Markov decision process

Tianxiao Zheng
SAIF

1. Introduction

Markov decision processes (MDPs) are an extension of Markov process; the difference is the addition of actions (allowing choice) and rewards (giving motivation). More broadly, a Markov decision process is a stochastic game with only one player.

MDPs provide a mathematical framework for modeling decision making in situations where outcomes are *both random and under the control of a decision maker*. MDPs are useful for studying a wide range of optimization problems solved via dynamic programming and reinforcement learning. MDPs were known at least as early as the 1950s (cf. Bellman, 1957). They are used in a wide area of disciplines, including robotics, automatic control, economics, *etc.*

Modern macroeconomic dynamics can usually be summarized as a MDP in which an agent chooses an action in a stochastic dynamical system so as to maximize some objectives.

2. Markov decision process

2.1. Elements

- Decision epochs: $\mathcal{T} = 0, 1, 2, \ldots, T$, $T \leq \infty$
- States: $s_t \in \mathbb{S}$, called state space
- Actions: $a_t \in \mathbb{A}$; actions can be chosen deterministically or randomly (*e.g.* mixed strategy in game theory). Here, only deterministic action is considered.
- Constraint on this action: $\Gamma(s_t) \subset \mathbb{A}$ is the set of feasible actions given current state s_t; $\Gamma : \mathbb{S} \to \mathbb{A}$. It is independent on histories.

After choose the action a_t at state s_t at time t, the decision maker receives

- One-period reward u_t, as a result of choosing action a_t at state s_t, the agent receives a one-period reward $u_t(s_t, a_t)$, (independent of histories of a_t or s_t)

and the system state evolves according to

- the stochastic kernel P_{t+1}. It is affected by the agent's current action and state, independent of future or past.

These elements constitute what is called Markov decision process. We have assume that $\mathbb{S}, \mathbb{A}$, and Γ are independent of time.

- If $T < \infty$, we will assume that no decision is made in the last period so that the last reward u_T is a function of states only. (bequest value)
- If $T = \infty$, we will assume that $u_t(s_t, a_t)$ and P_{t+1} are independent of time. In this case, the Markov decision process is stationary.

2.2. Optimality criterion

We assume that the agent maximizes total discounted expected rewards, where the expectation is over realizations of the states s_t. We do not discuss *non-expected utility* models (to address Allais paradox and the Ellsberg paradox) and *non-additive criterion* such as recursive utility.

2.3. Policy (Plan)

A policy is a sequence of actions $(\pi_0, \pi_1, ..., \pi_{T-1})$ such that $\pi_0 \in \mathbb{A}$ and $\mathbb{S}^t \to \mathbb{A}$ is measurable for $t = 1, 2, ..., T-1$. A policy is Markovian if π_t is independent of states from time 0 to $t-1$. It is stationary if π_t is independent of time t.

Here, the history of states are expressed by $s^t = (s_1, ..., s_t)$, and $\mathbb{S}^t$ is the associated measurable space.

3. Optimal stopping

Optimal stopping is a useful class of MDPs. Examples of optimal stopping are abundant

> An investor decides whether and when to invest in a project with exogenously given stochastic payoffs; A firm decides whether and when to enter or exit an industry; A worker decides whether and when to accept a job offer or to quit his job.

Its characteristics are

- The decision maker's choice does not affect the stochastic kernel. The only power the agent has is to choose when to stop;
- The decision is irreversible to some extent;
- There is uncertainty about future rewards or costs.

3.1. Formulation

- States: $\mathbb{S} = \mathbb{Z} \cup \{\hat{s}\}$. At time t, after observing the exogenous state z_t, the agent decides to stop or to continue. $\hat{s}$ is the additional state we introduce to represent stopping.

- Actions: $\mathbb{A} = \{0, 1\}$ has only two elements 1 (continue) and 0 (stop). Therefore, we have

$$\Gamma(s) = \begin{cases} \{0,1\}, & \text{if } s \in \mathbb{Z}, \\ \{1\}, & \text{if } s = \hat{s}. \end{cases}$$

- One period reward:

$$u_t(s,a) = \begin{cases} f_t(s), & \text{if } s \in \mathbb{Z}, a = 1 \\ g_t(s), & \text{if } s \in \mathbb{Z}, a = 0, t < T, \\ 0, & \text{if } s = \hat{s} \end{cases}$$

  That is continuation at date t generates a payoff $f_t(z_t)$, while stopping at date t yields an immediate payoff $g_t(z_t)$ and zero payoff in the future. $a = 0$ only at the date of stopping. The last period reward u_T is given by $u_T(s) = h(s), s \in \mathbb{Z},$ if $T < \infty$ and the decision maker has not chosen to stop before.
- Stochastic kernels: given exogenously and after stopping always remain in state $\hat{s}$.
- Optimality criterion: maximization of total discounted expected rewards with discount factor β

Example 3.1. *Firm exit problem*

This is a classic problem in macroeconomics and industrial organization. We may describe a stylized infinite-horizon exit model as follows.

- *states z_t is stationary and may be interpreted as a demand shock or productivity shock.*
- *one period reward: $f_t = \psi(z_t) - c_f$; $g_t = \xi$. Staying in business at date t generates profits $\psi(z_t)$ and incurs a fixed cost $c_f > 0$. The owner may decide to exit and seek outside opportunities. Let the outside opportunity value be a constant $\xi > 0$.*

Example 3.2. *Secretary problem*

This is also a well-known problem. Here, we follow Puterman, 1994.

Consider an employer's decision to hire an individual to fill a vacancy for a secretarial position. There are N candidates or applicants for this job, with different abilities. Candidates are interviewed sequentially. After an interview, the employer decides whether or not to offer the job to the current candidate. If he does not offer the job to this candidate, the candidate must seek employment elsewhere. The employer wishes to maximize the probability of giving an offer to the best candidate.

We label the N candidates from 1 to N by their abilities with 1 *being the best candidate. The difficulty of formulating this problem as an optimal stopping problem is to find the stochastic kernel associated with the evolution of the states. This requires some creativity.*

- *state space: $\mathbb{Z} = \{0, 1\}$. 1 denotes that the current one is the best seen so far, and 0 denotes that a previous one was better. The transition function is given by,*

$$Q_{t+1}(z_t; 1) = \frac{1}{t+1}; \qquad Q_{t+1}(z_t; 0) = \frac{t}{t+1}.$$

  Note that Q is independent of current state z_t, but is dependent on time.

- *action space:* $\mathbb{A} = \{0, 1\}$. *$a = 0$ means select the current object, and $a = 1$ means do not select current object and continue the search.*
- *one period rewards:*

$$u_t(z_t) = \begin{cases} f_t(z_t) = 0, \quad \text{\textit{if} } t < N \\ g_t(z_t) = \begin{cases} 0, & \text{\textit{for} } z_t = 0 \\ t/N, & \text{\textit{for} } z_t = 1 \end{cases}, \quad \text{\textit{if} } t < N \\ h(z_T) = \begin{cases} 0, & \text{\textit{for} } z_T = 0 \\ 1, & \text{\textit{for} } z_T = 1 \end{cases}, \quad \text{\textit{if} } t = N \end{cases}$$

The reward function is very easy to understand. If stopped at time t, the probability of choosing the best candidate is not 0 only when *this candidate is the best among the first t candidates. The probability of this candidate being the best of all N candidates is given by*

$$Pr(\textit{best of } N|\textit{best of } t) = \frac{Pr(\textit{best of } t|\textit{best of } N)}{Pr(\textit{best of } t)} Pr(\textit{best of } N) = \frac{1/N}{1/t} = \frac{t}{N}$$

where we have used Bayes' formula.

The optimal stopping problem is a particular example of a class of broader problems known as **discrete choice**, where decision maker's choice can in general affect the evolution of the state or its stochastic kernel. A classic example may be Rust, 1985.

4. Optimal control

Optimal control is a especially useful class of MDP in macroeconomics. Optimal control problems describe the evolution of the state system by difference equations instead of transition kernels. (more conveniently in continuous state space Markov process)

Suppose the state s of the system consists of an endogenous state $x \in \mathbb{X} \subset \mathbb{R}^{n_x}$ and an exogenous state $z \in \mathbb{Z} \subset \mathbb{R}^{n_z}$, *i.e.*, $s = (x, z)$. Time is denoted by $t = 0, 1, 2, ..., T \leq \infty$. The initial state (x_0, z_0) is given. The exogenous state evolves according to a time-homogeneous Markov process with the stationary transition function Q. The endogenous state evolves according to the following difference equation:

$$x_{t+1} = \phi_t(x_t, a_t, z_t, z_{t+1}), \quad t = 0, 1, ..., T-1, (x_0, z_0) \text{ given},$$

where $\phi_t : \mathbb{X} \times \mathbb{A} \times \mathbb{Z} \times \mathbb{Z} \to \mathbb{X}$ is a measurable function. Note that the decision maker's choice $a_t \in \mathbb{A} \subset \mathbb{R}^{n_a}$ may affect the above state transition equation. The action a_t is vector of control variables.

4.1. Infinite horizon ($T = \infty$)

We typically consider the time-homogeneous case in which $u_t = u$ and $\phi_t = \phi$. The decision maker's objective is to choose a feasible policy from (x_0, z_0) to maximize the total discounted expected reward:

$$\max_{\{a_t\}_{t=0}^{\infty}} E\left[\sum_{t=0}^{\infty} \beta^t u(x_t, z_t, a_t)\right]$$

subject to

$$x_{t+1} = \phi(x_t, a_t, z_t, z_{t+1})$$

In practice, there may also include intra-temporal constraints on the states and actions each period.

4.2. Finite horizon ($T < \infty$)

The decision maker's objective is to choose a feasible policy $(a_t)_{t=0}^{T-1}$ from (x_0, z_0) to maximize the total discounted expected reward:

$$\max_{\{a_t\}_{t=0}^{T-1}} E\left[\sum_{t=0}^{T-1} \beta^t u_t(x_t, z_t, a_t) + \beta^T u_T(x_T, z_T)\right]$$

subject to $x_{t+1} = \phi_t(x_t, a_t, z_t, z_{t+1})$.

When the action set $\mathbb{A}$ is a finite set, the control problem is a discrete choice problem. When actions involve both discrete and continuous choices, the problem is often called a mixed stopping and control problem. Examples include inventory management and optimal investment with fixed costs. With fixed costs, the decision maker has to decide when to make adjustments as well as what size of adjustments to make.

5. Multi-armed Bandit

A bandit process is a special type of MDP in which there are just two possible actions:

- $a = 1$ (continue) produces reward $u(x_t)$ and the state changes to x_{t+1}, according to Markov dynamics $Q(x_t, x_{t+1})$.
- $a = 0$ (freeze) produces no reward and the state does not change (hence the term 'freeze').

A multi-armed Bandit is a collection of Bandit processes. At each time, $t \in \{0, 1, 2, ...T\}$,

- One bandit process is to be activated (pulled/continued). If arm i activated then it changes state: $x_i(t)x_i(t+1)$ with probability $Q_i(x, y)$ and produces reward $u_i(x_i(t))$.
- All other bandit processes remain passive (not pulled/frozen).
- Objective: maximize the expected total β-discounted reward

$$\max_{\{a_t\}} \mathbb{E}\left[\sum_{t=0}^{T} \beta^t u_{i_t}(x_{i_t}(t))\right]$$

where i_t is the arm pulled at t.

There are many applications of the bandit model in economics and operations research, including task selection, search, resource allocation, and choice of R&D processes, etc.

Example 5.1. *Task selection*

$s_t^i \in [0,1]$ *represents the degree of completion of task* i*,* $i = 1, 2, ..., k$*. If the decision maker takes on task* i*, he receives an expected reward* $u_t^i(s_t^i)$*.* s_t^i *is modeled as a Markov process.* $p^i(C|s_t^i)$ *is the conditional probability of completing task* i *given the state* s_t^i*. Therefore,* $u_t^i = R^i p^i(C|s_t^i)$*, where* R^i *represents the reward if task* i *is completed. The decision maker can work on only one task at any time.*

Appendix A. Continuous-time Markov decision process

Continuous-time MDPs have applications in queueing systems, epidemic processes, and population processes. Recently, continuous time modeling is becoming more popular in economics. Here we introduce very briefly the continuous-time MDPs.

For continuous-time MDPs, decisions can be made at any time the decision maker chooses. In comparison to discrete-time Markov decision processes, continuous-time Markov decision processes can better model the decision making process for a system that has continuous dynamics, i.e., the system dynamics is defined by partial differential equations (PDEs).

If the state space and action space are finite, we could use linear programming to find the optimal policy, which was one of the earliest approaches applied. If the state space and action space are continuous, the optimal criterion could be found by solving Hamilton-Jacobi-Bellman (HJB) partial differential equation.

References

Bellman, R., 1957. Dynamic Programming. Princeton University Press, Princeton, NJ, USA, first ed.

Puterman, M. L., 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley.

Rust, J., 1985. Stationary equilibrium in a market for durable assets. Econometrica 53, 783–805.