Markov Decision Processes **January 16, 2024**

Module 2.A: Introduction

Lecturer: Avik Kar

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In our day-to-day lives, we try to bring decision-based events in our favor by making the decisions we think to be the best. Often, decisions are to be made in sequence, and the decisions have both immediate and long-term impacts. In *sequential decision-making*, the long-term impact of a decision is chained through subsequent situations and decisions. In reality, we often anticipate these chained impacts of decisions and try to increase the chance of being in a good situation throughout. From our common wisdom, we would probably agree that it is quite a complex task. In this course, we will study a model of the *sequential decision-making* problem.

1 The Sequential Decision-Making Model

In a sequential decision-making problem, there are mainly two entities: the decision maker/agent/controller that takes the decision and the environment/system that reacts to the decision. Since the problem is sequential, each problem has a time sequence involved, called the set of decision epochs. At each decision epoch, the agent observes the system state (formal term for 'situation') and chooses action (formal term for 'decision') based on its observations. The action choice produces two results: the agent immediately receives a reward, and the system evolves to a new state which is the system state for the next decision epoch.

The key ingredients of this sequential decision-making model are the following.

1. A set of system states, namely state space.
2. A set of available actions, namely action space.
3. A set of decision epochs.
4. An initial state distribution.
5. A state and action dependent transition law.
6. A state and action dependent immediate reward map.

In this course, we will confine ourselves to finite state spaces, finite action spaces, and at most countable set of decision epochs. The last two points mean that the immediate reward and the next system state depend only on the current system state and the subsequent action. This property makes the model 'independent of past given the present' or what is popularly known as the 'Markov property.' So, we call this sequential decision-making model the 'Markov Decision Process' (MDP). We denote the random variables taking values of the system state, the action played, and the reward received at decision epoch t by X_t, U_t, and R_t, respectively, and denote their realized values by x_t, u_t and r_t respectively. Now we will introduce a general definition of MDPs.

Definition 1.1 (Markov Decision process)**.** We refer to the tuple $(X, U, \mathcal{T}, \mu_0, p, r) := M$ as a Markov decision process where

1. X is the state space of the system.
2. U is the action space. At system state $x \in X$, the agent can take action from $U_x \subseteq U$.
3. An ordered set $\mathcal{T}$ is called the set of decision epochs.
4. μ_0 is denotes the initial state distribution, i.e., $X_0 \sim \mu_0$.

5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space and $\mathcal{B}_X$ be a collection measurable sets in X. The collection of transition kernels, $\{p_t : X \times U \times \mathcal{B}_X \to [0,1] \mid t \in \mathcal{T}\}$, governs the state transitions as

$$p_t(x,u,B) := \mathbb{P}\left(X_{t+1} \in B \mid X_t = x, U_t = u\right) \ \forall (x,u,B) \in X \times U \times \mathcal{B}_X. \tag{1}$$

6. $\{\bar{r}_t : X \times U \to \mathbb{R} \mid t \in \mathcal{T}\}$ be the collection of reward maps such that at the decision epoch, $t \in \mathcal{T}$ the agent receives reward $\bar{r}_t(X_t, U_t) =: R_t$.

One may consider that the reward R_t is sampled from a reward distribution parameterized by (X_t, U_t), i.e., $R_t \sim \mathcal{D}_t(X_t, U_t)$, or even parameterized by (X_t, U_t, X_{t+1}), i.e., $R_t \sim \mathcal{D}_t(X_t, U_t, X_{t+1})$. These considerations complicate the discussion mathematically and add no insight into the topic; hence will be avoided.

Example 1.2 (Inventory control)**.** Let us consider the problem of inventory control of one particular type of item. Assume a store has the capacity to keep $C \in \mathbb{Z}_+$ unit of that item. At the beginning of t^{th} time period, the store has X_t number of items. Storekeeper orders $U_t \in \{0,1,\ldots,C - X_t\}$ number of items which arrives immediately. Based on the demand, the number of items in the $(t+1)^{th}$ time step, X_{t+1} evolves. Let the demand at period t is described by the random variable W_t. We can easily see that

$$X_{t+1} = (X_t + U_t - W_t)_+$$

and

$$p(x,u,y) = \mathbb{P}(X_{t+1} = y \mid X_t = x, U_t = u) = \begin{cases} 0 & \text{if } y > x+u \\ \mathbb{P}(W_t = x+u-y) & \text{if } 0 < y \leq x+u \\ \sum_{w=x+u}^{\infty} \mathbb{P}(W_t = w) & \text{if } y = 0. \end{cases}$$

Let the unit purchasing price and unit selling price of the item be b and s units, respectively. Let the storekeeper makes a loss of l unit per excess item and also per unfilled demand in every time step. So, the reward incurred in step t is

$$\bar{r}'(X_t, U_t, W_t) = s \min\{W_t, X_t + U_t\} - bU_t + l\,|X_t + U_t - W_t|.$$

Note that $\bar{r}'$ is not only a function of the current state and the current action but also of the demand. We can take expectations with respect to the demand W_t to fit the reward function into our formulation. Define

$$\bar{r}(x,u) \triangleq s\left(\sum_{w=0}^{x+u-1} w\mathbb{P}(\{W_t = w\}) + (x+u)\sum_{w=x+u}^{\infty} \mathbb{P}(\{W_t = w\})\right) + bu + l\sum_{w=0}^{x+u-1} |x+u-w|\,\mathbb{P}(\{W_t = w\}).$$

Notice that in this model, $p, \bar{r}$ are not time-varying. So the MDP corresponding to the problem is $(X := [C], U := \{U_x = [C-x]\}, \mathcal{T}, \mu_0, p, \bar{r})$. For the time being, we are not specifying $\mathcal{T}$ and μ_0.

1.1 Decision rules and policies

Definition 1.3 (Natural Filtration associated to an MDP)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space. For an MDP, $(X, U, \mathcal{T}, p, r)$, its natural filtration is $\mathcal{F}_\bullet := \{\mathcal{F}_t \mid t \in \mathcal{T}\}$ where $\mathcal{F}_t := \sigma(X_s, U_s : s < t, X_t)$.

Definition 1.4 (History till time t)**.** The history till time t, H_t of $(X, U, \mathcal{T}, p, r)$ is random variable defined as $H_t := (X_s, U_s : s < t, X_t)$. The realized history of the system is $h_t = (x_s, u_s : s < t, x_t) \in \mathcal{F}_t$.

A decision rule ϕ_t prescribes a procedure for action selection at a specified decision epoch, $t \in \mathcal{T}$. Based on the domain and the range, decision rules are of four types.

1. A *deterministic-Markovian* decision rule, $\phi_t : X \to U$, selects action based on only the system state of that particular decision epoch, i.e., $u_t = \phi_t(x_t)$.

2. A *deterministic-history dependent* decision rule, $\phi_t : \mathcal{F}_t \to U$, selects action based on the history of the system till the decision epoch, i.e., $u_t = \phi_t(h_t)$.

3. A *randomized-Markovian* decision rule, $\phi_t : X \to \Delta_U$, specifies a probability distribution $\phi_t(\cdot \mid x_t)$ to sample the action from, i.e., $u_t \sim \phi_t(\cdot \mid x_t)$.

4. A *randomized-history dependent* decision rule, $\phi_t : \mathcal{F}_t \to \Delta_U$, specifies a probability distribution $\phi_t(\cdot \mid h_t)$ to sample the action from, i.e., $u_t \sim \phi_t(\cdot \mid h_t)$.

It is easy to see that deterministic decision rules are a special case of randomized decision rules, and Markovian decision rules are a special case of history-dependent decision rules.

Definition 1.5 (Policy)**.** A policy, π, is a sequence of decision rules enumerated by the set of decision epochs, $\mathcal{T}$, i.e.,

$$\pi = \{\phi_t \mid t \in \mathcal{T}\}.$$

Definition 1.6 (Stationary policy)**.** A policy, $\pi = \{\phi_t \mid t \in \mathcal{T}\}$ is called a stationary policy if $\phi_t = \phi$ for all $t \in \mathcal{T}$.

Note that a stationary policy must be Markovian. Based on the type of decision rules and stationarity, there are six categories of policies as follows.

1. Deterministic, stationary policies denote the set by Π^{SD}.

2. Deterministic, Markovian policies denote the set by Π^{MD}.

3. Deterministic, history-dependent policies denote the set by Π^{HD}.

4. Randomized, stationary policies denote the set by Π^{SR}.

5. Randomized, Markovian policies denote the set by Π^{MR}.

6. Randomized, history-dependent policies denote the set by Π^{HR}.

Exercise 1.7. Show the followings:

1. $\Pi^{SD} \subset \Pi^{SR} \subset \Pi^{MR} \subset \Pi^{HR}$.

2. $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{MR} \subset \Pi^{HR}$.

3. $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{HD} \subset \Pi^{HR}$.

In the rest of the course, history-dependent policies will hardly be mentioned due to the following result.

Proposition 1.8 (Adequacy of Markov policies)**.** *Assume that the action space U is at most countable, and consider the initial state distribution has countable support. The probability distribution of each pair (x_t, u_t) and the expected cost of each stage corresponding to a randomized history-dependent policy can also be obtained with a randomized Markov policy.*

Proof. Let $\pi = \{\phi_t | t \in \mathcal{T}\}$. Without loss of generality, we consider $\mathcal{T} = [H]$ and $\mathcal{T} = \mathbb{Z}_+$ in the case of finite horizon MDP and infinite horizon MDP, respectively. Let $\xi_t(x_t)$ and $\zeta_t(x_t, u_t)$ be the corresponding distribution of x_t and (x_t, u_t), respectively. Consider a randomized Markov policy $\bar{\pi} = \{\bar{\phi}_t | t \in \mathcal{T}\}$, where $\bar{\phi}_t$ is defined for all x_t with $\xi(x_t) > 0$ by

$$\bar{\phi}_t(u_t \mid x_t) = \frac{\zeta_t(x_t, u_t)}{\xi_t(x_t)}. \tag{2}$$

Let $\bar{\xi}_t(x_t)$ and $\bar{\zeta}_t(x_t, u_t)$ be the corresponding distributions of x_t and (x_t, u_t), respectively. We will show by induction that for all t, x_t and u_t, we have

$$\xi_t(x_t) = \bar{\xi}_t(x_t), \qquad \zeta_t(x_t, u_t) = \bar{\zeta}_t(x_t, u_t). \tag{3}$$

It is sufficient to show this for all t, x_t and u_t such that $\zeta_t(x_t, u_t) > 0$.

Indeed for $t = 0$, $\xi_0(x_0)$ and $\bar{\xi}_0(x_0)$ are both equal to $\mu(x_0)$, while

$$\bar{\zeta}_0(x_0, u_0) = \bar{\xi}_0(x_0)\phi_0(u_0 \mid x_0) = \bar{\xi}_0(x_0)\frac{\zeta_0(x_0, u_0)}{\xi_0(x_0)} = \zeta_0(x_0, u_0). \tag{4}$$

Suppose that (3) holds for some $t \in \mathcal{T}$. Then, we have

$$\begin{aligned}
\bar{\xi}_{t+1}(x_{t+1}) &= \sum_{x_t,u_t:\zeta(x_k,u_k)>0} \bar{\zeta}_t(x_t,u_t)p(x_t,u_t,x_{t+1}) \\
&= \sum_{x_t,u_t:\zeta(x_k,u_k)>0} \bar{\xi}_t(x_t)\bar{\phi}_t(u_t \mid x_t)p(x_t,u_t,x_{t+1}) \\
&= \sum_{x_t,u_t:\zeta(x_k,u_k)>0} \bar{\xi}_t(x_t)\frac{\zeta_t(x_t,u_t)}{\xi_t(x_t)}p(x_t,u_t,x_{t+1}) \\
&= \sum_{x_t,u_t:\zeta(x_k,u_k)>0} \zeta_t(x_t,u_t)p(x_t,u_t,x_{t+1}) \\
&= \xi_{t+1}(x_{t+1}),
\end{aligned}$$

where the fourth equality uses the induction hypothesis. Furthermore

$$\begin{aligned}
\bar{\zeta}_{t+1}(x_{t+1},u_{t+1}) &= \bar{\xi}_{t+1}(x_{t+1})\bar{\phi}_{t+1}(u_{k+1} \mid x_{k+1}) \\
&= \bar{\xi}_{t+1}(x_{t+1})\frac{\zeta_{t+1}(x_{t+1},u_{t+1})}{\xi_{t+1}(x_{t+1})} \\
&= \zeta_{t+1}(x_{t+1},u_{t+1}),
\end{aligned}$$

thereby completing the induction. Thus π and $\bar{\pi}$ generate the same state-control pair distributions. From this, it also follows that the corresponding expected rewards of every state are equal. □

Going forward, our discussion will only involve four classes of policies: Π_{MR}, Π_{MD}, Π_{SR}, and Π_{SD}.

1.2 Types of MDPs

In this course, we will categorize MDPs based on the size of the set of decision epochs and the criterion the agent wants to optimize. A finite horizon MDP (alternatively called episodic MDP) is where $\mathcal{T}$ is a finite set and $\mathcal{T}$ is a countably infinite set in the case of an infinite horizon MDP. Without loss of generality, we consider $\mathcal{T} = [H]$ and $\mathcal{T} = \mathbb{Z}_+$ in the case of finite horizon MDP and infinite horizon MDP, respectively. We say that the horizon length of a finite horizon MDP is H if $\mathcal{T} = [H]$ and refer to ∞ as the horizon length of an infinite horizon MDP.

For either type of MDP based on horizon length, the agent tries to maximize a measurable function of the rewards collected in the stretch of the horizon in expectation over the policies. Note that the rewards are random variables, and so is the criterion. The maximization of a random variable does not make sense, so the criterion must be the expectation of the measurable function of the rewards. We will focus on three variants of MDPs as follows:

1. Finite horizon MDP with the expected total reward, $\mathbb{E}_\pi\left[\sum_{t=0}^{H}\bar{r}_t(X_t,U_t)\right]$ as the criterion.
2. Infinite horizon MDP with the expected discounted sum of rewards, $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t\bar{r}_t(X_t,U_t)\right]$ as the criterion.
3. Infinite horizon MDP with long-term expected average reward, $\liminf_{T\to\infty}\frac{1}{T}\mathbb{E}_\pi\left[\sum_{t=0}^{T-1}\bar{r}_t(X_t,U_t)\right]$ as the criterion.

$\mathbb{E}_\pi$ denotes the expectation taken with respect to the policy π.

In the rest of the course, we will spend most of the time to find the answer to the following question:

How to find an optimal policy for a given MDP?