y \mathbf{1}$. This is a very strong property, replacing y by a vector with just one 1 (standard basis vectors), we get that every column of M^n is a constant vector.

Since every column is a constant vector, every row of M^n is some fixed vector w. Notice the difference between fixed and constant vector. Fixed implies, that all rows are same (vector w), individual entries of these rows need not be same.

Exercise 10. Show that entries of w sum to 1. Moreover, it is a probability vector.

Now, we show that w is the stationary distribution for M. Let u be a probability vector, then $u^T M^n = w$ (all rows of M^n are w). So, we can start in any state, ultimately the Markov chain will end up in w.

Also $w^T M^n = w$, by taking u to be w. You can also see that $M^{n+1} = M \cdot M^n$ has all rows w. So, $w = wM^{n+1} = wM^n M = wM$, proving that w is the stationary distribution.

3 Page rank algorithm

We will find an algorithm to rank pages on internet using ideas from Markov chains. This *page rank* algorithm motivated the actual algorithm for ranking pages on Google. The actual algorithm is kept secret by Google :).

What is the problem? Internet has a huge collection of pages. We need to rank pages according to their importance. Once done, we can output search result on a query by outputting high ranking pages containing that query. It could also be a more involved combination of page rank and similarity of the page to the query.

For now, we will only worry about ranking pages on the internet irrespective of the query. What information tells us that a page is more important than other? We have the information about connections between pages, i.e., which pages link to each other.

In other words, we are given a *directed* graph $G = (V, E)$, where V is the set of pages and $(i, j) \in E$ is a directed edge if i links to j. The importance of page i should be more if many pages link to it. Let p_i be the *importance score* of page i. Our first attempt could be,

$$p_i = \sum_{j:(j,i)\in E} 1.$$

This is a solution, but not a great one. We need to consider two more factors in calculating importance.

- If there are many link from page j, its contribution in the importance of page i should be small.
- A link from an important page should be weighted more.

The first point is easy to handle, say m_j is the total number of links from page j. Then,

$$p_i = \sum_{j:(j,i)\in E} 1/m_j.$$

How should we take care of the second point? We could multiply the contribution of j by p_j,

$$p_i = \sum_{j:(j,i)\in E} p_j/m_j. \tag{1}$$

Though, this makes our equation circular. We don't even know whether it will have a solution or not.

Let us look at this problem (and these equations) from another viewpoint, the one of Markov chains. Assume that a person is surfing the web. She starts from a random page. At each time step, she randomly moves to another web-page by clicking on a link from the current web-page.

This should remind you of a Markov chain. The state space will be the set of all pages. The transition probability from page j to i is $1/m_j$. From our discussion about the stationary distribution, wherever this person starts, she will eventually end up at the stationary distribution. Our final importance score for a page should be the corresponding entry in the stationary distribution.

Exercise 11. Let M be the transition matrix for the above chain. Show that $pM = p$ is same as Eq. 1.

So, to find the importance score, we just need to find the stationary distribution for the Markov chain. There is a catch, what is the guarantee that the chain M is regular. Otherwise, we might not have a stationary distribution (non-unique) or the surfer might not converge to the stationary distribution.

Exercise 12. Construct a Graph of pages, such that, the corresponding chain is not regular.

There is a simple fix to the problem above. We can assume that the surfer at any time step could just get bored with probability p and start fresh. In other words, with probability p she can go to an arbitrary web-page, and with probability $(1-p)$ she moves like before (using the links on the page).

The new transition matrix is given by,

$$M'_{ji} = \begin{cases} p/n + (1-p)/m_j & \text{if } j \text{ links to } i \\ p/n & \text{if } j \text{ does not link to } i \end{cases}$$

This is equivalent to the transformation,

$$M' = (p/n)\, J + (1-p)M,$$

where M is the old transition matrix and J is the all 1's matrix.

Since every entry of M' is positive (why?), it defines a regular Markov chain.

Exercise 13. Verify that M' defines a Markov chain.

So, a stationary distribution exists for M'. The entries in the stationary distribution give the importance of the web-page. They can be sorted or used with other measures to output search results.

Another way to modify the Markov chain: There is one more way to modify the Markov chain over web-pages to make it regular. A Markov chain is called *Ergodic* if for all i and j there exist an n, such that, $M^n_{ij} > 0$. Notice the difference between regular and Ergodic, in case of Ergodic, n can depend upon i and j.

Exercise 14. Construct a Markov chain which is Ergodic but not regular.

The method described below works for any Ergodic Markov chain and converts it into a regular chain. In our surfing example, Ergodic means that the graph is connected. We can assume that at every time step, with half the probability surfer stays at the same webpage or moves to another page via a link on the current web-page.

Exercise 15. Show that the new transition matrix is $M' = 1/2I + 1/2M$.

We will show that if M is Ergodic then M' is regular. You will show in the assignment that both M and M' have the same stationary distribution.

The idea is pretty simple.

- Suppose there are r states in the state space of M. Then, for all i and j, there exists $m < r$ such that $M^m_{ij} > 0$.
- If $M'^n_{ij} > 0$, then $M'^m_{ij} > 0$ for all $m \geq n$.

The first step follows from the graph theoretic fact that if there is a path between i and j, then there is a path between i and j with length less than the number of vertices. For the second step, notice that all diagonal entries of M' are positive (there is a positive chance that we stay put at a web-page). So, once we arrive at a web-page, there will always be a positive probability of being at that web-page at any future time step.

Exercise 16. Prove the second step formally using induction.

So, we can always convert an Ergodic Markov chain to a regular one and find the stationary distribution for the regular Markov chain.

Computing stationary distribution: There is another computational question here. How will we compute the stationary distribution? There are billions of web-pages on the internet. A usual algorithm to find the eigenvector will take lot of time. Luckily, our graph is very sparse (a single web-page is connected to only few web-pages). Solving linear equations for sparse matrices form a big area of study in theoretical computer science.

One of the way to compute the stationary distribution would be to compute $M'^n u$ for any initial distribution u and large n. We know that it should converge to the stationary distribution. Again, we need techniques for fast multiplication of sparse matrices to compute $M'^n u$.

4 Assignment

Exercise 17. Prove that T^n_{ij} is the probability of being in state j after n steps, if we start from state i.

Exercise 18. Read about Ergodic Markov chains.

Exercise 19. Let $\{X_i\}_i$ be a set of independent random variables. Show that $Y_i = \max\{X_1, X_2, \cdots, X_i\}$'s and $Z_i = \max\{X_1, X_2, \cdots, X_i\}$'s are Markov processes. Is $\{Y_i + Z_i\}_i$ a Markov process?

Exercise 20. Show that T and $1/2I + 1/2T$ have the same stationary distribution.

References

1. H. Tijms Understanding Probability. *Cambridge University Press*, 2012.
2. C. Grinstead and J. Snell Introduction to Probability. *American Mathematical Society*, 2003.
3. D. Stirzaker. Elementary Probability. *Cambridge University Press*, 2003.
4. D. Kahneman. Thinking, Fast and Slow. *Farrar, Straus and Giroux*, 2011.