Markov Decision Processes: Model Formulation and Policies

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Markov Decision Processes

1.1 Introduction

A Markov Decision Process (MDP) is a mathematical model for sequential decision making under uncertainty. It describes a system that evolves over time, where the evolution depends not only on random effects but also on deliberate actions chosen by a decision maker.

An MDP generalizes a Markov chain by allowing the transition probabilities between states to depend on the action selected at each decision epoch.

At each decision epoch:

1. the system occupies a state,
2. an action is selected,
3. an immediate reward (or cost) is incurred,
4. the system transitions to a new state according to a controlled probability law.

As the process unfolds, the decision maker observes a sequence of rewards. The objective of a sequential decision problem is to choose actions over time so as to optimize a prescribed function of this reward sequence.

1.2 Decision Epochs

Decisions are made at discrete points in time called *decision epochs*. The time index is denoted by

$$t = 1, 2, \ldots, T.$$

- If the time between decision epochs is constant, the model is called a *Markov Decision Process*.
- If the time between decision epochs is random, the model is called a *Semi-Markov Decision Process*.

In these notes, we restrict attention to discrete decision epochs. Continuous-time decision problems are typically treated within control theory.

1.3 State Space

At each decision epoch, the system occupies a state that summarizes all relevant information needed for future decision making.

We denote the state space by

$$S,$$

and assume it is finite, with cardinality

$$|S| = N.$$

The defining feature of the state is that, given the current state and the action chosen, the future evolution of the system is conditionally independent of the past.

1.4 Action Sets

When the system is in state $s \in S$, the decision maker chooses an action from a set of feasible actions.

We denote the set of available actions in state s by

$$A_s.$$

The set of all actions in the model is

$$A = \bigcup_{s \in S} A_s.$$

The dependence of the action set on the current state allows the model to capture physical, logical, or operational constraints.

1.5 Immediate Rewards

If action $a \in A_s$ is chosen when the system is in state s at decision epoch t, an immediate reward is received.

The reward is denoted by

$$r_t(s, a).$$

The reward depends only on:

- the decision epoch t,
- the current state s,
- the chosen action a,

and does not depend on the past history of the process.

Depending on the application, rewards may represent profits, utilities, or negative costs.

1.6 Transition Probabilities and the Markov Property

After choosing action $a \in A_s$ in state s at time t, the system transitions to a new state at the next decision epoch.

The transition probability is denoted by

$$P_t(s' \mid s, a), \qquad s' \in S.$$

This probability specifies the conditional distribution of the next state given the current state and action.

The defining *Markov property* of an MDP is that

$$\mathbb{P}(X_{t+1} = s' \mid X_1, \ldots, X_t, Y_1, \ldots, Y_t) = P_t(s' \mid X_t, Y_t).$$

Thus, the future evolution of the system depends on the past only through the current state and the chosen action.

1.7 Planning Horizon

The number of decision epochs is called the planning horizon.

- If the horizon $T < \infty$, the model is a *finite-horizon MDP*.
- If $T = \infty$, the model is an *infinite-horizon MDP*.

In a finite-horizon MDP, no action is chosen at the terminal epoch T. Instead, a terminal reward depending only on the final state is received.

This terminal reward is denoted by

$$r_N(s_N).$$

1.8 MDP Specification

A finite-horizon Markov Decision Process is completely specified by the collection

$$(S, \{A_s\}_{s \in S}, \{P_t(\cdot \mid s, a)\}, \{r_t(s, a)\}, r_N)\,.$$

These objects define the dynamics, rewards, and admissible decisions of the system.

1.9 Decision Rules and Policies

At each decision epoch t, the system occupies a state $s_t \in S$, and an action must be selected from the feasible action set A_{s_t}.

A *decision rule* specifies how this action is determined at decision epoch t. It is designed by the decision maker and governs action selection at that single decision epoch.

A *policy* is a sequence of decision rules,

$$\pi = (d_1, d_2, \ldots, d_T),$$

where d_t is the decision rule used at decision epoch t. Thus, a policy specifies how actions are determined at all decision epochs over the planning horizon.

1.10 Information and Action Specification

There are two independent aspects that characterize how a decision rule specifies actions:

(1) **Information:** What information is used as input to the decision rule when specifying an action?

(2) **Action specification:** Does the decision rule specify a single action, or does it specify a probability distribution over the feasible action set?

The first aspect distinguishes *Markovian* and *history-dependent* decision rules. The second distinguishes *deterministic* and *randomized* decision rules.

1.11 Histories and Available Information

A decision rule may depend on more than just the current state. To formalize the notion of past information, we introduce the concept of history.

The *history* of the system up to time t is defined as

$$h_t = (s_1, a_1, s_2, a_2, \ldots, s_{t-1}, a_{t-1}, s_t),$$

where s_u and a_u denote the state and action at decision epoch u.

Equivalently,

$$h_t = (h_{t-1}, a_{t-1}, s_t).$$

Let H_t denote the set of all possible histories up to time t. Then

$$H_1 = S, \qquad H_t = H_{t-1} \times A \times S \quad (t \geq 2).$$

More explicitly,

$$H_t = (S \times A)^{t-1} \times S.$$

A decision rule is called *Markovian* if it depends only on the current state s_t. It is called *history-dependent* if it depends on the entire history h_t.

1.12 The Four Classes of Decision Rules

Decision rules are classified according to the information they use and the manner in which actions are specified.

1.12.1 Deterministic Markovian Decision Rules (MD)

A deterministic Markovian decision rule depends only on the current state and specifies a single action with certainty:

$$d_t : S \to A, \qquad d_t(s) \in A_s \quad \text{for all } s \in S.$$

Such a rule is *Markovian* because it uses only the current state, and *deterministic* because the action is uniquely determined once the state is known.

The rule may depend on time. Consequently, the same state may lead to different actions at different decision epochs.

1.12.2 Deterministic History-Dependent Decision Rules (HD)

A deterministic history-dependent decision rule uses the entire history and specifies a single action:

$$d_t : H_t \to A, \qquad d_t(h_t) \in A_{s_t},$$

where s_t is the current state contained in the history h_t.

The decision rule may use any aspect of the past evolution of the system, but the selected action must be feasible in the current state.

1.12.3 Randomized Markovian Decision Rules (MR)

A randomized Markovian decision rule depends only on the current state and specifies a probability distribution over feasible actions:

$$d_t : S \to \mathcal{P}(A), \qquad d_t(s) \in \mathcal{P}(A_s).$$

At state s_t, the rule specifies a probability distribution over the set A_{s_t}. The action realized at the decision epoch is governed by this distribution.

The set $\mathcal{P}(A_s)$ denotes the collection of all probability distributions on the finite set A_s:

$$\mathcal{P}(A_s) = \Big\{ q : A_s \to [0,1] \;\Big|\; \sum_{a \in A_s} q(a) = 1 \Big\}.$$

Randomization arises from the decision rule itself and is distinct from the randomness present in the system dynamics.

1.12.4 Randomized History-Dependent Decision Rules (HR)

A randomized history-dependent decision rule uses the full history and specifies a probability distribution over feasible actions:

$$d_t : H_t \to \mathcal{P}(A), \qquad d_t(h_t) \in \mathcal{P}(A_{s_t}).$$

This is the most general class of decision rules considered in finite-state, discrete-time Markov decision process theory.

1.13 Deterministic Rules as Special Cases

A deterministic decision rule can be viewed as a special case of a randomized decision rule in which the probability distribution is degenerate.

For a deterministic Markovian decision rule d_t, the associated distribution is

$$q_t(a \mid s) = \begin{cases} 1, & a = d_t(s), \\ 0, & \text{otherwise.} \end{cases}$$

An analogous representation holds for deterministic history-dependent decision rules.

1.14 Decision Rule Sets and Classification Summary

Decision rules are classified as:

- MD: Markovian and deterministic,
- MR: Markovian and randomized,
- HD: history-dependent and deterministic,
- HR: history-dependent and randomized.

Let $D_t^{(K)}$ denote the set of all decision rules at time t belonging to class $K \in \{\mathrm{MD}, \mathrm{MR}, \mathrm{HD}, \mathrm{HR}\}$. The set $D_t^{(K)}$ is called a *decision rule set*.

1.15 Fundamental Question of MDP Theory

The central question in Markov decision process theory is:

> For a given optimality criterion, under what conditions is it sufficient to restrict attention to deterministic Markovian decision rules?

1.16 Policies

A *policy* (also called a contingency plan, plan, or strategy) specifies which decision rule is to be used at *every* decision epoch.

Intuitively, a policy provides the decision maker with a complete prescription for action selection under all possible future evolutions of the system, whether those evolutions are described by states alone or by full histories.

Formally, a policy π is a sequence of decision rules,

$$\pi = (d_1, d_2, \dots, d_{T-1}),$$

where $d_t \in D_t^{(K)}$ for each decision epoch t, and where $K \in \{\mathrm{MD}, \mathrm{MR}, \mathrm{HD}, \mathrm{HR}\}$ denotes the class of decision rules being used.

Let $\Pi^{(K)}$ denote the set of all policies of class K. Thus,

$$\Pi^{(K)} = D_1^{(K)} \times D_2^{(K)} \times \cdots \times D_{T-1}^{(K)}.$$

The distinction between different classes of policies is entirely inherited from the class of decision rules they contain.

1.17 Stationary Policies

A policy is called *stationary* if the same decision rule is used at every decision epoch.

That is, a policy π is stationary if

$$d_t = d \quad \text{for all } t.$$

In this case, the policy has the form

$$\pi = (d, d, d, \dots),$$

and is often denoted simply by d^∞ or, when no confusion arises, by d.

Stationarity means that the action-selection mechanism does not explicitly depend on time. The same rule is applied whenever the system visits a given state, regardless of when that visit occurs.

Stationary policies play a fundamental role in the theory of infinite-horizon Markov decision processes.

1.18 Stationary Markovian Randomized Policies

Consider a stationary Markovian randomized decision rule

$$d : S \to \mathcal{P}(A), \qquad d(s) \in \mathcal{P}(A_s).$$

Such a rule assigns to each state s a probability distribution over the feasible action set A_s.

Since both the state space and action space are finite, a stationary Markovian randomized policy can be represented as a matrix

$$D = \big[d(s,a)\big]_{s\in S,\, a\in A},$$

where

$$d(s,a) = \mathbb{P}(a \mid s).$$

This matrix has the following properties:

- $d(s,a) \geq 0$ for all $s \in S$ and $a \in A$,
- $d(s,a) = 0$ for all $a \notin A_s$,
- for each $s \in S$,

$$\sum_{a\in A_s} d(s,a) = 1.$$

Thus, each row of D corresponds to a state and is a probability distribution over the actions feasible in that state. The matrix D is therefore *row-stochastic*.

We sometimes refer to D as the *decision matrix* associated with the policy.

1.19 Stationary Markovian Deterministic Policies

A stationary Markovian deterministic policy is a special case of a stationary randomized policy.

In this case, for each state $s \in S$, the policy selects a single action $a = d(s) \in A_s$ with certainty.

The associated decision matrix D satisfies

$$d(s,a) = \begin{cases} 1, & a = d(s), \\ 0, & \text{otherwise.} \end{cases}$$

Thus, each row of D contains exactly one entry equal to 1, with all other entries equal to 0.

Stationary deterministic policies are sometimes called *pure policies*. We denote the set of stationary deterministic policies by Π^{SD} and the set of stationary randomized policies by Π^{SR}.

1.20 Relationship Between Policy Classes

The various classes of policies are related by inclusion.

In particular,

$$\Pi^{\mathrm{SD}} \subset \Pi^{\mathrm{SR}} \subset \Pi^{\mathrm{MR}} \subset \Pi^{\mathrm{HR}},$$

and

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

Thus, randomized history-dependent policies are the most general, while stationary deterministic policies are the most restrictive.

A central theme of Markov decision process theory is to identify conditions under which optimal policies exist within these smaller, more structured classes.