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 $x_{k-1}, \ldots, x_0$. 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

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

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

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

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.

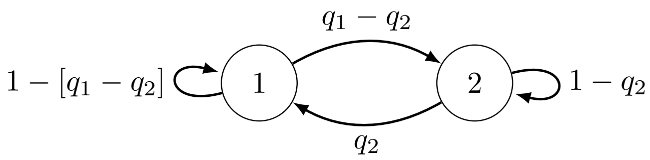

Figure 4.1: The state transition diagram of a controlled Markov process.

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 $\{1, \dots, I\}$. 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 $\boldsymbol{U}$,

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

with the interpretation that

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

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 $\boldsymbol{g} = \{\boldsymbol{g}_0, \boldsymbol{g}_1, \dots\}$ 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: 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

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

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,

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

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

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

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

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

The last equality in 4.17 follows from 4.15. Thus the best Markov policy $\boldsymbol{g}$ is the one that minimizes $J(\boldsymbol{g}) = \sum_{k=0}^{N} \boldsymbol{p}_0 \, (\boldsymbol{P}_0^{\boldsymbol{g}} \times \, \dots \, \times \boldsymbol{P}_{k-1}^{g}) \boldsymbol{c}_k^{g}$. 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 $k, \dots, N$ when $x_k = i$. That is,

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

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

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

starting with the final condition

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

Proof: