Definition and Examples of DTMCs

Natarajan Gautam
Department of Industrial and Systems Engineering
Texas A&M University
235A Zachry, College Station, TX 77843-3131
Email: gautam@tamu.edu
Phone: 979-845-5458
Fax: 979-847-9005

August 11, 2009

Abstract

The objective of this chapter is to describe a special type of stochastic process called discrete time Markov chain (DTMC). We first provide definitions for DTMCs using simple language as well as mathematics. Then we describe a systematic technique to model a system as a DTMC. Finally we provide examples to both illustrate the modeling technique as well as to motivate the breadth of real-world applications DTMCs can be used for.

A stochastic process describes the state of a system and its random evolution over time. Broadly speaking, stochastic processes are of four types depending on whether the system is observed continuously or at discrete time points, and whether the observations or states are discrete quantities or continuous. For example:

1. *Discrete state and discrete observation*: the number of cereal boxes on a grocery store shelf observed at the beginning of a day;

2. *Discrete state and continuous observation*: the number of emails in someone's *InBox* observed at all times;

3. *Continuous state and discrete observation*: the amount of water in a reservoir observed immediately after a rain shower;

4. *Continuous state and continuous observation*: the temperature inside a car observed all the time.

A discrete time Markov chain (DTMC) is a special type of stochastic process belonging to the first category, i.e. system states are discrete and system is observed at discrete times. All DTMCs are discrete state and discrete observation stochastic processes but for a stochastic process to be a DTMC, some other conditions need to be satisfied which will be provided

in the next section. However, it is crucial to point out that the observations do not have to be periodic (although in most cases they would be), such as every day or every hour, etc. It is also critical to realize that the notion of "time" is not rigid, for example the observations can be done over space (but that is certainly atypical). In summary, DTMCs can be applied in fairly general settings to situations commonly found in real-life. They have gained significant attention in the queueing [1] and inventory [4] literature, and have been used extensively in the analysis of production systems, computer-communication systems, transportation systems, biological systems, etc.

We first describe the definition and notations used in modeling a system as a DTMC. This is the focus of Section 1. Then in Section 2 we provide a methodical technique to model a given system using a DTMC. Finally in Section 3, we describe several examples to illustrate the modeling framework as well as to motivate the wide variety of applications where DTMCs are used. In addition, there are several texts (viz. [2] and [3]) and websites (http://www.sosmath.com/matrix/markov/markov.html) that provide additional examples that would enhance the understanding of concepts included in this chapter.

1 Definition and Notation

Consider a system that is randomly evolving in time. Let X_n be the state of the system at the n^{th} observation. In essence, X_n describes the system based on the n^{th} observation. For example, X_n could be: (a) the status of a machine (up or down) at the beginning of the n^{th} hour; (b) the number of repairs left to be completed in a shop at the end of the n^{th} day; (c) the cell phone company a person is with when he/she received his/her n^{th} phone call; (d) the number of boys and the number of girls in a classroom at the beginning of the n^{th} class period; (e) the number of gas stations at the n^{th} exit on a highway.

It is crucial to understand that although all the examples above are discrete states and discrete observations, it is not necessary that the stochastic process $\{X_0, X_1, X_2, \ldots\}$ is a DTMC. For that some additional properties are necessary. However, notice from the examples that the states can be one-dimensional or multi-dimensional, they can take finite or infinite values, they could be over time or space, they do not have to be equally spaced (periodic), and, they do not have to be numerical values. We define the *state space* S as the set of all possible values of X_n for all n. In the previous paragraph, in example (a) we have $S = \{Up, Down\}$, in example (b) we have $S = \{0, 1, 2, \ldots\}$, and in example (d) we have $S = \{(0,0), (0,1), (1,0), (1,1), (0,2), (2,0), \ldots\}$. In summary, we require X_n and S to be discrete-valued (or also called countable).

The next question to ask is: When can a stochastic process $\{X_0, X_1, X_2, \ldots\}$ be called a DTMC? A stochastic process $\{X_0, X_1, X_2, \ldots\}$ which we will henceforth denote as $\{X_n, n \geq 0\}$, is called a DTMC if it satisfies the *Markov property*. In words, Markov property states that if the current state of the system is known, the future states are independent of the past. In other words, to predict (probabilistically) the states in the future, all one needs to know is the present state and nothing from the past. For example, if X_n denotes the amount of inventory of a product at the beginning of the n^{th} day (say today), then it is conceivable that to predict the inventory at the begninning of tomorrow and day after, all one would

need is today's inventory. Yesterday's inventory or how you got to today's inventory would not be necessary.

Mathematically, Markov property is defined as

$$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, \ldots, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\}.$$

In other words, once X_n is known, we can predict (probabilistically) X_{n+1}, X_{n+2}, etc. without needing to know X_{n-1}, X_{n-2}, etc. In summary, a stochastic process $\{X_n, n \geq 0\}$ defined on a countable state space S is called a DTMC if it satisfies the Markov property. That is technically the definition of a DTMC. Now that we have defined a DTMC, in the next section we explain how a system can be modeled as a DTMC so that the system can be analyzed.

2 Modeling a System as a DTMC

Before presenting a systematic approach to modeling a system as a DTMC, we provide some more assumptions and notation. Besides the Markov property there is another property that would need to be satisfied in order to be able to effectively analyze a DTMC. A DTMC is called *time-homogeneous* if the probability of transitioning from a state to another state does not vary with time. In other words, for a DTMC with state space S to be time-homogeneous, for every $i \in S$ and $j \in S$,

$$P\{X_{n+1} = j | X_n = i\} = P\{X_1 = j | X_0 = i\}.$$

We denote the above expression as p_{ij}, the one step transition probability of going from state i to j. Therefore for a time-homogeneous DTMC, for $i \in S$ and $j \in S$,

$$p_{ij} = P\{X_{n+1} = j | X_n = i\}.$$

Notice that p_{ij} is not a function of n which is essentially because the DTMC is time-homogeneous.

Using the transition probabilities p_{ij} for every $i \in S$ and $j \in S$, we can build a square matrix $P = [p_{ij}]$. The matrix P is called *transition probability matrix*. The rows of the matrix correspond to the given current state, while the columns correspond to the next state. The sum of the elements of each row adds to 1 because given that the DTMC is in state i (for some $i \in S$), the next state ought to be one of the states in S hence for every $i \in S$,

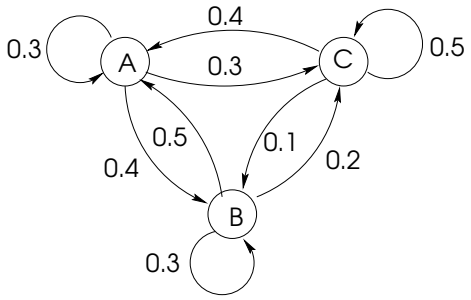
$$\sum_{j \in S} p_{ij} = 1.$$

Oftentimes a transition diagram is used to represent the transition probabilities, as opposed to the matrix P. A transition diagram is a directed graph with nodes corresponding to the states (thereby the set of nodes is indeed S), arcs corresponding to possible 1-step transitions and arc values being transition probabilities. To draw a transition diagram, for every $i \in S$ and $j \in S$, draw an arc from node i to node j if $p_{ij} > 0$. Now we illustrate an example of X_n, S, the transition probability matrix P and transition diagram.

Consider three long-distance telephone companies A, B and C. Everytime a sale is announced, users switch from one company to another. Let X_n denote the long-distance company with which a particular user named Jill is just before the n^{th} sale announcement. We have $S = \{A, B, C\}$. Based on the customers' switching in the past it is estimated that Jill's switching patterns follow the following transition probability matrix:

$$P = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{bmatrix} 0.3 & 0.4 & 0.3 \\ 0.5 & 0.3 & 0.2 \\ 0.4 & 0.1 & 0.5 \end{bmatrix} \end{array}.$$

For example, if Jill is with B before a sale is announced, she would switch to A with probability 0.5 and C with probability 0.2 or stay with B with probability 0.3 as evident from the second row in P. Now, converting the above P matrix into a transition diagram, we get the picture shown below.



Having described all the definitions, notation and requirements, we are now ready to model a system as a DTMC. There are 5 steps involved in systematically modeling a system as a DTMC: (1) Define X_n; (2) Write down S; (3) Verify that the Markov property is satisfied; (4) Verify that the DTMC is time homogeneous; (5) Obtain P or draw the transition diagram.

Next we present some examples where systems are modeled as DTMCs using the 5 steps described above. Note that although modeling a system as a DTMC is considered an art, the key scientific aspect is that X_n must be chosen carefully so that S is as small as possible and steps (3), (4) and (5) can be performed. If not, one must consider revising X_n. In particular, one of the most important steps in modeling is to choose X_n so that Markov property is satisfied. One approach is to defined X_n in such a way that it carries all the past information needed to forecast future states, or that the future state could be written in terms of only the present state and some other factors that are independent of the past and present states. We illustrate this in some of the examples in the next section.

3 Examples

In the examples that follow, the objective is to model the system described as a DTMC by stating X_n and S as well as obtaining P. The order of states in the rows and columns of the P matrices are identical to those in the order of the corresponding S.

Example 1 *A company uses two forecasting tools for making demand predictions. Tool i is effective with probability p_i (for $i = 1, 2$). If the n^{th} prediction uses tool i and it is observed to be effective, then the $(n+1)^{st}$ prediction is also done using the same tool; if it is observed to be ineffective, then the $(n+1)^{st}$ prediction is made using the other tool. This is sometimes known as "play the winner rule".*

To model this system as a DTMC, we let X_n to be the tool used for the n^{th} prediction. The state space is $S = \{1, 2\}$. It can be verified that Markov property is satisfied because the tool used for the $n+1^{st}$ prediction only depends on what was used for the n^{th} prediction and the outcome of the n^{th} prediction but not the history. Further, p_i for $i = 1, 2$ remains constant over time, hence the DTMC is time-homogeneous. Clearly from the description,

$$P = \begin{bmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{bmatrix}.$$

Example 2 *Probe vehicles are sent through two similar routes from source A to destination B to determine the travel times. If a route was congested when a probe vehicle is sent, then it will be congested with probability q when the next probe vehicle needs to be sent. Likewise, if a route was not congested when a probe vehicle is sent, it will not be congested with probability p when the next probe vehicle is sent. The routes behave independently of each other and since they are similar we assume that p and q do not vary between the routes.*

Let X_n be the number of uncongested routes when the n^{th} set of probe vehicles are sent. The state space is $S = \{0, 1, 2\}$. It can be verified that Markov property is satisfied because the congestion states of each route depends only on whether they were congested during the previous probe but not the history. Further p and q remain constant over time, hence the DTMC is time-homogeneous. From the description it is possible to show that,

$$P = \begin{bmatrix} q^2 & 2q(1-q) & (1-q)^2 \\ q(1-p) & pq + (1-p)(1-q) & p(1-q) \\ (1-p)^2 & 2p(1-p) & p^2 \end{bmatrix}.$$

It may be worthwhile to describe some of the above p_{ij} values. In particular $p_{02} = (1-q)^2$ since the probability of going from both routes being congested to no route congested is if both congested routes become uncongested, each happeneing with probability $1-q$. Likewise $p_{10} = q(1-p)$ because the probability that the congested route remains congested is q and the probability that the uncongested route becomes congested is $1-p$. In a similar fashion all the p_{ij} values can be obtained using a similar logic.

Example 3 *Consider the following weather forecasting model: if today is sunny and it is the n^{th} day of the current sunny spell, then it will be sunny tomorrow with probability p_n regardless of what happened before the current sunny spell started. Likewise, if today is rainy and it is the n^{th} day of the current rainy spell, then it will be rainy tomorrow with probability q_n regardless of what happened before the current rainy spell started.*

Oftentimes one is tempted to say that the state of the system is the type of day (sunny or rainy). However, one of the concerns is that Markov property would not be satisfied since to predict the next state one needs to know the history to figure out how long the spell has been. Therefore one needs to carry the spell information too in the state. In that light let X_n be a 2-dimensional state denoting the type of day on the n^{th} day and how many days the current spell has lasted. Letting s and r to denote rainy and sunny respectively, the state space S is
$S = \{(s,1),(r,1),(s,2),(r,2),(s,3),(r,3),\ldots\}$. It can easily be verified that Markov property is satisfied. The DTMC is time homogeneous, although one must not mistake n in p_n or q_n to be the n in X_n. Consider that the system is in state (s,i) during a day. Then from the description, on day $n+1$ the system would be in state $(s,i+1)$ with probability p_i if it does not rain and $(r,1)$ with probability $1-p_i$ if it does rain. A similar argument can be made for (r,i). Therefore the P matrix in the order the states are represented in S is

$$P = \begin{bmatrix} 0 & 1-p_1 & p_1 & 0 & 0 & 0 & 0 & 0 & \ldots \\ 1-q_1 & 0 & 0 & q_1 & 0 & 0 & 0 & 0 & \ldots \\ 0 & 1-p_2 & 0 & 0 & p_2 & 0 & 0 & 0 & \ldots \\ 1-q_2 & 0 & 0 & 0 & 0 & q_2 & 0 & 0 & \ldots \\ 0 & 1-p_3 & 0 & 0 & 0 & 0 & p_3 & 0 & \ldots \\ 1-q_3 & 0 & 0 & 0 & 0 & 0 & 0 & q_3 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Example 4 *Many biological processes especially at the cellular level are a result of polymerization and depolymerization of organic compounds. A polymer in a cell is made up of N monomers that are attached back to back with one end of the chain anchored to the cell and the other end grows or shrinks. For example, a polymer $M_0 - M - M - M$ is made of 4 monomers and M_0 indicates it is anchored. In each time unit with probability p, a new monomer joins the growing end and with probability q, the non-anchored end monomer leaves the polymer. For example, the $M_0 - M - M - M$ chain in the next time unit becomes $M_0 - M - M - M - M$ with probability p, $M_0 - M - M$ with probability q, and stays as $M_0 - M - M - M$ with probability $1 - p - q$.*

We consider a single polymer and let X_n denote the length of the polymer (in terms of number of monomers) at the beginning of the n^{th} time unit. Therefore the state space is $S = \{1,2,3,4,\ldots\}$. Since one side in anchored we assume that the last monomer would never break away. Verify that Markov property is satisfied and $\{X_n, n \geq 0\}$ is time-homogeneous. The P matrix in the order the states are represented in S is

$$P = \begin{bmatrix} 1-p & p & 0 & 0 & 0 & 0 & 0 & \ldots \\ q & 1-p-q & p & 0 & 0 & 0 & 0 & \ldots \\ 0 & q & 1-p-q & p & 0 & 0 & 0 & \ldots \\ 0 & 0 & q & 1-p-q & p & 0 & 0 & \ldots \\ 0 & 0 & 0 & q & 1-p-q & p & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

Example 5 *A machine produces two items per day. The probability that an item is nondefective is p. Successive items are independent. Defective items are thrown away instantly. The demand is one item per day which occurs at the end of a day. Any demand that cannot be satisfied immediately is lost.*

Let X_n be the number of items in storage at the beginning of the n^{th} day (before production and demand of that day). Since demand takes place after production on a given day, we have: if $X_n > 0$,

$$X_{n+1} = \begin{cases} X_n + 1 & \text{w.p. } p^2 \\ X_n & \text{w.p. } 2p(1-p) \\ X_n - 1 & \text{w.p. } (1-p)^2 \end{cases}$$

and, if $X_n = 0$,

$$X_{n+1} = \begin{cases} X_n + 1 & \text{w.p. } p^2 \\ X_n & \text{w.p. } 1 - p^2 \end{cases}$$

Since X_{n+1} only depends on X_n this is a DTMC with state space $\{0, 1, 2, \ldots\}$. The transition probabilities p_{ij} for all $i \geq 0$ and $j \geq 0$ is given by

$$p_{ij} = \begin{cases} p^2 & \text{if } j = i + 1 \\ 2p(1-p) & \text{if } j = i \text{ and } i > 0 \\ 1 - p^2 & \text{if } j = i = 0 \\ (1-p)^2 & \text{if } j = i - 1 \\ 0 & \text{otherwise.} \end{cases}$$

Example 6 *Consider a time division multiplexer from which packets are transmitted at times 0, 1, 2, etc. Packets arriving between time n and n + 1 have to wait until time n + 1 to be transmitted. However at most one packet can be transmitted at a time. Let Y_n be the number of packets that arrive during time n to n + 1. Assume that $a_i = P\{Y_n = i\}$ and that there is infinite room in the waiting space.*

Let X_n be the number of packets awaiting transmission just before time n. Clearly we have the state space $S = \{0, 1, 2, \ldots\}$. Define the transition probabilities p_{ij} as (for $i > 0$)

$$\begin{aligned} p_{ij} &= P\{X_{n+1} = j | X_n = i\} \\ &= a_{j-i+1}. \end{aligned}$$

Similarly,

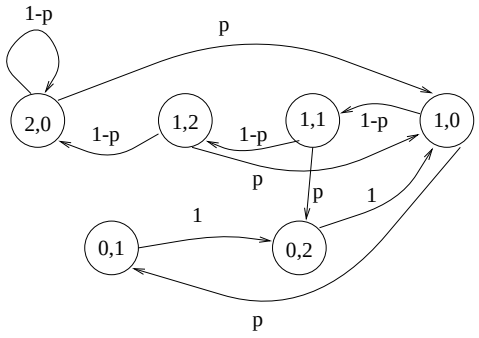
$$p_{0j} = P\{X_{n+1} = j | X_n = 0\} = P\{Y_n = j\} = a_j.$$

Verify that Markov property is satisfied and $\{X_n, n \geq 0\}$ is time-homogeneous. Therefore the transition probability matrix $P = [p_{ij}]$ is

$$P = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \ldots \\ a_0 & a_1 & a_2 & a_3 & \ldots \\ & a_0 & a_1 & a_2 & \ldots \\ & & a_0 & a_1 & \ldots \\ & & & a_0 & \ldots \end{bmatrix}.$$

Example 7 *Consider a manufacturing system of 2 identical machines. If both machines are in working condition, only one is in use and the other one is on standby. The probability that a machine that is in use fails during an hour is p (assume that a machine in standby mode does not fail). The system is observed once every hour. It takes just a little less than 3 hours to repair a failed machine. Only one machine can be repaired at a time and repairs are according to first come first serve basis.*

Let X_n = number of machines in working condition at beginning of time unit n.
Y_n = number of time units of repair complete at beginning of time unit n.
Notice that X_n and Y_n are chosen such that Markov property is satisfied and $\{(X_n.Y_n), n \geq 0\}$ is time-homogeneous. Therefore, $\{(X_n, Y_n), n \geq 0\}$ is a DTMC with the following transition diagram:



References

[1] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory.* 3rd Ed., John Wiley and Sons Inc., New York, 1998.

[2] V.G. Kulkarni. *Modeling and Analysis of Stochastic Systems.* Texts in Statistical Science Series. Chapman and Hall, Ltd., London, 1995.

[3] S.M. Ross. *Introdution to Probability Models.* Academic Press, San Diego, CA, 2003.

[4] P.H. Zipkin. *Foundations of Inventory Management.* McGraw Hill and Company Inc., 2000.