CS 839: Probability & Learning in High Dimension

Lecture 18: Markov Decision Processes

Yudong Chen

UW-Madison CS

April 6, 2022

Reading

- 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)

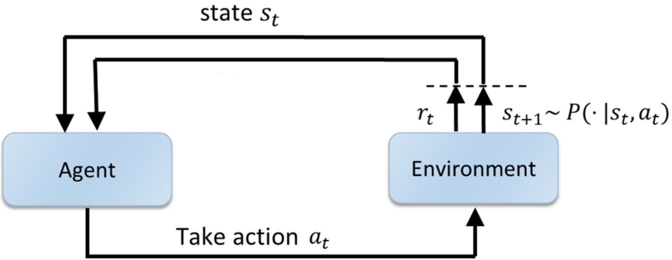
- Interaction between agent and environment is described by an MDP $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$.
- $\mathcal{S}$: *state space*; finite
- $\mathcal{A}$: *action space*; finite
- $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the *transition kernel*, 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
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 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 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$
  - $\pi(a|s)$ = probability of taking action a at state s

- Initial state $s_0 \sim \mu \in \Delta(\mathcal{S})$

- 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 $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$ at time $t + 1$



Value and Q-Functions

- Fix a policy π
- The **value function** $V^\pi : \mathcal{S} \to \mathbb{R}$ is defined to be

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

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

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

- Relationship:

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

- The objective isto find a policy π that maximizes $V^\pi(s)$

Optimal Value Function and Policy

- Let Π be the set of all randomized policies. Define the **optimal value and Q functions** as

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

Theorem 1 (Puterman '94, Thm 6.2.7)

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

$$V^{\pi^*}(s) = V^*(s), \qquad \forall s \in \mathcal{S},$$
$$Q^{\pi^*}(s,a) = Q^*(s,a), \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}$$

π^* is called an ***optimal policy****.*

- π^* maximizes $V^\pi(s)$ *simultaneously* for all $s \in \mathcal{S}$.

Bellman Equations

Theorem 2 (Bellman Equations for π)

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

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

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

Theorem 3 (Bellman Optimality Equations)

V^* and Q^* are the unique functions that satisfy

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

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

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

$$Q^{(k+1)}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q^{(k)}(s',a') \right], \quad \text{for each } s, a. \tag{6}$$

- 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 computed:

$$\pi^*(s) = \operatorname*{argmax}_{a \in \mathcal{A}} Q^*(s,a), \quad s \in \mathcal{S}.$$

Finite-Horizon MDP

- $M = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \boldsymbol{H})$.
- $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $\boldsymbol{H}$**: 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** $\boldsymbol{h}$
- $r = (r_h : h = 1, \ldots, H)$ are the *reward functions*
  - $r_h(s,a)$ = one-step reward when current state-action is (s,a) a **at step** $\boldsymbol{h}$
- Policy $\pi = (\pi_h : h = 1, \ldots H)$
  - $\pi_h(a|s)$ = probability of taking action a at state s and **step** $\boldsymbol{h}$
- Value and Q functions: for $h = 1, \ldots, H$

$$V_h^\pi(s) := \mathbb{E}_\pi\Big[\sum_{t=h}^{H} r_t(s_t, a_t) | s_h = s\Big], \quad Q_h^\pi(s,a) := \mathbb{E}_\pi\Big[\sum_{t=h}^{H} r_t(s_t, a_t) | s_h = s, a_h = a\Big]$$

- π^*, V_h^*, Q_h^* defined analogously