An introduction to Markov Decision Process

Alireza Kavoosi
School of Industrial Engineering, University of Tehran

February 23, 2025

Overview

1. Introduction
2. Optimality equation
3. Some examples

What is MDP?

- Markov decision processes (MDPs), often referred to as stochastic dynamic programming, have been the focus of extensive research since their introduction in [Bellman, 1957].
- Dynamic programming has seen significant advancement since the late 1950s, thanks to numerous outstanding contributions from researchers.
- MDPs are primarily utilized to model and address dynamic decision-making challenges over multi-periods in stochastic environments.
- In [Watkins, 1989], the full integration of dynamic programming strategies was demonstrated, and its approach to reinforcement learning using the MDP formulation has been broadly accepted in the field.
- The simplest form of MDPs is the discrete-time Markov decision process, which can be represented as follows:

$$\{S, A(i), p_{ij}(a), r(i, a), V\}$$

System Overview

$\{S, A(i), p_{ij}(a), r(i,a), V\}$

- The system has a state space S observed at discrete time periods $n = 0, 1, \ldots$.
- When in state $i \in S$:
  - Choose an action a from the action set $A(i)$.
  - Outcomes:
    - Receive a reward $r(i,a)$.
    - Transition to state j with probability $p_{ij}(a)$.
- The objective V is defined later.
- Assume S and all $A(i)$ are countable.
- Define $\Gamma = \{(i,a) | i \in S, a \in A(i)\}$ as the set of state-action pairs.

Decision Functions and Policies

- Define $A := \bigcup_{i \in S} A(i)$ as the union of action sets.
- A decision function $f : S \to A$:
  - $f(i) \in A(i)$ for $i \in S$.
  - Action $f(i)$ is chosen when state i is observed.
- Let F be the set of all decision functions, $F = \times_{i \in S} A(i)$.

Policies

- A policy determines actions based on history and observation period.
- Define history sets:
  - $H_n = \Gamma^{n-1} \times S$ for $n > 0$.
  - $H_0 = S$.
- A policy $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$:
  - For any $n \geq 0$ and history $h_n = (i_0, a_0, \ldots, i_n) \in H_n$:
  - $\pi_n(h_n)$ is a probability distribution on $A(i_n)$.

Question 1 What is a deterministic policy?

Question 2 What is a Markov policy?

Stochastic Process and Decision Criteria

- For $n \geq 0$:
  - X_n: State at period n.
  - Δ_n: Action chosen at period n.
- The process $\{X_n, \Delta_n, n \geq 0\}$ is well-defined under any policy $\pi \in \Pi$.
- Under a Markov policy $\pi \in \Pi_M$ Forms a discrete-time Markov chain.
- For each $\pi \in \Pi$ and $i \in S$:
  - $P_{\pi,i}$: Probability under policy π with initial state i.
  - $E_{\pi,i}$: Expectation under policy π with initial state i.
- Reward structure:
  - Reward $r(X_n, \Delta_n)$ at period n is random.

Question How to compare different policies?

Markovian decision process

Let X_n, Δ_n denote the state and the action taken (by the system) at period n. The total reward is:

Discounted criterion/total reward criterion

$V_\beta(\pi, i) = \sum_{n=0}^{\infty} \beta^n \mathbf{E}_{\pi,i}(r(X_n, \Delta_n)), \qquad i \in S, \pi \in \Pi$

In the literature, the discount rate $\beta \in [0, 1]$ is often assumed. Why?
The optimal value function for this criterion is defined by:

$$V_{\beta,N}(i) = \sup_{\pi \in \Pi} V_{\beta,N}(\pi, i), \qquad i \in S$$

Condition 1

$V_\beta(\pi, i)$ is well-defined for all $\pi \in \Pi$ and $i \in S$.

Definitions

We define the optimal value function as

$$V_\beta(i) = \sup\{V_\beta(\pi, i) \mid \pi \in \Pi\} \quad \text{for } i \in S.$$

For a given $\varepsilon > 0$, a policy $\pi^* \in \Pi$, and a state $i \in S$, we say that π^* is ε-optimal at state i if:

- $V_\beta(\pi^*, i) \geq V_\beta(i) - \varepsilon$ when $V_\beta(i) < +\infty$, or
- $V_\beta(\pi^*, i) \geq \frac{1}{\varepsilon}$ when $V_\beta(i) = +\infty$.

Here, we adopt the convention that $\frac{1}{0} = +\infty$.
If π^* is ε-optimal for all states $i \in S$, we refer to π^* as an ε-optimal policy.
A policy that is 0-optimal is termed an optimal policy.

Validity of optimality equation

Condition 2

For any policy $\pi = (\pi_0, \pi_1, \dots) \in \Pi$ *and state* $i \in S$,

$$V_\beta(\pi, i) = \int_{A(i)} \pi_0(da|i)\{r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(\pi^{i,a}, j)\},$$

where $\pi^{i,a} = (\sigma_0, \sigma_1, \dots) \in \Pi$ *with* $\sigma_n(\cdot|h_n) = \pi_{n+1}(\cdot|i, a, h_n)$ *for* $n \geq 0$.

This condition shows that any process under a policy π splits naturally into the first period and the following periods—a key idea behind the optimality equation. It also guarantees that the summation $\sum_j$ and the integration $\int_{A(i)}$ are well defined.

Many works verify Condition 2 under assumptions that $r(i, a)$ is nonnegative, nonpositive, or bounded.

State Subsets

We now partition the state space into three parts. Define

$$S_\infty := \{i \mid V_\beta(i) = +\infty\}, \quad S_{-\infty} := \{i \mid V_\beta(i) = -\infty\},$$

and

$$S_0 := S \setminus (S_\infty \cup S_{-\infty}) .$$

These represent, respectively, states with positive infinite, negative infinite, and finite optimal values. Moreover, let

$$S_{=\infty} := \{i \mid \text{there is } \pi \in \Pi \text{ such that } V_\beta(\pi, i) = +\infty\}.$$

Clearly, $S_{=\infty} \subset S_\infty$.

Lemma 1

Under Conditions 1 and 2, $\sum_{j \in S_0} p_{ij}(a) V_\beta(j)$ is well defined for any $(i, a) \in \Gamma$.

Validity of optimality equation

Proof of lemma 1

By Condition 2, for any $(i, a) \in \Gamma$ and policy $\pi \in \Pi$, the series $\sum_j p_{ij}(a)V_\beta(\pi, j)$ is well-defined. In particular, consider the policy (f, π) with $f(i) = a$, which uses f in the first period and π thereafter.

Now, for any $\varepsilon > 0$ and $j \in S_0$, choose a policy $\pi(\varepsilon, j)$ such that

$$V_\beta(\pi(\varepsilon, j), j) \geq V_\beta(j) - \varepsilon,$$

and let $\pi(\varepsilon)$ be the policy that selects $\pi(\varepsilon, j)$ when the initial state is j. Then, $\sum_{j\in S_0} p_{ij}(a)V_\beta(\pi(\varepsilon, j), j) = \sum_{j\in S_0} p_{ij}(a)V_\beta(\pi(\varepsilon), j)$ is well-defined. Consequently, for any subset $S'' \subset S_0$, we have

$$\sum_{j\in S''} p_{ij}(a)V_\beta(j) \leq \sum_{j\in S''} p_{ij}(a)[V_\beta(\pi(\varepsilon), j) + \varepsilon].$$

This confirms that the series $\sum_{j\in S_0} p_{ij}(a)V_\beta(j)$ is well defined. $\square$

Optimality equation

Theorem 1

Provided that Condition 1 and Condition 2 are true and that $\sum_j p_{ij}(a)V_\beta(j)$ is well defined for any $(i,a) \in \Gamma$, then V_β satisfies the following optimality equation:

$$V_\beta(i) = \sup_{a \in A(i)} \left\{ r(i,a) + \beta \sum_j p_{ij}(a) V_\beta(j) \right\}, \quad i \in S.$$

Setup and Statement of the Proof

Proof

Given By Condition 2, for all $i \in S$, we have

$$V_\beta(i) \leq \sup_{a \in A(i)} \Big\{ r(i,a) \, + \, \beta \sum_j p_{ij}(a)\, V_\beta(j) \Big\}. \tag{1}$$

Policy Construction: For any $\varepsilon > 0$, define a policy $\pi(\varepsilon, i)$ such that:

1. If $i \in S_0$, then $V_\beta(\pi(\varepsilon, i), i) \geq V_\beta(i) - \varepsilon$.
2. If $i \in S_\infty$, then $V_\beta(\pi(\varepsilon, i), i) \geq 1/\varepsilon$.
3. If $i \in S_{-\infty}$, then $V_\beta(\pi, i) = -\infty$ for *any* policy π.

Let $\pi(\varepsilon)$ be a policy that chooses $\pi(\varepsilon, j)$ whenever the initial state is $j \in S_0 \cup S_\infty$.
Goal: We want to show that for any $(i, a) \in \Gamma$,

$$V_\beta(i) \; \geq \; r(i,a) \, + \, \beta \sum_j p_{ij}(a)\, V_\beta(j). \tag{2}$$

Proof (Part 2)

Case 0: If $i \in S_\infty$, then (2) is *trivial*.

Case 1: Suppose $i \in S - S_\infty$. Pick any f with $f(i) = a$. By Condition 2.2, we have

$$V_\beta(i) \;\geq\; V_\beta\big((f, \pi(\varepsilon)), i\big) \;=\; r(i, a) \;+\; \beta \sum_j p_{ij}(a)\, V_\beta\big(\pi(\varepsilon), j\big).$$

Thus, it follows that

$$V_\beta(i) \;\geq\; r(i, a) \;+\; \beta \sum_j p_{ij}(a)\, V_\beta(j). \tag{3}$$

We now verify that this implies (2) by considering three subcases:

Proof (Part 3)

Subcase 1A:

$$p_{iS_\infty}(a) \; := \; \sum_{j \in S_\infty} p_{ij}(a) \; > \; 0 \quad \text{or} \quad \sum_{j \in S_0} p_{ij}(a)\, V_\beta(j) \; = \; -\infty.$$

- If $\sum_{j \in S_0} p_{ij}(a)\, V_\beta(j)$ is well-defined and equals $-\infty$, then directly (2) holds.
- If $p_{iS_\infty}(a) > 0$ and $\sum_{j \in S_0} p_{ij}(a)\, V_\beta(j)$ is finite, we will see it forces an infinite value.

Proof (Part 4)

Subcase 1B:

$$p_{iS_\infty}(a) \;=\; 0 \quad \text{and} \quad \sum_{j\in S_0} p_{ij}(a)\, V_\beta(j) \;>\; -\infty \quad \text{but} \quad p_{iS_\infty}(a) \;>\; 0.$$

Here,

$$\sum_j p_{ij}(a)\, V_\beta(j) \;=\; \sum_{j\in S_0} p_{ij}(a)\, V_\beta(j) \;+\; \sum_{j\in S_\infty} p_{ij}(a)\, V_\beta(j) \;=\; +\infty.$$

From (3),

$$V_\beta(i) \;\geq\; r(i,a) \;+\; \beta \sum_{j\in S_0} p_{ij}(a) \left[V_\beta(j) - \varepsilon\right] \;+\; \beta\, p_{iS_\infty}(a) \left(1/\varepsilon\right).$$

Letting $\varepsilon \to 0^+$ forces $V_\beta(i) = +\infty$, so (2) holds.

Proof (Part 5)

Subcase 1C:

Neither of the above conditions (1A or 1B) hold.

Then from (3), we have

$$V_\beta(i) \;\geq\; r(i,a) \;+\; \beta \sum_{j \in S_0} p_{ij}(a)\, V_\beta(j) \;-\; \varepsilon.$$

Since ε is arbitrary, this again implies (2).

Conclusion of the Cases:

- In all scenarios, (2) holds.
- Hence, (2) implies The theorem from (1), using the arbitrariness of i and a.

Therefore, the proof is complete. $\square$

Continuous Problems with Exact Solutions [Powell, 2022]

A Gambling Problem [Ross, 2014]

At each play of the game, a gambler can bet any nonnegative amount up to his present fortune and will either win or lose that amount with probabilities p and $q = 1 - p$, respectively. The gambler is allowed to make n bets and his objective is to maximize the expectations of the logarithm of his final fortune. What strategy achieves this end?

Continuous Problems with Exact Solutions

Sequential Investment Problem

Suppose one has an amount M of money and considers investing this money over N future periods. However, the opportunity for investment is not deterministic. At each period, an investment opportunity occurs with probability p, which is independent of the past and the amount of remaining money. When an investment opportunity occurs, if he invests x, he will earn a revenue $r(x)$, including his investment. Assume that both his investment and his return at any period cannot be reinvested in the future. What is the optimal strategy for this problem?
Let $V_n(X)$ be the maximal expected profit when there are n periods remaining, X money available for future investment, and an investment opportunity occurs.

1. Write the optimality equation.
2. Assume that $r(x)$ is nondecreasing, concave, and satisfies $r(0) = 0$. Show that $V_n(X)$ is also concave in X.

A Stock Option Model

Consider the problem of buying an option for a given stock. Let P_n be the price of the stock on the nth day for $n = 0, 1, 2, \ldots$. Suppose that $\{P_n\}$ satisfies the random-walk model, meaning there are independent random variables $\xi_1, \xi_2, \ldots$ with an identical distribution function F such that

$$P_{n+1} = P_n + \xi_n, \quad n \geq 0.$$

Here, P_0 is the initial price and is independent of $\{\xi_n, n \geq 0\}$. Moreover, we assume that one has the option to buy one share of the stock at a fixed price r on the initial day and then exercise the option on some day in the future. A strategy is to determine when to exercise the option, which is obviously based on the price of the stock, given r.

Expected Revenue and Strategy

A Stock Option Model- Continue

Let $V(p)$ be the maximal expected revenue when the current price of the stock is p. The problem is to find a strategy that maximizes the expected profit from exercising the option.

1. Write the optimality equation.
2. Show that $V(p) - p$ is decreasing in p under the condition that the mean of ξ_n is finite.
3. Show that under the condition that the mean of ξ_n is finite, the optimal strategy is as follows: if the current price is p, then exercise the option if and only if $p \geq p^*$ for some number p^*.
4. Discuss whether the results above still hold when the mean of ξ_n is infinite.

References

Bellman, R. (1957).
A markovian decision process.
Journal of mathematics and mechanics, pages 679–684.

Powell, W. B. (2022).
Exact Dynamic Programming, chapter 14, pages 731–794.
John Wiley & Sons, Ltd.

Ross, S. M. (2014).
Introduction to stochastic dynamic programming.
Academic press.

Watkins, C. J. C. H. (1989).
Learning from delayed rewards.

Thank you for your attention