11 Stochastic Systems and Controlled Markov Chains

Stochastic System Model

- The dynamic behavior of a (discrete-time) deterministic system is usually modeled by an equation of the form $x_{t+1} = f_t(x_t, u_t), t = 0, 1, 2, \ldots$, where $x_t \in \mathbb{R}^n$ is the state and $u_t \in \mathbb{R}^m$ is the input at time t. Usually there is an output $y_t \in \mathbb{R}^p$ modeled by the equation $y_t = h_t(x_t), t = 0, 1, 2, \ldots$.

- An obvious but important property of deterministic system is that the current state x_t and the input sequence $u_t, u_{t+1}, \ldots, u_{t+m}$ determine the state x_{t+m} independently of the past values of state $x_0, \ldots, x_{t-1}$ and input $u_0, \ldots, u_{t-1}$, i.e., $x_{t+m+1} = f_{t+m+1,t}(x_t, u_t, \ldots, u_{t+m})$.

- In practice, a system may have uncertainty, including noise for the dynamic, noise for the observation and uncertainty in the initial condition. Therefore, a **stochastic system model** is an equation of the form

$$X_{t+1} = f_t(X_t, U_t, W_t), \quad Y_t = h_t(X_t, V_t), \quad t = 0, 1, 2, \ldots$$

- To make the stochastic system concrete, we need to specify
  1. the dynamic equation f_t and the observation equation h_t for each $t \geq 0$, and
  2. the joint probability distribution of the **primitive/basic random variables**

  $$X_0, W_0, W_1, \ldots, V_0, V_1, \ldots$$

  where X_0 is the initial state, $W_0, W_1, \ldots$ are the input disturbances, and $V_0, V_1, \ldots$ are the measurement noise. Usually, we assume they are mutually independent.

- Suppose that $u_0, u_1, \ldots$ are specified deterministic sequence of inputs. Then we have

$$X_1 := f_1(X_0, u_0, W_0), \quad X_2 = f_2(f_1(X_0, u_0, W_0), u_1, W_1), \ldots$$

  Therefore, X_{t+1} is a random variable depends upon the input sequence $u_{0:t} = (u_0, \ldots, u_t)$ as well as the basic random variables $X_0, W_0, \ldots, W_t$. We call the stochastic process $\{X_t\}$ the **state process**. Similarly, we have the **observation process** $\{Y_t\}$.

- Now, it remains to describe the control/action process $\{U_t\}$. In general, $\{U_t\}$ is determined by a **control strategy/control law/decision strategy**

$$g = (g_0, g_1, \ldots, g_t, \ldots), \text{ where } \quad U_t = g_t(Y_{0:t}, U_{0:t-1})$$

  Therefore, given a control strategy g, we can completely determine the state process $\{X_t^g\}$ and the observation process $\{Y_t^g\}$ by

$$\begin{aligned}
X_1^g &= f_0(X_0, U_0, W_0) = f_0(X_0, g_0(Y_0), W_0) = f_0(X_0, g_0(h_0(X_0, V_0)), W_0) = \tilde{f}_0^g(X_0, V_0, W_0) \\
Y_1^g &= h_1(X_1, V_1) = h_1(\tilde{f}_0(X_0, V_0, W_0), V_1) = \tilde{h}_1^g(X_0, V_0, W_0, V_1) \\
X_2^g &= f_1(X_1, U_1, W_1) = \tilde{f}_1^g(X_0, W_0, W_1, V_0, V_1) \\
Y_2^g &= \tilde{h}_2^g(X_0, W_0, W_1, V_0, V_1, V_2)
\end{aligned}$$

  Therefore, we conclude that $X_t^g = \tilde{f}_{t-1}^g(X_0, W_{0,t-1}, V_{0:t-1})$ and $Y_t^g = \tilde{f}_t^g(X_0, W_{0,t-1}, V_{0:t})$.

Controlled Markov Chain

- For a stochastic system, the state-space is $\mathbb{R}^n$ in general. Recall that in finite Markov chain, we assume a finite state-space $S = \{0, 1, \dots, I\}$. Furthermore, we assume process $\{X_t\}$ satisfies the Markov property, i.e.,

$$\forall t \geq 0, \forall B \in \mathcal{B}(\mathbb{R}^n) : P(X_{t+1} \in B \mid X_t = x_t, \dots, X_0 = x_0) = P(X_{t+1} \in B \mid X_t = x_t)$$

  Then under the time-homogeneous assumption, we can define the the matrix of transition probability (MOTP) $\mathbb{P} = [P_{ij}]_{i,j \in S}$, where $P_{ij} = P(X_{t+1} = j \mid X_t = i)$. If we define the state-distribution vector

$$\pi_t = (\pi_t(0), \pi_t(1), \dots, \pi_t(I)), \quad \text{where} \quad \pi_t(i) = P(X_t = i)$$

  Then the CK-Equation tells that that $\pi_{t+1} = \pi_t \mathbb{P}$.

- In a Markov chain, the MOTP $\mathbb{P}$ is invariant. In a **controlled Markov chain** (also called **Markov decision process**), we assume that the MOTP depends on the control action. Assume the action space $\mathcal{U}$ is finite, then for each $u \in \mathcal{U}$, $\mathbb{P}(u) = [P_{ij}(u)]_{i,j \in S}$ is a MOTP, where

$$P_{ij}(u) = P(X_{t+1} = j \mid X_t = i, u_t = u)$$

- Hereafter, we assume the case of **perfect observation**, i.e., $Y_t = X_t, \forall t$. Then a control strategy $g = (g_0, g_1, \dots, g_t, \dots)$, in general, is in the form of

$$U_t = g_t(X_{0:t}, U_{0:t-1}) = \tilde{g}_t(X_{0:t})$$

  Therefore, when g is fixed, we have the state process under control $\{X_t^g\}$ defined by

$$P(X_{t+1} = j \mid X_t = i, U_t = u) = P(X_{t+1} = j \mid X_t = i, U_t = \tilde{g}_t(X_{0:t}) = u)$$

  Compared with the general model of stochastic system, we do not need input disturbance W_t because this information has been captured by the MOTP $\mathbb{P}(u)$.

- Note that, the above general form of control policy is both history dependent and time-variant. We say a control strategy $g = (g_0, g_1, \dots, g_t, \dots)$ is
  - **Markov** if $U_t = g_t(X_t), \forall t \geq 0$; and
  - **stationary** if $g_0 = g_1 = g_2 = \cdots$.

Example of Controlled Markov Chain

- Consider a machine whose condition at time t is described by the state X_t which can take the values 1 or 2 meaning it is in an operational or failed condition, respectively. If $X_t = 1$, the there is a probability $q > 0$ to fail in the next period. Also, a failed machine continues to remain failed. Then $\{X_t\}$ is a Markov chain whose MOTP is $\mathbb{P} = \begin{pmatrix} 1-q & q \\ 0 & 1 \end{pmatrix}$.

- We now introduce two control actions: u_k^1 is the intensity of machine use at time t taking values $0, 1$ or 2, and u_t^2 is the intensity of machine maintenance effort taking values 0 or 1. The effects of these two control actions, intensity of machine use and maintenance, can be modeled as a controlled MOTP $\mathbb{P}(u_t^1, u_t^2) = \begin{pmatrix} 1 - q_1(u_t^1) + q_2(u_t^2) & q_1(u_t^1) - q_2(u_t^2) \\ q_2(u_t^2) & 1 - q_2(u_t^2) \end{pmatrix}$.

Finite Horizon Problem

- Given a controlled Markov chain, a Markov policy g determines the state process $\{X_k\}$ and the control process $\{U_t = g_t(X_t)\}$. Clearly, different policies will lead to different processes and one is interested in finding the best or optimal policy.

- To this end, one needs to compare different policies. This is done by specifying a **cost function**, which is a sequence of real valued functions of the state and control,

  $$c_t(i, u), i \in S = \{1, \dots, I\}, u \in \mathcal{U}, t \geq 0$$

  The interpretation is that $c_t(i, u)$ is the cost to be paid if at time t, $X_t = i$ and $U_t = u$.

- The **cost incurred** by g up to the time horizon T is $\sum_{t=0}^{T} c_t(X_t, U_t)$. Note that this cost is a random variable because X_t and U_t are. Then by fixing a **Markov policy** (MP) g, this cost is just a random variable of the state process $\{X_t\}$ and the **expected total cost** of MP g is

  $$J^g = E^g\left(\sum_{t=0}^{T} c_t(X_t, U_t)\right) = E^g\left(\sum_{t=0}^{T} c_t(X_t, g_t(X_t))\right)$$

Infinite Horizon Problem

- Note that the time horizon above is finite. In some applications, one is interested in the infinite horizon when $T \to \infty$. For this case, the above expected total cost usually is meaningless because one can get $J^g = \infty$ for every g. There are two ways to treat the infinite horizon problem.

- One approach is to consider the **expected discounted cost**

  $$J^g = E^g\left(\sum_{t=0}^{\infty} \beta^t c_t(X_t, U_t)\right)$$

  where $0 < \beta < 1$ is a discount factor. Therefore, if c_t is bounded, then J^g is finite. Since the cost incurred at time t is weighted by β^t, present costs are more important than future costs. For example, in an economic context, $\beta = (1 + r)^{-1}$, where $r > 0$ is the interest rate.

- Another approach is to consider the **average cost per unit time**

  $$J^g = \lim_{T\to\infty} \frac{1}{T+1} E^g\left(\sum_{t=0}^{T} c_t(X_t, g_t(X_t))\right)$$

- For the infinite horizon case, in addition to the assumption that $\mathbb{P}(u)$ is time-invariant, hereafter, we also assume that the cost function c_t is time-invariant. Furthermore, we only consider stationary Markov policy $g = (g, g, \dots)$. Then by fixing the a stationary MP g, the controlled Markov chain becomes a standard (autonomous) Markov chain $\mathbb{P}^g$ defined by $\mathbb{P}^g_{i,j} = P_{i,j}(g(i)) = P(j \mid i, g(i))$.

Computation of Finite Horizon Cost

- **Method 1: Forward Computation**
  As we mentioned earlier, for the case of finite horizon, we assume Markov policy g. One can show that the state process $\{X_t^g\}$ is then a (non-time-homogeneous) Markov chain, where the one-step MOTP at time t is $\mathbb{P}_t^g$, where $[\mathbb{P}_t^g]_{i,j\in S} = P(j \mid i, g_t(i))$. Its m-step MOTP at time t is $\mathbb{P}_t^g \cdots \mathbb{P}_{t+m-1}^g$. Therefore, the probability distribution satisfies

$$\pi_{t+m}^g = \pi_t^g \mathbb{P}_t^g \cdots \mathbb{P}_{t+m-1}^g, \text{ where } \pi_t^g = (\pi_t^g(0), \ldots, \pi_t^g(I)) \text{ and } \pi_t^g(i) = P(X_t^g = i)$$

  Based on the above, analysis, we can easily write cost J^g in terms of the MOTP $\mathbb{P}_t^g$ as:

$$J^g = E^g\left(\sum_{t=0}^{T} c_t(X_t, g_t(X_t))\right) = \sum_{t=0}^{T}\sum_{i\in S}\pi_t^g(i)c_t(i, g_t(i)) = \sum_{t=0}^{T}\pi_0\left(\mathbb{P}_0^g \cdots \mathbb{P}_{t-1}^g\right)\underbrace{\begin{pmatrix} c_t(0, g_t(0)) \\ \vdots \\ c_t(I, g_t(I)) \end{pmatrix}}_{=:c_t^g}$$

- **Method 2: Backward Computation**
  Actually, it is more insightful to compute J^g by backward recursion. To this end, we define the expected cost incurred during $t, \ldots, T$ when $X_t = i$, i.e.,

$$V_t^g(i) = E^g\left(\sum_{l=t}^{T} c_l(X_l, g_l(X_l)) \mid X_t = i\right) \quad \Rightarrow \quad J^g = \sum_{i\in S}\pi_0(i)V_0^g(i)$$

  The functions $V_t^g(i)$ can be calculated by backward recursion as follows

$$\begin{aligned}
V_t^g(i) =& E^g\left(\sum_{l=t}^{T} c_l(X_l, g_l(X_l)) \mid X_t = i\right) \\
=& c_t(i, g_t(i)) + E^g\left(\sum_{l=t+1}^{T} c_l(X_l, g_l(X_l)) \mid X_t = i\right) \\
\overset{E(X|Y)=E(E(X|Y,Z)|Y)}{=}& c_t(i, g_t(i)) + E^g\left(E^g\left(\sum_{l=t+1}^{T} c_l(X_l, g_l(X_l)) \mid X_{t+1}, X_t = i\right) \mid X_i = i\right) \\
\overset{\text{Markov property}}{=}& c_t(i, g_t(i)) + E^g\left(E^g\left(\sum_{l=t+1}^{T} c_l(X_l, g_l(X_l)) \mid X_{t+1}\right) \mid X_i = i\right) \\
=& c_t(i, g_t(i)) + E^g\left(V_{t+1}^g(X_{t+1}) \mid X_i = i\right) \\
=& c_t(i, g_t(i)) + \sum_{j\in S} P(j \mid i, g_t(i))V_{t+1}^g(j)
\end{aligned}$$

  Note the terminal condition is $V_T^g(i) = c_T(i, g_T(i))$. Put into the vector form, we have

$$\begin{cases} V_T^g = c_T^g \\ V_t^g = c_t^g + \mathbb{P}_t^g V_{t+1}^g \\ J^g = \pi_0 V_0^g \end{cases}, \text{ where } V_t^g = \begin{pmatrix} V_t^g(0) \\ \vdots \\ V_t^g(I) \end{pmatrix} \text{ and } c_t^g = \begin{pmatrix} c_t(0, g_t(0)) \\ \vdots \\ c_t(I, g_t(I)) \end{pmatrix}$$

Computation of Infinite Horizon Expected Discount Cost

- Similar to the finite horizon case, we define the expected discount cost incurred during $t, \ldots, \infty$ when $X_t = i$, i.e.,

$$V_t^g(i) = E^g\left(\sum_{l=t}^{\infty} \beta^l c(X_l, U_l) \mid X_t = i\right)$$

Under the assumptions that $c(X, u)$ is time-invariant and g is a stationary MP, we have the followings

$$\begin{aligned}
V_t^g(i) =& E^g\left(\sum_{l=t}^{\infty} \beta^l c(X_l, g(X_l)) \mid X_t = i\right)\\
=& E^g\left(\beta^t c(X_t, g(X_l)) \mid X_t = i\right) + E^g\left(\sum_{l=t+1}^{\infty} \beta^l c(X_l, g(X_l)) \mid X_t = i\right)\\
=& \beta^t c(i, g(i)) + E^g\left(E^g\left(\sum_{l=t+1}^{\infty} \beta^l c(X_l, g(X_l)) \mid X_{t+1}, X_t = i\right) \mid X_t = i\right)\\
=& \beta^t c(i, g(i)) + E^g\left(E^g\left(\sum_{l=t+1}^{\infty} \beta^l c(X_l, g(X_l)) \mid X_{t+1}\right) \mid X_t = i\right)\\
=& \beta^t c(i, g(i)) + E^g\left(V_{t+1}^g(X_{t+1}) \mid X_t = i\right)\\
=& \beta^t c(i, g(i)) + \sum_{j \in S} P(j \mid i, g(i)) V_{t+1}^g(j) \qquad (\star)
\end{aligned}$$

- According to the definition of $V_t^g(i)$, we have that

$$V_t^g(i) = \beta^t V_0^g(i) \qquad (\star\star)$$

Therefore, combining $(\star)$ and $(\star\star)$, we have

$$\begin{aligned}
\beta^t V_0^g(i) &= \beta^t c(i, g(i)) + \sum_{j \in S} P(j \mid i, g(i)) \beta^{t+1} V_0^g(i)\\
\Rightarrow V_0^g(i) &= c(i, g(i)) + \beta \sum_{j \in S} P(j \mid i, g(i)) V_0^g(i)
\end{aligned}$$

- Put into the vector form, we need to solve equation

$$V_0^g = c^g + \beta \mathbb{P}^g V_0^g, \quad \text{where} \quad V_0^g = \begin{pmatrix} V_0^g(1) \\ \vdots \\ V_0^g(I) \end{pmatrix}, c^g = \begin{pmatrix} c(1, g(1)) \\ \vdots \\ c(I, g(I)) \end{pmatrix}$$

and the cost is

$$J^g = \sum_{i \in S} \pi_0(i) V_0^g(i), \text{ where } V_0^g = (I - \beta \mathbb{P}^g)^{-1} c^g$$

Actually, one can show that matrix $I - \beta \mathbb{P}^g$ is always invertible.

Computation of Average Cost Per Unit Time

- Note that $\mathbb{P}^g$ is actually a Markov chain because we assume that g is a stationary MP and $\mathbb{P}(u)$ is time-homogeneous. Therefore, the Cesaro limit always exists

$$\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}(\mathbb{P}^g)^t =: \Pi^g$$

  Therefore, we have

$$J^g = \lim_{T\to\infty}\frac{1}{T+1}E^g\left(\sum_{t=0}^{T}c(X_t, g_t(X_t))\right) = \lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\pi_0(\mathbb{P}^g)^t c^g = \pi_0\Pi^g c^g$$

  where we have $c^g = (c(1, g(1)), \cdots, c(I, g(I)))^{\mathrm{T}}$.

- As we have discussed previously, Π^g may be initially state dependent in the sense that $\Pi^g_{i,j} \neq \Pi^g_{k,j}$. However, under the assumption that, $\mathbb{P}^g$ is irreducible, we know that $\lim_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}(\mathbb{P}^g)^t_{i,j} \to \pi_j$, where $\pi = (\pi_0, \ldots, \pi_I)$ is the unique solution to $\pi = \pi\mathbb{P}^g$.

- Therefore, under the irreducible assumption, we know that Π^g is initial state independent and we have

$$J^g = \pi_0\Pi^g c^g = \pi c^g$$

- What happens if Π^g is not irreducible? For this case, you can compute the probability of going to each irreducible component (SCC) and applies the above procedure for each irreducible component induced sub-MC.