6.4110
Spring 2026 Recitation: MDPs

Author: Sunshine Jiang

Markov Decision Processes (MDPs)

An MDP is defined by:

- States: $\mathcal{S}$
- Actions: $\mathcal{A}$
- Transition model: $P(s' \mid s, a)$
- Reward: $R(s, a, s')$
- Discount factor: γ

Objective:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Key idea: A policy $\pi(s)$ specifies which action to take in each state.

Value Function and Q-Function

Value function:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, \pi\right]$$

Q-function:

$$Q^{\pi}(s, a) = \sum_{s'} P(s'|s, a)\left[R(s, a, s') + \gamma V^{\pi}(s')\right]$$

Relationship:

$$V^{\pi}(s) = Q^{\pi}(s, \pi(s))$$

Key idea:

- $V(s)$: how good a state is
- $Q(s, a)$: how good an action is at this state

Policy Evaluation (Bellman Expectation)

Given a fixed policy π:

$$V^{\pi}(s) = \sum_{s'} P(s'|s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V^{\pi}(s') \right]$$

Interpretation:

- Follow the policy $\pi(s)$
- Enumerate possible next states
- Compute reward + discounted future value
- Take expectation

Important: No maximization — the policy is fixed.

Bellman Optimality Equation

$$V^{*}(s) = \max_{a} \sum_{s'} P(s'|s, a) \left[R(s, a, s') + \gamma V^{*}(s') \right]$$

Structure:

$$\text{Expectation over } s' \quad \rightarrow \quad \max \text{ over } a$$

Interpretation:

- Try each action a
- Compute expected outcome
- Choose the best action

Bellman Update

$$V(s) \leftarrow \max_{a} \sum_{s'} P(s'|s, a) \left[R(s, a, s') + \gamma V(s') \right]$$

Meaning:

- Perform one-step lookahead
- Update $V(s)$ using current estimates of $V(s')$

Value Iteration

Update rule:

$$V_{k+1}(s) = \max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_k(s')\right]$$

Algorithm:

- Initialize $V_0(s) = 0$
- Repeatedly apply Bellman updates

Policy extraction:

$$\pi^*(s) = \operatorname*{argmax}_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V^*(s')\right]$$

1 Practice Problem: TV MDP

After your finals, you are relaxing on the couch watching TV. There is a remote control that you can reach while sitting on the couch to turn on the TV. Unfortunately, both the TV and the remote are buggy: the TV may turn off randomly, and clicking the remote may fail. You can also get up and turn on the TV manually.

We model this as an infinite-horizon MDP:

- States: $\mathcal{S} = \{\texttt{tvOn}, \texttt{tvOff}\}$
- Actions: $\mathcal{A} = \{\texttt{clickRemote}, \texttt{pressTV}, \texttt{doNothing}\}$
- Discount factor: $\gamma = 0.9$

Transition model $P(s'|s,a)$:

- If $a = \texttt{clickRemote}$:
  - Switch state with probability 0.4
  - Stay in same state with probability 0.6
- If $a = \texttt{pressTV}$:
  - Always switch state
- If $a = \texttt{doNothing}$:
  - If TV is on: turns off with probability 0.05
  - If TV is off: stays off with probability 1.0

Reward function $R(s,a,s')$:

- If $s' = \texttt{tvOn}$: reward $+3$
- If $a = \texttt{clickRemote}$: reward -1
- If $a = \texttt{pressTV}$: reward -2

You are given the policy:

$$\pi(\texttt{tvOn}) = \texttt{doNothing}, \quad \pi(\texttt{tvOff}) = \texttt{clickRemote}$$

(a) Policy Evaluation

Write the system of equations for $V^{\pi}(s)$.

(b) Value Iteration

Suppose:

$$V(\texttt{tvOn}) = 2, \quad V(\texttt{tvOff}) = 1$$

Compute one Bellman update for $V(\texttt{tvOn})$.