# Summary: Stochastic Systems - Estimation, Identification, and Adaptive Control

# Chapter 4

# Controlled Markov Chain Model

In the case of complete observations and feedback laws depending only on the current state, then we can call that the state process is a Markov chain.

## 4.1 An Example

The state takes values in $R^n$, but in many situations it is more appropriate to permit the state to take on only a finite number of values. Consider a machine whose condition at time $k$ is described by the state $x_k$ which can take the values 1 or 2 with the interpretation that $x_k = 1$ or $x_k = 2$ depending on whether the machine is in an operational or failed condition. For the moment there are no control actions allowed so that the machine behavior is autonomous. Suppose the machine is operational at time $k$, so $x_k = 1$, and there is a probability $q > 0$ that it will fail in the next period, so $x_{k+1} = 2$; with probability $1 - q$ it will continue to remain operational, so $x_{k+1} = 1$. Suppose further that $q$ does not depend upon previous values $x_{k-1}, \dots, x_0$. Finally, suppose that a failed machine continues to remain failed, so that $x_{k+1} = 2$ with probability 1, if $x_k = 2$. Then $\{x_k, k > 0\}$ is a Markov chain whose transition probabilities are described by the matrix $\boldsymbol{P} = \{\boldsymbol{P}_{ij}\}$:

$$\boldsymbol{P} = \begin{bmatrix} 1-q & q \\ 0 & 1 \end{bmatrix}. \tag{4.1}$$

The transition probability matrix $\boldsymbol{P}$ has the property that all its elements are non-negative and the sum of the elements in every row is 1. Such a matrix is said to be a *stochastic matrix*. The Markov property is expressed by

$$Prob\{x_{k+1} = j \,|\, x_k = i, x_{k-1}, \dots, x_0\} = \boldsymbol{P}_{ij}, \;\; i, j \in \{1, 2\}. \tag{4.2}$$

**Control Actions**: Let $\boldsymbol{u}_k^1$ denote the intensity of machine use at time $k$. It takes on values $\boldsymbol{u}_k^1 = 0, 1$ or 2 accordingly as the machine is not used, is in light use, or is in heavy use. Suppose that the greater the intensity of use, the larger is the likelihood of machine failure. Let $\boldsymbol{u}_k^2$ denote the intensity of machine maintenance effort. Suppose it takes only two values 0 or 1, the higher value denoting greater maintenance. The idea is that maintenance reduces the likelihood of machine failure and permits a failed machine to become operational.

The effects of these two control actions, intensity of machine use and maintenance, can be modeled as a controlled transition probability as follows. Let $\boldsymbol{u}_k \doteq (\boldsymbol{u}_k^1, \boldsymbol{u}_k^2)$. Then

$$\begin{aligned}
&Prob\{x_{k+1} = 2 \,|\, x_k = 1, x_{k-1}, \dots, \boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \dots\} = q_1(\boldsymbol{u}_k^1) - q_2(\boldsymbol{u}_k^2)\\
&Prob\{x_{k+1} = 1 \,|\, x_k = 1, x_{k-1}, \dots, \boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \dots\} = 1 - [q_1(\boldsymbol{u}_k^1) - q_2(\boldsymbol{u}_k^2)]\\
&Prob\{x_{k+1} = 1 \,|\, x_k = 2, x_{k-1}, \dots, \boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \dots\} = q_2(\boldsymbol{u}_k^2)\\
&Prob\{x_{k+1} = 2 \,|\, x_k = 2, x_{k-1}, \dots, \boldsymbol{u}_k, \boldsymbol{u}_{k-1}, \dots\} = 1 - q_2(\boldsymbol{u}_k^2).
\end{aligned} \tag{4.3}$$

These transition probabilities can be put in matrix form similar to 4.1, except that they will be functions of the control $\boldsymbol{u}$:

$$\boldsymbol{P}(\boldsymbol{u^1}, \boldsymbol{u^2}) = \begin{bmatrix} 1 - q_1(\boldsymbol{u}^1) + q_2(\boldsymbol{u}^2) & q_1(\boldsymbol{u}^1) - q_2(\boldsymbol{u}^2) \\ q_2(\boldsymbol{u}^2) & 1 - q_2(\boldsymbol{u}^2) \end{bmatrix}. \tag{4.4}$$

Equation 4.3 is illustrated in the state transition diagram below.

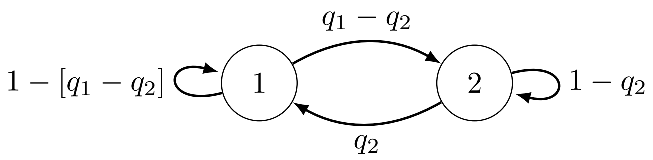

Figure 4.1: The state transition diagram of a controlled Markov process.

Of course the values of $q$ are such that $q_1(0) < q_1(1) < q_1(2)$ and $q_2(0) < q_2(1)$ because a lightly used or better maintained machine is less likely to fail than a heavily used or less well-maintained machine. We also expect that the last two probabilities in 4.3 should not depend on $\boldsymbol{u}_k^1$ because when the machine has failed, it cannot be used.

Suppose the state is observed and consider a feedback policy $\{\boldsymbol{g}_0, \boldsymbol{g}_1, \dots\}$ which is time-invariant, that is, $\boldsymbol{g}_k \equiv \boldsymbol{g}$, and let $\boldsymbol{u}_k = \boldsymbol{g}(x_k)$. This results in the transition probability matrix $\boldsymbol{P^g} = \{\boldsymbol{P}_{ij}^g\}$ where

$$\boldsymbol{P}_{ij}^{\boldsymbol{g}} \doteq \boldsymbol{P}_{ij}\big(\boldsymbol{g}(i)\big), \;\; i, j \in \{1, 2\}. \tag{4.5}$$

For example, if $\boldsymbol{g}(1) = (2, 0)$, i.e the machine is in heavy use and smaller maintenance is required, and $\boldsymbol{g}(2) = (0, 1)$, the machine is not used and greater maintenance is required. Then

$$\boldsymbol{P^g} = \begin{bmatrix} 1 - q_1(2) + q_2(0) & q_1(2) - q_2(0) \\ q_2(1) & 1 - q_2(1) \end{bmatrix}. \tag{4.6}$$

The resulting process $\{x_k\}$ is a Markov chain with stationary transition probability $\boldsymbol{P^g}$. The joint probability distribution of $x_k$ can be written as the row vector

$$\boldsymbol{p}_k \doteq (Prob\{x_k = 1\}, Prob\{x_k = 2\}).$$

By the Markov property 4.3

$$\boldsymbol{p}_{k+m} \doteq \boldsymbol{p}_k [\boldsymbol{P^g}]^m, \;\; m \geq 0, \tag{4.7}$$

and, in particular,

$$\boldsymbol{p}_k \doteq \boldsymbol{p}_0 [\boldsymbol{P^g}]^k, \tag{4.8}$$

where $\boldsymbol{p}_0$ is the initial distribution of $x_0$.

Often, as $k \to \infty$, $\boldsymbol{p}_k$ converges to a probability distribution $\boldsymbol{p} = (\boldsymbol{p}(1), \boldsymbol{p}(2))$ that does not depend on the initial distribution $\boldsymbol{p}_0$. We then say that it is an *ergodic* chain. The limiting probability distribution is called the *steady-state* or *equilibrium* or *invariant* distribution. It is the solution of the following linear equations

$$\begin{aligned} \boldsymbol{p} &= \boldsymbol{p}\boldsymbol{P}^{\boldsymbol{g}}, \\ \boldsymbol{p}(1) &+ \boldsymbol{p}(2) = 1. \end{aligned} \tag{4.9}$$

Equation 4.9 always has a solution. In the ergodic case the solution is unique and the limiting distribution $\boldsymbol{p} = \big(\boldsymbol{p}(1), \boldsymbol{p}(2)\big)$ has the following interpretation:

$$\boldsymbol{p}(i) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} I(x_k = i) \text{ with probability } 1, \tag{4.10}$$

where $I$ is the *indicator function*—i.e., $I(x_k = i) = 1$ if $x_k = i$, and $I(x_k = i) = 0$ if $x_k \neq i$. Thus $\boldsymbol{p}(1)$ is the average proportion of time that the machine is operational and $\boldsymbol{p}(2)$ the average proportion of time it spends in the failed state.

From 4.9 it is evident that the steady state probability $\boldsymbol{p}$ depends on the feedback law $g$. So by changing the policy $g$, that is, by changing the use and maintenance of the machine, we can alter the number of times it fails. One may then ask the following question: Which policy $\boldsymbol{g}$ leads to the best p?

## 4.2 Finite State Controlled Markov Chains

The preceding example generalizes to the case of an arbitrary finite state controlled Markov chain whose state $x_k$ takes values in $\{1, \ldots, I\}$. The control $\boldsymbol{u}_k$ takes values in a pre-specified set $\boldsymbol{U}$. $\boldsymbol{U}$ may be finite or infinite. The transition probabilities are specified by the $I \times I$ matrix valued function on $\boldsymbol{U}$,

$$\boldsymbol{u} \to \boldsymbol{P}(\boldsymbol{u}) \doteq \{\boldsymbol{P}_{ij}(\boldsymbol{u}),\ 1 \leq i, j \leq I\} \tag{4.11}$$

with the interpretation that

$$Prob\{x_{k+1} = j \,|\, x_k = i, x_{k-1}, \ldots, x_0, \boldsymbol{u}_k, \ldots, \boldsymbol{u}_0\} = \boldsymbol{P}_{ij}(\boldsymbol{u}_k). \tag{4.12}$$

The matrix $\boldsymbol{P}(\boldsymbol{u})$ has the property that every element is non-negative, and the sum of the elements in every row is 1 - i.e., it is a stochastic matrix. We must also specify the probability distribution of the initial state $x_0$. In case the observation $y_k \not\equiv x_k$, one must also specify the observation probability

$$P(y|i) \doteq Prob\{y_k = y \,|\, x_k = i\}. \tag{4.13}$$

## 4.3 Complete Observations and Markov Policies

Consider the controlled Markov chain model 4.11. Suppose the state is observed, $y_k \equiv x_k$. Let $\boldsymbol{g} = \{\boldsymbol{g}_0, \boldsymbol{g}_1, \ldots\}$ be a feedback policy such that $\boldsymbol{g}_k$ depends only on the current state

$x_k$ (and not on $x_{k-1}, x_{k-2}, \ldots$). We call such a $\boldsymbol{g}$ a *Markov policy*.

Let $\boldsymbol{g}$ be a Markov policy, and let $\{x_k\}$ be the resulting state process. Denote the probability distribution of $x_k$ by the $I$-dimensional row vector

$$\boldsymbol{p}_k^{\boldsymbol{g}} \doteq (Prob\{x_k = 1\}, \ldots, Prob\{x_k = I\}), \tag{4.14}$$

the superscript $\boldsymbol{g}$ emphasizes the dependence on $\boldsymbol{g}$. However, in the sequel we drop the superscript since the $\boldsymbol{g}$-dependence will be clear from the context.

*Lemma 4.1*: When a Markov policy $\boldsymbol{g}$ is employed, the resulting state process $\{x_k\}$ is a Markov process. Its one-step transition probability at time $k$ is given by the matrix

$$\boldsymbol{P}_k^{\boldsymbol{g}} \doteq \{(\boldsymbol{P}_k^{\boldsymbol{g}})_{ij} \doteq \boldsymbol{P}_{ij}\big(\boldsymbol{g}_k(i)\big),\ 1 \leq i, j \leq I)\},$$

Its $m$-step transition probability at time $k$ is given by the matrix

$$\boldsymbol{P}_k^{\boldsymbol{g}} \times \ldots \times \boldsymbol{P}_{k+m-1}^{\boldsymbol{g}},$$

so its $ij$-th element is the probability that the state will be $j$ at time $k+m$ given that it is $i$ at time $k$. Hence

$$\boldsymbol{p}_{k+m} = \boldsymbol{p}_k(\boldsymbol{P}_k^{\boldsymbol{g}} \times \ldots \times \boldsymbol{P}_{k+m-1}^{\boldsymbol{g}}).$$

In particular,

$$\boldsymbol{p}_k = \boldsymbol{p}_0(\boldsymbol{P}_0^{g} \times \ldots \times \boldsymbol{P}_{k-1}^{g}), \tag{4.15}$$

where $\boldsymbol{p}_0$ is the probability distribution of the initial state $x_0$.

*Proof*: The proof is immediate from the Markov property 4.12. Since the transition probability matrix $\boldsymbol{P}_k^{\boldsymbol{g}}$ depends on the time $k$, we say that $\{x_k\}$ is a Markov chain with nonstationary transition probability.

## 4.4 The Cost of a Markov Policy

A Markov policy $\boldsymbol{g}$ determines the probability distribution of the state process $\{x_k\}$ and the control process $\{\boldsymbol{u}_k = \boldsymbol{g}_k(x_k)\}$. Different policies will lead to different probability distributions. In optimal control problems one is interested in finding the best or optimal policy. To do this one needs to compare different policies. This is done by specifying a *cost function*. This is a sequence of real valued functions of the state and control,

$$c_k(i, \boldsymbol{u}),\ 1 \leq i \leq I,\ \boldsymbol{u} \in \boldsymbol{U},\ k \geq 0.$$

The interpretation is that $c_k(i, \boldsymbol{u})$ is the cost to be paid if at time $k$, $x_k = i$ and $\boldsymbol{u}_k = \boldsymbol{u}$. Fix a Markov policy $\boldsymbol{g}$. The cost incurred by $\boldsymbol{g}$ up to the time horizon $N$ is $\sum_{k=0}^{N} c_k(x_k, \boldsymbol{u}_k)$. This is a random variable since $x_k$ and $\boldsymbol{u}_k$ are random. Hence the expected cost is

$$J(\boldsymbol{g}) \doteq \mathbb{E}^{\boldsymbol{g}}\left[\sum_{k=0}^{N} c_k(x_k, \boldsymbol{u}_k)\right] = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{k=0}^{N} c_k\big(x_k, \boldsymbol{g}_k(x_k)\big)\right], \tag{4.16}$$

here $\mathbb{E}^{\boldsymbol{g}}$ denotes expectation with respect to the probability distribution of $\{x_k\}$, $\{\boldsymbol{u}_k\}$ determined by $\boldsymbol{g}$. $J(\boldsymbol{g})$ can be readily evaluated in terms of the transition probability matrices $\boldsymbol{P}_k^{\boldsymbol{g}}$ as follows. From 4.16 it can be obtained

$$\begin{aligned} J(\boldsymbol{g}) &= \sum_{k=0}^{N} \sum_{i=1}^{I} Prob\{x_k = i\} c_k\big(i, \boldsymbol{g}_k(i)\big) \\ &= \sum_{k=0}^{N} \boldsymbol{p}_k\, \boldsymbol{c}_k^{\boldsymbol{g}} = \sum_{k=0}^{N} \boldsymbol{p}_0\, (\boldsymbol{P}_0^{\boldsymbol{g}} \times \,\ldots\, \times \boldsymbol{P}_{k-1}^{\boldsymbol{g}}) \boldsymbol{c}_k^{\boldsymbol{g}}, \end{aligned} \tag{4.17}$$

where $\boldsymbol{c}_k^{\boldsymbol{g}}$ is the $I$-dimensional column vector

$$\boldsymbol{c}_k^{\boldsymbol{g}} \doteq \Big(c_k\big(1, \boldsymbol{g}_k(1)\big), \,\ldots,\, c_k\big(I, \boldsymbol{g}_k(I)\big)\Big)^T. \tag{4.18}$$

The last equality in 4.17 follows from 4.15. Thus the best Markov policy $\boldsymbol{g}$ is the one that minimizes $J(\boldsymbol{g}) = \sum_{k=0}^{N} \boldsymbol{p}_0\, (\boldsymbol{P}_0^{\boldsymbol{g}} \times \,\ldots\, \times \boldsymbol{P}_{k-1}^{g}) \boldsymbol{c}_k^{g}$. A *Dynamic Programming* can be an approach for computing the best $\boldsymbol{g}$. Central to dynamic programming is a recursive technique for calculating the cost of a Markov policy $\boldsymbol{g}$. Since the technique depends only on the fact that the state process corresponding to $\boldsymbol{g}$ is Markov, which is introduced here. For each time $1 \leq k \leq N$, and state $1 \leq i \leq I$, let $V_k^{\boldsymbol{g}}(i)$ denote the expected cost incurred during $k, \,\ldots,\, N$ when $x_k = i$. That is,

$$V_k^{\boldsymbol{g}}(i) \doteq \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{l=k}^{N} c_l\big(x_l, \boldsymbol{g}_l(x_l)\big) \,\middle|\, x_k = i \right]. \tag{4.19}$$

Observe that with this notation the total cost 4.17 is

$$J(\boldsymbol{g}) = \sum_{i=1}^{I} (\boldsymbol{p}_0)_i\, V_0^{\boldsymbol{g}}(i). \tag{4.20}$$

*Lemma 4.2*: The functions $V_k^{\boldsymbol{g}}(i)$ can be calculated by backward recursion,

$$V_k^{\boldsymbol{g}}(i) = c_k\big(i, \boldsymbol{g}_k(i)\big) + \sum_{j=1}^{I} \left(\boldsymbol{P}_{\boldsymbol{k}}^{\boldsymbol{g}}\right)_{ij} V_{k+1}^{\boldsymbol{g}}(j), \;\; 0 \leq k < N. \tag{4.21}$$

starting with the final condition

$$V_N^{\boldsymbol{g}}(i) = c_N\big(i, \boldsymbol{g}_N(i)\big). \tag{4.22}$$

*Proof*: From the definition we immediately get 4.22. Next

$$\begin{aligned} V_k^{\boldsymbol{g}}(i) &= \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{l=k}^{N} c_l\big(x_l, \boldsymbol{g}_l(x_l)\big) \,\middle|\, x_k = i \right] \\ &= c_k\big(i, \boldsymbol{g}_k(i)\big) + \mathbb{E}^{\boldsymbol{g}} \left[ \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{l=k+1}^{N} c_l\big(x_l, \boldsymbol{g}_l(x_l)\big) \,|\, x_{k+1},\, x_k = i \right] \,\middle|\, x_k = i \right] \\ &= c_k\big(i, \boldsymbol{g}_k(i)\big) + \mathbb{E}^{\boldsymbol{g}} \left[ V_{k+1}^{\boldsymbol{g}}(x_{k+1}) \,|\, x_k = i \right] \;\; \text{(by 4.12)} \end{aligned}$$

$$= c_k\big(i, \boldsymbol{g}_k(i)\big) + \sum_{j=1}^{I} V_{k+1}^{\boldsymbol{g}}(j)\; Prob\{x_{k+1} = j \,|\, x_k = i\}$$

which is 4.19 once we recall the definition of $(\boldsymbol{P}_k^{\boldsymbol{g}})_{ij}$.

The previous equations can be expressed in a convenient vector notation. Denote the $I$-dimensional column vector

$$V_k^{\boldsymbol{g}} \doteq \big(V_k^{\boldsymbol{g}}(1),\, \ldots,\, V_k^{\boldsymbol{g}}(I)\big)^T. \tag{4.23}$$

Then, using 4.18, we can express 4.21, 4.22 and 4.20, respectively, as

$$V_k^{\boldsymbol{g}} = \boldsymbol{c}_k^{\boldsymbol{g}} + \boldsymbol{P}_k^{\boldsymbol{g}}\, V_{k+1}^{\boldsymbol{g}},\ \ 0 \le k < N, \tag{4.24}$$
$$V_N^{\boldsymbol{g}} = \boldsymbol{c}_N^{\boldsymbol{g}}, \tag{4.25}$$
$$J(\boldsymbol{g}) = \boldsymbol{p}_0\, V_0^{\boldsymbol{g}}. \tag{4.26}$$

Here, the time horizon $N$ is finite. Often one is interested in the infinite horizon. This is not an immediate extension, since if one simply sets $N = \infty$ in 4.16, in most cases one gets $J(\boldsymbol{g}) = \infty$ for every $\boldsymbol{g}$. The notion of best $\boldsymbol{g}$ then becomes meaningless. **There are two ways to treat the infinite horizon problem.** The first approach is to introduce a discount factor $\beta$, $0 < \beta < 1$, and to consider the expected discounted cost

$$J(\boldsymbol{g}) = \mathbb{E}^{\boldsymbol{g}}\left[\sum_{k=0}^{\infty} \beta^k\, c_k(x_k, \boldsymbol{u}_k)\right].$$

Observe that if $c_k$ is bounded, then $J(\boldsymbol{g})$ will be finite. Since the cost incurred at time $k$ is weighted by $\beta^k$, present costs are more important than future costs. In an economic context, $\beta = (1+r)^{-1}$, where $r > 0$ is the interest rate. With this interpretation, $J(\boldsymbol{g})$ is the present value of the cost. From 4.17 it follows that

$$J(\boldsymbol{g}) = \sum_{k=0}^{\infty} \beta^k\, \boldsymbol{p}_0\, (\boldsymbol{P}_0^{g} \times \cdots \times \boldsymbol{P}_{k-1}^{g})\, \boldsymbol{c}_k^{\boldsymbol{g}}.$$

Define

$$V_k^{\boldsymbol{g}}(i) \doteq \mathbb{E}^{\boldsymbol{g}}\left[\sum_{l=k}^{\infty} \beta^l\, c_l(x_l, \boldsymbol{g}_l(x_l)) \,\big|\, x_k = i\right];$$

then, using the notation 4.18, the counterparts of 4.24 and 4.26 are

$$V_k^{\boldsymbol{g}} = \beta^k\, \boldsymbol{c}_k^{\boldsymbol{g}} + \boldsymbol{P}_k^{\boldsymbol{g}}\, V_{k+1}^{\boldsymbol{g}},\ \ k \ge 0, \tag{4.27}$$
$$J(\boldsymbol{g}) = \boldsymbol{p}_0\, V_0^{\boldsymbol{g}}. \tag{4.28}$$

However, in the infinite horizon problem, there is no counterpart of the final condition 4.25. The second approach is followed when discounting is inappropriate. A policy is then evaluated according to its average cost per unit time,

$$J(\boldsymbol{g}) = \lim_{N\to\infty} \frac{1}{N}\, \mathbb{E}^{\boldsymbol{g}}\left[\sum_{k=0}^{N-1} c_k\left(x_k, \boldsymbol{g}_k(x_k)\right)\right]. \tag{4.29}$$

Using 4.17 this cost equals

$$J(\boldsymbol{g}) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \boldsymbol{p}_0 \, (\boldsymbol{P}_0^g \times \cdots \times \boldsymbol{P}_{k-1}^g) \, \boldsymbol{c}_k^{\boldsymbol{g}}. \tag{4.30}$$

From this expression we see that if $\boldsymbol{P}_k^g$ varies with $k$, then the limit above need not exist. If the transition matrix does not depend on $k$, then the limit always exists. So far, the total cost 4.16 is the sum of the costs incurred in each time period. The next exercise shows that other cost functions can be put into this additive form.

---

*Exercise 4.1*: Suppose the cost incurred by a Markov policy $\boldsymbol{g} \doteq \{\boldsymbol{g}_0, \ldots, \boldsymbol{g}_N\}$ is

$$J(\boldsymbol{g}) \doteq \mathbb{E}^g \left[ I \left( \max_{0\le k\le N} h(x_k) \ge \alpha \right) \right], \tag{4.31}$$

where $I$ is the indicator function, and $h$ and $\alpha$ are specified early. [Thus $J(\boldsymbol{g})$ is the probability that $h(x_k)$ exceeds $\alpha$ at some time $k$.] Show that 4.31 can be put into the additive form 4.16.
[Hint: Define the new chain $\boldsymbol{z}_k \doteq (x_k, y_k)$, with $y_k \in 0, 1$. The transition probability of $x_k$ is exactly as before, whereas

$$\begin{aligned} &Prob\{y_{k+1} = 1 \,|\, y_k = 0,\, x_k,\, \boldsymbol{z}_{k-1},\, \ldots\, \boldsymbol{z}_0,\, \boldsymbol{u}_k,\, \ldots\, \boldsymbol{u}_0\} = \begin{cases} 1, & \text{if } h(x_k) \ge \alpha \\ 0, & \text{if } h(x_k) < \alpha \end{cases} \\ &Prob\{y_{k+1} = 1 \,|\, y_k = 1,\, x_k,\, \boldsymbol{z}_{k-1},\, \ldots\, \boldsymbol{z}_0,\, \boldsymbol{u}_k,\, \ldots\, \boldsymbol{u}_0\} = 1. \end{aligned}$$

Now let $c_k(\boldsymbol{z}_k, \boldsymbol{u}_k) = 0$ for $k < N$, and $c_N(x, y, \boldsymbol{u}) = y$.]

---

## 4.5 Stationary Markov Policy

A Markov policy $\boldsymbol{g} = \boldsymbol{g}_0, \boldsymbol{g}_1, \ldots$ is *stationary* or *time-invariant* if $\boldsymbol{g}_0 = \boldsymbol{g}_1 = \cdots = \boldsymbol{g}$, with a slight abuse of notation. Let $g$ be stationary; then the transition probability matrix is stationary, $P_k^g = P^g$. Suppose the cost functions are also time-invariant, $c_k = c$. Fix a discount $0 < \beta < 1$. Then

$$\begin{aligned} V_k^{\boldsymbol{g}}(i) =& \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{l=k}^{\infty} \beta^l \, c_l(x_l, \boldsymbol{g}_l(x_l)) \,\Big|\, x_k = i \right] \\ =& \beta^k \, \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{l=0}^{\infty} \beta^l \, c_l(x_l, \boldsymbol{g}_l(x_l)) \,\Big|\, x_0 = i \right] \\ =& \beta^k \, V_0^g(i). \end{aligned}$$

Using this in 4.27 gives

$$\beta^k \, V_0^g = \beta^k \, \boldsymbol{c}^{\boldsymbol{g}} + \beta^{k+1} \, \boldsymbol{P}^{\boldsymbol{g}} \, V_0^g,$$

or, in matrix notation,

$$[\boldsymbol{I} - \beta \, \boldsymbol{P}^{\boldsymbol{g}}] \; V_0^g = \boldsymbol{c}^g.$$

This is a set of $I$ linear equations in the $I$ unknowns $V_0^g(i)$, for $i = 1, 2, \ldots, I$. $[\boldsymbol{I} - \beta\, \boldsymbol{P}^{\boldsymbol{g}}]$ is invertible so that this system of linear equations has a unique solution $V_0^g$. Next the average cost 4.29 and 4.30 are studied when $\boldsymbol{g}$ and $\boldsymbol{c}$ are stationary. The next three lemmas are stated without proof.

*Lemma 4.3*: If $\boldsymbol{P}$ is a transition probability matrix, then the *Cesaro limit*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} \boldsymbol{P}^k \doteq \boldsymbol{\Pi},$$

always exists. The matrix $\boldsymbol{\Pi}$ is a stochastic matrix and it satisfies the equation

$$\boldsymbol{\Pi} = \boldsymbol{\Pi}\, \boldsymbol{P}.$$

Thus for stationary $g$ and time-invariant cost, the average cost per unit time 4.29 and 4.30, is

$$J(\boldsymbol{g}) = \lim_{N\to\infty} \frac{1}{N}\, \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{k=0}^{N-1} c\left(x_k, \boldsymbol{g}_k(x_k)\right) \right] = \boldsymbol{p}_0\, \boldsymbol{\Pi}\, \boldsymbol{c}^g. \tag{4.32}$$

Let $\boldsymbol{\pi}$ be one of the rows of $\boldsymbol{\Pi}$. Then

$$\boldsymbol{\pi} = \boldsymbol{\pi}\, \boldsymbol{P}.$$

Moreover since $\boldsymbol{\Pi}$ is a stochastic matrix, $\boldsymbol{\pi}$ can be regarded as a probability distribution. This has the interpretation that if the Markov chain has initial probability distribution given by $\boldsymbol{\pi}$, then the probability distribution of the state remains at $\boldsymbol{\pi}$ for all time. Thus it is said to be an *invariant probability distribution*. Clearly if the rows of $\boldsymbol{\Pi}$ are not all the same, then there is more than one invariant probability distribution.

---

*Exercise 4.2*: Give an example such that $J(\boldsymbol{g})$ in 4.32 depends on $\boldsymbol{p}_0$.
[Hint: Choose $\boldsymbol{P} = \boldsymbol{I}$. Note that there are several invariant probability distributions.]

---

In many cases 4.32 is independent of the initial distribution $\boldsymbol{p}_0$. An $I \times I$ transition probability matrix $\boldsymbol{P}$ is *reducible* or *decomposable* if there is a renumbering of the states $\{1, , \ldots, I\}$ for which $\boldsymbol{P}$ takes the form

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \\ 0 & \boldsymbol{P}_3 \end{bmatrix},$$

where $\boldsymbol{P}_1$, and $\boldsymbol{P}_3$ are square matrices. This means that it is not possible to make a transition from a state indexing a row of $\boldsymbol{P}_3$ to a state corresponding to $\boldsymbol{P}_1$. Hence if the initial state happens to lie in the set of states indexing rows of $\boldsymbol{P}_3$, then the Markov chain stays forever in the same set, with transition probabilities given by $\boldsymbol{P}_3$. Thus there is a Markov chain with this smaller state space. A transition matrix $\boldsymbol{P}$ which cannot be put in above-mentioned form by any renumbering of states is called *irreducible* or *indecomposable*.

*Lemma 4.4*: If $\boldsymbol{P}$ is an irreducible transition probability matrix, then there is a unique row vector $\boldsymbol{\pi}$ such that

$$\boldsymbol{\pi}\, \boldsymbol{P} = \boldsymbol{\pi}, \; \sum_{i=1}^{I} \boldsymbol{\pi}_i = 1.$$

Moreover $\boldsymbol{\pi}_i > 0$, all $i$. Finally, the matrix $\mathbf{\Pi}$ in 4.32 has all rows equal to $\boldsymbol{\pi}$. [$\boldsymbol{\pi}$ is called the steady state or invariant probability distribution of the Markov chain $\{x_k\}$.]

---

*Exercise 4.3*: Construct an example of a Markov chain that is reducible and that has several different invariant probability distributions. Is every component $\boldsymbol{\pi}_i > 0$ for every invariant probability distribution $\boldsymbol{\pi}$? [Hint: See previous exercise.]

---

*Exercise 4.4*: Show that $\boldsymbol{P}$ is irreducible if and only if for every $i$ and $j$ there is a sequence of states $i \doteq i_0,\, i_1,\, \ldots,\, i_{k-1},\, i_k \doteq j$ such that $\boldsymbol{P}i_l|i_{l+1} > 0$ for $l = 0,\, 1,\, \ldots, k-1$. Hence there is a path in the state space from every state to every other state that can be traversed by the Markov chain with positive probability. [Hint: If not, then group all the states which cannot be reached from a certain state into one set. Identify these states with the matrix $\boldsymbol{P}_1$.]

Thus if $\boldsymbol{P}^{\boldsymbol{g}}$ is irreducible, then

$$J(\boldsymbol{g}) = \boldsymbol{p}_0\, \mathbf{\Pi}\, \boldsymbol{c}^{\boldsymbol{g}} = \boldsymbol{\pi}\, \boldsymbol{c}^{\boldsymbol{g}} \tag{4.33}$$

since all the rows of $\mathbf{\Pi}$ are identical. The cost $J(\boldsymbol{g})$ is therefore independent of the initial distribution. A probability transition matrix $\boldsymbol{P}$ is said to be periodic if there is a renumbering of the states for which $\boldsymbol{P}$ takes the form

$$\boldsymbol{P} = \begin{bmatrix} \mathbf{0} & \boldsymbol{P}_1 & \mathbf{0} & \cdot & \cdot & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{P}_2 & \cdot & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdot & \cdot & \boldsymbol{P}_{n-1} \\ \boldsymbol{P}_n & \mathbf{0} & \mathbf{0} & \cdot & \cdot & \mathbf{0} \end{bmatrix}.$$

This means that one can partition the states into disjoint subsets $I_1,\, \ldots,\, I_n$ such that from states in $I_{m-1}$ transitions are possible only to states in $I_m$. If $\boldsymbol{P}$ cannot be put into this form it is said to be *aperiodic*.

---

*Lemma 4.5*: If $\boldsymbol{P}$ is irreducible and aperiodic, then

$$\lim_{k \to \infty} \boldsymbol{P}^k = \mathbf{\Pi},$$

where $\mathbf{\Pi}$ is the matrix with all rows equal to $\boldsymbol{\pi}$.

---

*Exercise 4.5*: Show that if $\boldsymbol{P}$ is periodic, then $\boldsymbol{P}_k$ cannot converge.

There is an interesting case of a cost over the infinite time horizon which requires neither a discount factor nor consideration of the average cost per unit time.

---

*Exercise 4.6*: Let $F \subset \{1,\, ,\ldots,\, I\}$ be a subset of states, and for a stationary Markov policy $\boldsymbol{g}$ let $\tau$ be the first time that $x_k$ enters $F$. That is

$$\tau = \begin{cases} \min\{k \geq 0 \,|\, x_k \in F\} \\ \infty, \;\text{ if } x_k \notin F \text{ for all } k. \end{cases}$$

Suppose

$$\mathbb{E}\{\tau \,|\, x_0 = i\} < \infty \text{ for all } i.$$

Let $c(i, \boldsymbol{u})$ be a stationary cost function, and define

$$V^{\boldsymbol{g}}(i) \doteq \mathbb{E}^{\boldsymbol{g}}\left[\sum_{k=0}^{\tau-1} \beta^l \, c(x_k, \boldsymbol{g}(x_k)) \,\Big|\, x_0 = i\right].$$

Show that $V^{\boldsymbol{g}}(i) < \infty$ and

$$V^{\boldsymbol{g}} = \boldsymbol{c}^{\boldsymbol{g}} + \boldsymbol{R}^{\boldsymbol{g}} \, V^{\boldsymbol{g}},$$

where $(\boldsymbol{R}^{\boldsymbol{g}})_{ij} \doteq (\boldsymbol{R}^{\boldsymbol{g}})_{ij}$ if $i \notin F$, and $(\boldsymbol{R}^{\boldsymbol{g}})_{ij} \doteq 0$ if $i \in F$. [The random time $\tau$ is called a stopping time, and problems of this type are called *stopping time problems*.]

---

## 4.6 Infinite State Markov Chains

The discussion in Sections 4.2, 4.3, and 4.4 carries over with obvious changes to the case of controlled Markov chains whose state $x_k$ takes values in the infinite set $\{1, 2, 3, \ldots\}$. The control takes values in $\boldsymbol{U}$, and for $\boldsymbol{u} \in \boldsymbol{U}$, the transition probabilities are specified by an infinite dimensional matrix

$$\boldsymbol{P}(\boldsymbol{u}) \doteq \{\boldsymbol{P}_{ij}(\boldsymbol{u}),\ 1 \le i, j < \infty\}.$$

with the same interpretation as 4.12

A Markov policy $\boldsymbol{g}$ is defined exactly as before, and the corresponding one-step and $m$-step transition probability matrix at time $k$ are

$$\begin{array}{ll} \text{one-step transtion probability} & : \boldsymbol{P}_k^{\boldsymbol{g}} \doteq \{((\boldsymbol{P}^{\boldsymbol{g}})_{ij}) \doteq \boldsymbol{P}_{ij}\left(\boldsymbol{g}_k(i)\right), 1 \le i,\, j < \infty\}, \\ m - \text{step transtion probability} & : \boldsymbol{P}_k^{\boldsymbol{g}} \times \cdots \times \boldsymbol{P}_{k+m-1}^{\boldsymbol{g}}. \end{array}$$

The probability distribution of $x_k$ is now the infinite row vector

$$\boldsymbol{p}_k^{\boldsymbol{g}} = \boldsymbol{p}_k \doteq (Prob\{x_k = 1\},\ Prob\{x_k = 2\},\ \ldots),$$

and by the Markov property,

$$\boldsymbol{p}_{k+m} = \boldsymbol{p}_k\, (\boldsymbol{P}_k^{\boldsymbol{g}} \times \cdots \times \boldsymbol{P}_{k+m-1}^{\boldsymbol{g}}),\ \text{and } \boldsymbol{p}_k = \boldsymbol{p}_0\, (\boldsymbol{P}_0^{\boldsymbol{g}} \times \cdots \times \boldsymbol{P}_{k-1}^{\boldsymbol{g}})$$

For cost functions $c_k(i, \boldsymbol{u})$, $1 \le i < \infty$, $\boldsymbol{u} \in \boldsymbol{U}$, the expected cost over a finite horizon is given by 4.16. If $V_k^{\boldsymbol{g}}$ is the infinite-dimensional column vector with components $V_k^{\boldsymbol{g}}(i)$ defined by 4.20, then the recursion analogous to 4.21 and 4.22 is

$$\begin{aligned} V_k^{\boldsymbol{g}}(i) &= c_k\big(i, \boldsymbol{g}_k(i)\big) + \sum_{j=1}^{\infty} (\boldsymbol{P}_k^{\boldsymbol{g}})_{ij}\, V_{k+1}^{\boldsymbol{g}}(j),\ 0 \le k < N, \\ V_k^{\boldsymbol{g}}(i) &= c_N(i, \boldsymbol{g}_N(i)). \end{aligned}$$

Similarly, the discounted cost over the infinite horizon is given by the recursion 4.27. If the policy $\boldsymbol{g}$ and the cost function $c(i, \boldsymbol{u})$ are time-invariant, then the infinite-dimensional vector $V_0^{\boldsymbol{g}}$ satisfies the linear equation

$$[\boldsymbol{I} - \beta\, \boldsymbol{P}^{\boldsymbol{g}}]\; V_0^g = \boldsymbol{c}^g. \tag{4.34}$$

---

*Exercise 4.7*: Let $\boldsymbol{l}_\infty$ be the set of all infinite-dimensional column vectors $\boldsymbol{x} = (\boldsymbol{x}_1,\, \boldsymbol{x}_1, \dots)^T$. For $\boldsymbol{x} \in \boldsymbol{l}_\infty$, define its norm $\|\boldsymbol{x}\| \doteq \sup\{|\boldsymbol{x}_1|,\, |\boldsymbol{x}_2|, \dots\}$. Show that $\|\boldsymbol{P}^{\boldsymbol{g}}\| \doteq \sup_{\boldsymbol{x} \neq \boldsymbol{0}} \frac{|\boldsymbol{P}^{\boldsymbol{g}}\, \boldsymbol{x}|}{\boldsymbol{x}} = 1$. Now show that the linear map defined on $\boldsymbol{l}_\infty$ by $\boldsymbol{x} \to [\boldsymbol{I} - \beta\, \boldsymbol{P}^{\boldsymbol{g}}]\, \boldsymbol{x}$ is invertible; in fact,

$$[\boldsymbol{I} - \beta\, \boldsymbol{P}^{\boldsymbol{g}}]^{-1}\, \boldsymbol{x} = \sum_{k=1}^{\infty} [\beta\, \boldsymbol{P}^{\boldsymbol{g}}]^k\; \boldsymbol{x}.$$

In particular, if $\boldsymbol{c}^g \in \boldsymbol{l}_\infty$ - i.e., if $\sup_i |\boldsymbol{c}^g(i)|$ is finite - then 4.34 has a unique solution.

---

Similarly, we can define the average cost per unit time as

$$\begin{aligned} J(\boldsymbol{g}) &\doteq \lim_{N\to\infty} \frac{1}{N}\, \mathbb{E}^{\boldsymbol{g}} \left[ \sum_{k=0}^{N} c\left(x_k, \boldsymbol{g}_k(x_k)\right) \right] \\ &= \boldsymbol{p}_0 \left[ \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} (\boldsymbol{P}^g)^k \right] \boldsymbol{c}^g. \end{aligned} \tag{4.35}$$

The only significant difference between the finite and infinite state cases arises at this point, because Lemmas (4.3), (4.4) and (4.5) do not hold in the infinite case.

---

*Exercise 4.8*: Consider a Markov chain with the following state transition diagram.

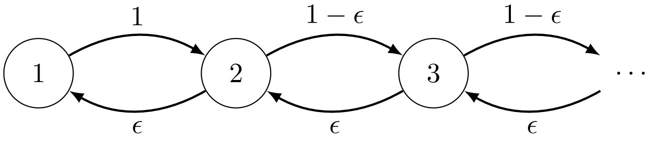

Figure 4.2: The state transition diagram of a infinite Markov process.

This gives the transition probability matrix

$$\boldsymbol{P} = \begin{bmatrix} 0 & 1 & . & . & . \\ \epsilon & 0 & 1-\epsilon & . & . \\ . & \epsilon & 0 & 1-\epsilon & . \\ . & . & \epsilon & 0 & . \\ . & . & . & \epsilon & . \\ . & . & . & . & . \end{bmatrix}.$$

Show that $\boldsymbol{P}$ is irreducible and periodic for $\epsilon > 0$. Show that there exists a steady-state distribution $\boldsymbol{\pi}$ with

$$\boldsymbol{\pi}\, \boldsymbol{P} = \boldsymbol{\pi}, \;\text{ and } \sum_{i=0}^{\infty} \boldsymbol{\pi}(i) = 1,$$

if and only if $\epsilon = \frac{1}{2}$. [Hint: First show that a positive solution to $\boldsymbol{\pi}\,\boldsymbol{P} = \boldsymbol{\pi}$ must be proportional to $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ with

$$\boldsymbol{\eta}(1) = 1, \text{ and } \boldsymbol{\eta}(i) = \frac{(1-\epsilon)^{i-2}}{\epsilon^{i-1}}, \ i > 1.$$

Now show that $\sum_i \boldsymbol{\pi}(i) = 1$ is possible if and only if $\sum_i \boldsymbol{\eta}(i) < \infty$ if and only if $\epsilon > \frac{1}{2}$.] If, however, the Cesaro limit in 4.35 exists (as it always does in the finite case),

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N} (\boldsymbol{P}^g)^k = \boldsymbol{\Pi},$$

then $J(\boldsymbol{g}) = \boldsymbol{p}_0\,\boldsymbol{\Pi}\,\boldsymbol{c}^{\boldsymbol{g}}$. If, furthermore, the chain is ergodic so that all rows of $\boldsymbol{\Pi}$ are equal to the steady state distribution $\boldsymbol{\pi}$, then the average cost is independent of the initial distribution, $J(\boldsymbol{g}) = \boldsymbol{\pi}\,\boldsymbol{c}^{\boldsymbol{g}}$.

## 4.7 Continuous Time Markov Chains

The discrete time example of Section 4.1 also makes sense if the transition of the machine state can occur at any continuous time $t$.