Lecture 18: Markov Decision Processes

Yudong Chen
UW-Madison CS
April 6, 2022
Section 1.1–1.3.3 in Reinforcement Learning: Theory and Algorithms, by Alekh Agarwal, Nan Jiang, Sham M. Kakade, Wen Sun, 2021. ("AJKS" book)
https://rltheorybook.github.io
Discounted Markov Decision Process (MDP)

- Interaction between agent and environment is described by an MDP $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$.

- $\mathcal{S}$: *state space*; finite

- $\mathcal{A}$: *action space*; finite

- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ is the *transition kernel*, where $\Delta(\mathcal{S}) :=$ space of probability distributions over $\mathcal{S}$.
  - $\mathbb{P}(s'|s, a) =$ probability of transitioning to next state $s'$ given the current state is $s$ and action $a$ is taken

- $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the *reward function*
  - $r(s, a) =$ one-step (deterministic) reward when current state is $s$ and action is $a$.

- $\gamma \in [0, 1)$: *discount factor*.
Policy and MDP Dynamics

- Agent adopts a *stochastic* / randomized policy \( \pi : S \rightarrow \Delta(A) \)
  - \( \pi(a|s) = \) probability of taking action \( a \) at state \( s \)

- Initial state \( s_0 \sim \mu \in \Delta(S) \)

- At each time \( t \), agent observes state \( s_t \), takes action \( a_t \sim \pi(\cdot|s_t) \), and receives reward \( r_t = r(s_t, a_t) \). System then transitions to state \( s_{t+1} \sim P(\cdot|s_t, a_t) \) at time \( t+1 \)
Fix a policy \( \pi \)

The **value function** \( V^\pi : S \rightarrow \mathbb{R} \) is defined to be

\[
V^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s \right]
\]

The **action-value (Q-value) function** \( Q^\pi : S \times A \rightarrow \mathbb{R} \) is defined as

\[
Q^\pi(s, a) := \mathbb{E}_\pi \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \middle| s_0 = s, a_0 = a \right]
\]

**Relationship:**

\[
V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot|s)}[Q^\pi(s, a)] \tag{1}
\]

The objective is to find a policy \( \pi \) that maximizes \( V^\pi(s) \)
Let $\Pi$ be the set of all randomized policies. Define the **optimal value** and **Q functions** as

\[
V^*(s) := \sup_{\pi \in \Pi} V^\pi(s), \quad \text{for each } s \in S,
\]
\[
Q^*(s, a) := \sup_{\pi \in \Pi} Q^\pi(s, a), \quad \text{for each } s \in S, a \in A.
\]

**Theorem 1 (Puterman ’94, Thm 6.2.7)**

Assume $S$ and $A$ are finite. There exists a (deterministic) policy $\pi^*$ such that

\[
V^{\pi^*}(s) = V^*(s), \quad \forall s \in S,
\]
\[
Q^{\pi^*}(s, a) = Q^*(s, a), \quad \forall s \in S, a \in A
\]

$\pi^*$ is called an **optimal policy**.

- $\pi^*$ maximizes $V^\pi(s)$ **simultaneously** for all $s \in S$. 
**Theorem 2 (Bellman Equations for \( \pi \))**

For each policy \( \pi : S \rightarrow \Delta(A) \), \( V^{\pi} \) and \( Q^{\pi} \) are the unique functions that satisfy

\[
V^{\pi}(s) = \mathbb{E}_{a \sim \pi(s), s' \sim P(\cdot | s, a)} [r(s, a) + \gamma V^{\pi}(s')] 
\]
(2)

\[
Q^{\pi}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a), a' \sim \pi(\cdot | s')} [Q^{\pi}(s', a')]. 
\]
(3)

**Theorem 3 (Bellman Optimality Equations)**

\( V^* \) and \( Q^* \) are the unique functions that satisfy

\[
V^*(s) = \max_{a \in A} \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} V^*(s') \right], \quad \forall s \in S. 
\]
(4)

\[
Q^*(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot | s, a)} \left[ \max_{a' \in A} Q^*(s', a') \right], \quad \forall s \in S, a \in A. 
\]
(5)
Value Iteration

- When $r$ and $\mathbb{P}$ are known, $Q^*$ can be computed using **Value Iteration**:
  - Specify an initial function $Q^{(0)} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
  - For $k = 0, 1, \ldots$, compute
    \[
    Q^{(k+1)}(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(.|s, a)} \left[ \max_{a' \in \mathcal{A}} Q^{(k)}(s', a') \right], \text{ for each } s, a. \tag{6}
    \]

- Note: $Q^*$ is a fixed point of (6) by Bellman Optimality Equation

- Value Iteration converges to $Q^*$ geometrically:
  \[
  \left\| Q^{(k+1)} - Q^* \right\|_{\infty} \leq \gamma^k \left\| Q^{(0)} - Q^* \right\|_{\infty}.
  \]

- Given $Q^*$, an optimal deterministic policy can be computed:
  \[
  \pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a), \quad s \in \mathcal{S}.
  \]
Finite-Horizon MDP

- \( M = (S, A, \mathbb{P}, r, H) \).
- \( S \): state space; \( A \): action space
- **\( H \): horizon** (number of steps in an *episode*)
- \( \mathbb{P} = (\mathbb{P}_h : h = 1, \ldots, H) \) are the *transition kernels*
  - \( \mathbb{P}_h(s'|s,a) \) = probability of next state \( s' \) when current state is \( s \) and action \( a \) is taken at step \( h \)
- \( r = (r_h : h = 1, \ldots, H) \) are the *reward functions*
  - \( r_h(s,a) \) = one-step reward when current state-action is \( (s,a) \) at step \( h \)
- Policy \( \pi = (\pi_h : h = 1, \ldots, H) \)
  - \( \pi_h(a|s) \) = probability of taking action \( a \) at state \( s \) and step \( h \)
- Value and Q functions: for \( h = 1, \ldots, H \)
  - \( V_\pi^h(s) := \mathbb{E}_\pi \left[ \sum_{t=h}^{H} r_t(s_t, a_t)|s_h = s \right] \), \( Q_\pi^h(s,a) := \mathbb{E}_\pi \left[ \sum_{t=h}^{H} r_t(s_t, a_t)|s_h = s, a_h = a \right] \)
- \( \pi^*, V_\pi^*, Q_\pi^* \) defined analogously