Chapter 3. Stochastic Processes

So far, when discussing random variables, random vectors and their distributions, we described the situation at a particular moment of time, as if someone had said "Freeze!" and everything stood still. But the real world is dynamic and many random variables develop and change in time (think stock prices, air temperatures, interest rates, football scores, CPU usage, the speed of internet connection, popularity of politicians, and so on).

Basically, stochastic processes are random variables that *evolve and change in time*.

1 Basic Notions

Definition 1.1. *A **stochastic process** is a random variable that also depends on time. It is denoted by $X(t, e)$ or $X_t(e)$, where $t \in \mathcal{T}$ is time and $e \in S$ is an outcome. The values of $X(t, e)$ are called **states**.*

If $t \in \mathcal{T}$ is fixed, then X_t is a random variable, whereas if we fix $e \in S$, X_e is a function of time, called a **realization** or **sample path** or **trajectory** of the process $X(t, e)$.

Definition 1.2. *A stochastic process is called **discrete-state** if $X_t(e)$ is a discrete random variable, for all $t \in \mathcal{T}$ and **continuous-state** if $X_t(e)$ is a continuous random variable, for all $t \in \mathcal{T}$. Similarly, a stochastic process is said to be **discrete-time** if the set $\mathcal{T}$ is discrete and **continuous-time** if the set of times $\mathcal{T}$ is a (possibly unbounded) interval in $\mathbb{R}$.*

Example 1.3.

1. Available memory, CPU usage, in percents, is a continuous-state, continuous-time process.
2. The CPU usage *per hour* is continuous-state, discrete-time.
3. In a printer shop, $X_n(e)$, the amount of time required to print the n^{th} job, is a discrete-time, continuous-state stochastic process, because $n = 1, 2, \ldots$ and $X \in (0, \infty)$.
4. On the other hand, $Y_n(e)$, the number of pages of the n^{th} printing job, is discrete-time and discrete-state. In this case, $Y = 1, 2, \ldots$, which is a discrete set.
5. The *actual* air temperature $X_t(e)$ at time t is a continuous-time, continuous-state stochastic process. Indeed, it changes smoothly and never jumps from one value to another.
6. However, $Y_t(e)$, the temperature reported *every hour* on radio or TV, is a discrete-time process. Moreover, since the reported temperature is usually rounded to the nearest degree, it is also a discrete-state process.

Throughout the rest of the course, we will omit writing e as an argument of a stochastic process (as it is customary when writing random variables).

2 Markov Processes and Markov Chains

2.1 Transition Probability Matrix

Definition 2.1. *A stochastic process X_t is **Markov** if for any times $t_1 < t_2 < \ldots < t_n < t$ and any sets $A_1, A_2, \ldots, A_n; A$,*

$$P(X_t \in A \mid X_{t_1} \in A_1, \ldots, X_{t_n} \in A_n) = P(X_t \in A \mid X_{t_n} \in A_n). \tag{2.1}$$

What this means is that the conditional distribution of X_t, given observations of the process *at several moments in the past*, is the same as the one given *only the latest* observation. In other words, knowing the *present*, we get no information from the *past* that can be used to predict the *future*:

$$P(\text{future} \mid \text{past, present}) = P(\text{future} \mid \text{present}).$$

Then, for the future development of a Markov process, only its present state is important, and it does not matter *how* the process arrived to this state.

Some processes satisfy the Markov property, others don't.

Example 2.2.
1. Let X_t be the total number of internet users registered by some internet service provider by the time t. If, say, there were 999 users connected by 10 o'clock, then their total number will be or exceed 1000 during the next hour *regardless* of when and how those 999 users connected to the internet in the past. The number of connections in an hour will only depend on the current number. This process *is* Markov.
2. Let Y_t be the value of some stock or some market index at time t. If we know $Y(t)$, do we also want to know $Y(t-1)$ in order to predict $Y(t+1)$? One may argue that if $Y(t-1) < Y(t)$, then the market is rising, therefore, $Y(t+1)$ is likely (but not certain) to exceed $Y(t)$. On the other hand, if $Y(t-1) > Y(t)$, we may conclude that the market is falling and may expect $Y(t+1) < Y(t)$. It looks like knowing the past, in addition to the present, did help us to predict the future. In this case, to make predictions about the future, we need a history (so the past, too, not just the present). Then, this process is *not* Markov.

Due to a well-developed theory and a number of simple techniques available for Markov processes, it is important to know whether a stochastic process is Markov or not.

Remark 2.3. The idea of Markov dependence was proposed and developed by Andrey A. Markov ($1856-1922$), who was a student of P. L. Chebyshev at St. Petersburg University (Russia).

Definition 2.4. *A discrete-state, discrete-time Markov stochastic process is called a* ***Markov chain****.*

To simplify the writing, we use the following notations: Since a Markov chain is a discrete-time process, we can consider the time set as $\mathcal{T} = \{0, 1, 2, \dots\}$ and the Markov chain as a sequence of random variables

$$\{X_0, X_1, \dots\},$$

where X_k describes the situation at time $t = k$.

It is also a discrete-state process, so we can denote the states (the values that the Markov chain takes) by $1, 2, \dots, n$. Sometimes we will start enumeration from state 0, and sometimes we might deal with a Markov chain with infinitely many (discrete) states, then we will have $n = \infty$.
Then the random variable X_k has the pdf

$$X_k \begin{pmatrix} 1 & 2 & \dots & n \\ P_k(1) & P_k(2) & \dots & P_k(n) \end{pmatrix}, \tag{2.2}$$

where

$$\begin{aligned} P_k(1) &= P(X_k = 1), \\ P_k(2) &= P(X_k = 2), \\ \dots, & \\ P_k(n) &= P(X_k = n). \end{aligned}$$

Since the states (the values of the random variable X_k) are the same for each k, one only needs the second row to describe the pdf. Then let

$$P_k = [P_k(1) \;\; P_k(2) \;\; \dots \;\; P_k(n)] \tag{2.3}$$

denote the vector on the second row of the pdf (2.2). Obviously,

$$\sum_{i=1}^{n} P_k(i) = 1.$$

So, in short, we can write the pdf of X_k as

$$X_k \begin{pmatrix} 1 & \dots & n \\ & P_k & \end{pmatrix}.$$

The Markov property (2.1) means that in predicting the value of X_{t+1}, i.e. in which state j it is and with what probability $P_{t+1}(j)$, only the value i of X_t matters. So (2.1) can now be written as

$$P(X_{t+1} = j \mid X_t = i, X_{t-1} = l, \dots) \;=\; P(X_{t+1} = j \mid X_t = i), \text{ for all } t \in \mathcal{T}. \tag{2.4}$$

We summarize this information in a matrix.

Definition 2.5.

- *The conditional probability*

$$p_{ij}(t) \;=\; P(X_{t+1} = j \mid X_t = i) \tag{2.5}$$

is called a ***transition probability****; it is the probability that the Markov chain transitions from state i to state j, at time t. The matrix*

$$P(t) \;=\; \left[p_{ij}(t)\right]_{i,j=\overline{1,n}} \tag{2.6}$$

is called the ***transition probability matrix*** *at time t.*

- *Similarly, the conditional probability*

$$p_{ij}^{(h)}(t) \;=\; P(X_{t+h} = j \mid X_t = i) \tag{2.7}$$

is called an ***h-step transition probability****, i.e. the probability that the Markov chain moves from state i to state j in h steps, and the matrix*

$$P^{(h)}(t) \;=\; \left[p_{ij}^{(h)}(t)\right]_{i,j=\overline{1,n}} \tag{2.8}$$

is the ***h-step transition probability matrix*** *at time t.*

Definition 2.6. *A Markov chain is* ***homogeneous (or stationary)*** *if all transition probabilities are* *independent of time,*

$$\begin{aligned} p_{ij}(t) &= p_{ij}, \\ P(t) &= P = [p_{ij}]_{i,j=\overline{1,n}}, \\ p_{ij}^{(h)}(t) &= p_{ij}^{(h)}, \\ P^{(h)}(t) &= P^{(h)} = \left[p_{ij}^{(h)}\right]_{i,j=\overline{1,n}}. \end{aligned}$$

Being homogeneous means that transition from i to j has the same probability *at any time*.

By the Markov property, each next state can be predicted from the previous state *only*. So, when working with Markov chains, we will need to know:

- X_0, its initial situation, i.e. the distribution of its initial state, P_0;
- the mechanism of transitions from one state to another, i.e. the matrix P.

Based on this, we want to find:

- h-step transition probabilities $p_{ij}^{(h)}$ and $P^{(h)}$;
- the distribution of states at time h, X_h, i.e. P_h, which will be our forecast;
- possibly the limit of $P^{(h)}$ and P_h as $h \to \infty$, i.e. a *long-term* forecast; as we will see, when making forecasts for *many* transitions ahead, computations will become rather lengthy, and thus, it will be more efficient to take the limit.

In order to better understand the ideas and the computations, let us start with a simple example and then discuss the general formulas.

Example 2.7. In Rainbow City, each day is either sunny or rainy. A sunny day is followed by another sunny day with probability 0.7, while a rainy day is followed by a sunny day with probability 0.4. Suppose it rains on Monday. Make forecasts for Tuesday.

Solution. This process has two states, $1 =$ "sunny" and $2 =$ "rainy", so it is **discrete-state**. The time set $\{$Monday, Tuesday, $\ldots\}$ is also discrete, so it is **discrete-time**.
Since the weather forecast for each day depends *only* on the weather the previous day, it is a **Markov** process and, hence, a **Markov chain**.
Finally, since transition probabilities are the same for *any* two consecutive times (days), it is also **homogeneous**.

Thus, X_k, the weather situation on day k, is a homogeneous Markov chain with 2 states.
The initial situation (on Monday) is

$$X_0 \begin{pmatrix} 1 & 2 \\ 0 & 1 \end{pmatrix}, \;\; P_0(1) = 0, \; P_0(2) = 1, \;\; P_0 = [0 \;\; 1].$$

The transition probability matrix is

$$P \;=\; \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \;=\; \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}.$$

This can also be seen in a *transition diagram* (Figure 1). Arrows represent all possible one-step transitions, along with the corresponding probabilities.

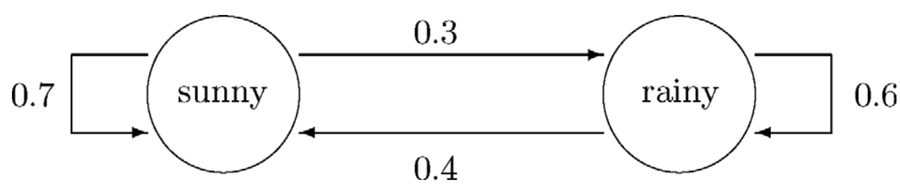

Fig. 1: Transition diagram for Example 2.7

Now, what is the prognosis for Tuesday ($t = 1$)? Since it rains on Monday, we only need to look at the second row in matrix P, the transition probabilities from state 2. Then the forecast for Tuesday is "sunny" with probability $p_{21} = 0.4$ (making a transition from a rainy to a sunny day) and "rainy" with probability $p_{22} = 0.6$. So for X_1, we have

$$X_1 \begin{pmatrix} 1 & 2 \\ 0.4 & 0.6 \end{pmatrix}, \;\; P_1(1) = 0.4, \; P_1(2) = 0.6, \;\; P_1 = [0.4 \;\; 0.6].$$

■

Now, before we go any further with our forecast, we need a little review. Recall the Total Probability Rule (Theorem 1.4j, in Lecture 1):

$$P(E) \;\;=\;\; \sum_{i \in I} P\left(E | A_i\right) P\left(A_i\right),$$

for any partition $\{A_i\}_{i \in I}$.

The same formula holds for a *conditional* probability, i.e.

$$P(E|B) \;=\; \sum_{i\in I} P\left(E|A_i\right) P\left(A_i|B\right), \tag{2.9}$$

if $\{A_i\}_{i\in I}$ is a partition of S and $P(B) \neq 0$.

Example 2.8. Assuming the same situation as before, make forecasts for Wednesday and Thursday.

Solution. To make forecasts for Wednesday, we need the 2-step transition probability matrix $P^{(2)}$, making one transition from Monday to Tuesday, X_0 to X_1, and another one from Tuesday to Wednesday, X_1 to X_2. We'll have to *condition* on the weather situation on Tuesday and use formula (2.9). Notice that the events $\Big\{\{\text{Tuesday is sunny}\}, \{\text{Tuesday is rainy}\}\Big\}$ form a partition. That is, $\{(X_1 = 1), (X_1 = 2)\}$ form a partition.

So, let us proceed:

$$\begin{aligned}
p_{21}^{(2)} &= P\left(\text{Wednesday is sunny} \mid \text{Monday is rainy}\right) \\
&= P(X_2 = 1|X_0 = 2) \\
&= P(X_2 = 1|X_1 = 1)P(X_1 = 1|X_0 = 2) \\
&+ P(X_2 = 1|X_1 = 2)P(X_1 = 2|X_0 = 2) \\
&= p_{11} \cdot p_{21} + p_{21} \cdot p_{22} \\
&= 0.7 \cdot 0.4 + 0.4 \cdot 0.6 \;=\; 0.52.
\end{aligned}$$

Obviously,

$$\begin{aligned}
p_{22}^{(2)} &= P\left(\text{Wednesday is rainy} \mid \text{Monday is rainy}\right) \\
&= 1 - P\left(\text{Wednesday is sunny} \mid \text{Monday is rainy}\right) \\
&= 1 - p_{21}^{(2)} \;=\; 0.48.
\end{aligned}$$

Thus, we have the second row of $P^{(2)}$, which is *all* we need to know in order to make forecasts for Wednesday:

$$X_2 \begin{pmatrix} 1 & 2 \\ 0.52 & 0.48 \end{pmatrix}, \;\; P_2(1) = 0.52,\; P_2(2) = 0.48,\;\; P_2 = [0.52 \;\; 0.48].$$

So, for Wednesday there is 52% chance of sun and 48% chance of rain.

For the Thursday forecast, we need to compute 3-step transition probabilities $p_{ij}^{(3)}$, because it takes *three* transitions to move from Monday to Thursday. We have to use the Total Probability Rule conditioning on *both* Tuesday and Wednesday. This corresponds to a sequence of states

$$2 \rightarrow i \rightarrow j \rightarrow 1.$$

Luckily, we have already computed the 2-step transition probabilities $p_{21}^{(2)}$ and $p_{22}^{(2)}$, describing transition from Monday to Wednesday. It remains to add *one* transition to Thursday. Thus,

$$\begin{aligned} p_{21}^{(3)} &= p_{21}^{(2)} \cdot p_{11} + p_{22}^{(2)} \cdot p_{21} \\ &= 0.52 \cdot 0.7 + 0.48 \cdot 0.4 \;=\; 0.556 \end{aligned}$$

and then,

$$p_{22}^{(3)} \;=\; 1 - p_{21}^{(3)} \;=\; 0.444.$$

So, for Thursday, we predict a 55.6% chance of sun and a 44.4% chance of rain.

■

Remark 2.9. Obviously, more remote forecasts require more lengthy computations. For a t-day ahead forecast, we have to account for *all* t-step paths on diagram Figure 1. Or, we use the of Total Probability Rule, conditioning on *all* the intermediate states $X_1, X_2, \ldots, X_{t-1}$. To simplify the task, we will employ matrices.

Recall multiplication of matrices. For two $n \times n$ matrices, $A = [a_{ij}]_{i,j=\overline{1,n}}$, $B = [b_{ij}]_{i,j=\overline{1,n}}$, the product is computed by

$$[A \cdot B]_{ij} \;=\; \underbrace{[a_{i1} \;\; \ldots \;\; a_{in}]}_{i^{th} \text{ row of } A} \cdot \underbrace{\begin{bmatrix} b_{1j} \\ \ldots \\ b_{nj} \end{bmatrix}}_{j^{th} \text{ col. of } B} \;=\; \sum_{k=1}^{n} a_{ik} \cdot b_{kj}.$$

Let us notice that

$$\underline{P_0 \cdot P} \;=\; [0 \quad 1] \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix} \;=\; [0.4 \quad 0.6] \;=\; \underline{P_1}. \tag{2.10}$$

Now, to get back to our task, from all the previous computations, let us notice that

$$\underline{P_0 \cdot P^{(2)}} \;=\; [0 \quad 1] \begin{bmatrix} \cdots & \cdots \\ 0.52 & 0.48 \end{bmatrix} \;=\; [0.52 \quad 0.48] \;=\; \underline{P_2}. \tag{2.11}$$

Even though it wasn't necessary for the Wednesday forecast, let us still compute the first row of $P^{(2)}$, in order to draw some conclusions. We proceed in a similar way (but write fewer details). We have

$$\begin{aligned} p_{11}^{(2)} &= P(X_2 = 1 | X_0 = 1) \\ &= P(X_2 = 1 | X_1 = 1) P(X_1 = 1 | X_0 = 1) \\ &+ P(X_2 = 1 | X_1 = 2) P(X_1 = 2 | X_0 = 1) \\ &= p_{11} \cdot p_{11} + p_{21} \cdot p_{12} \\ &= (0.7)^2 + 0.3 \cdot 0.4 \;=\; 0.61 \end{aligned}$$

and, of course,

$$p_{12}^{(2)} \;=\; 1 - p_{11}^{(2)} \;=\; 0.39.$$

So, we notice that

$$\begin{aligned} p_{11}^{(2)} &= p_{11} \cdot p_{11} + p_{21} \cdot p_{12} \\ &= [p_{11} \quad p_{12}] \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}, \\ p_{21}^{(2)} &= p_{11} \cdot p_{21} + p_{21} \cdot p_{22} \\ &= [p_{21} \quad p_{22}] \begin{bmatrix} p_{11} \\ p_{21} \end{bmatrix}. \end{aligned}$$

If we had computed the other two probabilities *directly*, we would have found that

$$\begin{aligned} p_{12}^{(2)} &= p_{11} \cdot p_{12} + p_{12} \cdot p_{22} \\ &= [p_{11} \quad p_{12}] \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix} \text{ and} \\ p_{22}^{(2)} &= p_{21} \cdot p_{12} + p_{22} \cdot p_{22} \\ &= [p_{21} \quad p_{22}] \begin{bmatrix} p_{12} \\ p_{22} \end{bmatrix}. \end{aligned}$$

So, in fact, we see that

$$P^{(2)} \;=\; \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} \;=\; P^2,$$

the *second power* of P.
Also, from (2.10) and (2.11), we notice that

$$P_0 \cdot P^{(i)} \;=\; P_i, \;\; i = 1, 2.$$

Now, we can state the general result.

Proposition 2.10 (Chapman-Kolmogorov). *Let $\{X_0, X_1, \dots\}$ be a Markov chain. Then the following relations hold:*

$$P^{(h)} \;=\; P^h \, (= \underbrace{P \cdot P \cdot \ldots \cdot P}_{h \text{ times}}), \;\text{ for all }\; h = 1, 2, \dots \tag{2.12}$$

$$P_i \;=\; P_0 \cdot P^{(i)} \;=\; P_0 \cdot P^i, \;\text{ for all }\; i = 0, 1, \dots \tag{2.13}$$

Proof.
The proof of (2.12) goes by induction.
Obviously, relation (2.12) is true for $h = 1$. Assume $P^{(h-1)} = P^{h-1}$.

For a matrix M, we use the notation $[M]_{ij} = M(i,j)$ and, similarly, for a vector v, $(v)_i = v(i)$.

Since the events $\{(X_{h-1} = k)\}_{k=\overline{1,n}}$ form a partition, using the Total Probability Rule (2.9) with $E = (X_h = j), B = (X_0 = i), A_k = (X_{h-1} = k), k = \overline{1,n}$, for $[P^{(h)}]_{ij} = p_{ij}^{(h)}$ (the (i,j)-entry in matrix $P^{(h)}$), we have

$$\begin{aligned} p_{ij}^{(h)} \quad &= \quad P(X_h = j \mid X_0 = i) \\ &= \quad \sum_{k=1}^{n} \underbrace{P(X_h = j \mid X_{h-1} = k)}_{p_{kj}} \cdot \underbrace{P(X_{h-1} = k \mid X_0 = i)}_{p_{ik}^{(h-1)}} \\ &= \quad \sum_{k=1}^{n} p_{ik}^{(h-1)} \cdot p_{kj} \;\; = \;\; \left[P^{(h-1)} \cdot P\right]_{ij} \\ &\overset{\text{ind. hyp.}}{=} \quad \left[P^{h-1} \cdot P\right]_{ij}, \;\text{ for all }\; i, j = \overline{1,n}, \end{aligned}$$

so

$$P^{(h)} = P^h.$$

To prove the second relation (2.13), for each $j = \overline{1,n}$, we have $[P_i]_j = P_i(j) = P(X_i = j)$. Again, using the Total Probability Rule for the partition $\{(X_0 = k)\}_{k=\overline{1,n}}$, with $E = (X_i = j)$ and $A_k = (X_0 = k)$), we get for $[P_i]_j$

$$\begin{aligned} P(X_i = j) &= \sum_{k=1}^{n} \underbrace{P(X_i = j \mid X_0 = k)}_{p_{kj}^{(i)}} \cdot \underbrace{P(X_0 = k)}_{[P_0]_k} \\ &= \sum_{k=1}^{n} [P_0]_k \cdot p_{kj}^{(i)} \\ &= \left[P_0 \cdot P^{(i)}\right]_j, \end{aligned}$$

so, by the previous relation proved, (2.12), we obtain

$$P_i = P_0 \cdot P^i.$$

$\square$

Example 2.11. Assume the same situation as before, except for Monday the forecast is 80% chance of rain. Make forecasts for Wednesday and Friday.

Solution. What is different from the previous situation? The transition probability matrices P and $P^{(h)} = P^h$ are the same. What changes is the *initial* situation. Now, a sunny Monday (state 1) is *also possible* and the pdf of X_0 is

$$X_0 \begin{pmatrix} 1 & 2 \\ 0.2 & 0.8 \end{pmatrix}, \quad P_0 = [0.2 \quad 0.8].$$

So, for Wednesday ($t = 2$), we have

$$P_2 = P_0 \cdot P^{(2)} = P_0 \cdot P^2 = [0.2 \quad 0.8] \begin{bmatrix} 0.61 & 0.39 \\ 0.52 & 0.48 \end{bmatrix} = [0.538 \quad 0.462],$$

that means 53.8% chance of sun and 46.2% chance of rain.

For Friday, four days after Monday (so, at $t = 4$), we have

$$P_4 = P_0 \cdot P^{(4)} \;=\; P_0 \cdot P^4 \;=\; [0.2 \quad 0.8] \begin{bmatrix} 0.5749 & 0.4251 \\ 0.5668 & 0.4332 \end{bmatrix} \;=\; [0.5684 \quad 0.4316],$$

i.e. 56.84% chance of sun and 43.16% chance of rain.

■

Remark 2.12. Notice that in matrices P and $P^{(h)} (= P^h)$, the sum of all the probabilities on each row is 1. That is because from each state, a Markov chain makes a transition to *one and only one* state, i.e. state destinations are mutually exclusive and exhaustive events, thus forming a partition. Such matrices are called **stochastic**. **Caution!** In general, this property does not hold for column totals. Some states may be "more favorable" than others, then they are visited more often than others, thus their column total will be larger. In our weather example, that is the case for the state "sunny".

2.2 Simulation of Markov Chains

Many important characteristics of stochastic processes require lengthy complex computations. Thus, it is preferable to estimate them by means of Monte Carlo methods.

For Markov chains, to predict its future behavior, all that is required is the distribution of X_0, i.e. P_0 (the initial situation) and the pattern of change at each step, i.e. the transition probability matrix P.

Once X_0 is generated, it takes some value $X_0 = i$ (according to its pdf P_0). Then, at the next step, X_1 is a discrete random variable taking the values $j, j = 1, \ldots, n$ with probabilities p_{ij} from row i of the matrix P. Its pdf will be

$$X_1 \begin{pmatrix} 1 & 2 & \ldots & n \\ p_{i1} & p_{i2} & \ldots & p_{in} \end{pmatrix}$$

The next steps are simulated similarly.
Since, at each step, the generation of a discrete random variable is needed, we can use the algorithm that simulates an arbitrary discrete distribution, Algorithm 2.6 in Lecture 3.

Algorithm 2.13.
1. Given:

$N_M =$ sample path size (length of Markov chain),

$P_0 = [P_0(1) \ \ \ldots \ \ P_0(n)]$,

$P = [p_{ij}]_{i,j=\overline{1,n}}$.

2. Generate X_0 from its pdf P_0.

3. Transition: if $X_t = i$, generate X_{t+1}, with probabilities $p_{ij}, j = \overline{1,n}$ (i.e. the i^{th} row of P), using Algorithm 2.6 (L3).

4. Return to step 3 until a Markov chain of length N_M is generated.