COMP3211 Tutorial 7: Markov Decision Process

Fengming ZHU

Mar. 24, 2022

Department of CSE
HKUST

Outline

MDP V.S. Search

Value Functions

Bellman Expectation Equation

Bellman Optimality Equation

MDP V.S. Search

MDP V.S. Search

Search:

- A set of states $\mathcal{S}$, initial state I, goal state G
- A set of actions $\mathcal{A}$
- Deterministic transitions $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$
- cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Objective: a path p from I to G that minimizes $c(p)$

MDP V.S. Search

Search:

- A set of states $\mathcal{S}$, initial state I, goal state G
- A set of actions $\mathcal{A}$
- Deterministic transitions $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$
- cost function $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
- Objective: a path p from I to G that minimizes $c(p)$

MDP:

- A set of states $\mathcal{S}$, a terminating condition $End(s)$
- A set of actions $\mathcal{A}$
- Stochastic transitions $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
- Reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, with a discount factor γ
- Objective: maximize $\sum_t \gamma^t r_t$

Solution Concept: Policy

Question:
To make sure you can come up with an optimal solution, you'd like to know:

$S_0 \xrightarrow{a_0} S_1 \xrightarrow{a_1} \boxed{S_2} \xrightarrow{a_2}$

- (A) Only your current state S_2
- (B) All the history from the beginning up until now S_0, a_0, S_1, a_1, S_2
- (C) Need to know more

Solution Concept: Policy

Question:
To make sure you can come up with an optimal solution, you'd like to know:

- (A) Only your current state
- (B) All the history from the beginning up until now
- (C) Need to know more

Theorem:
Markov property holds: $P[S_{t+1}|S_t, a_t, \boxed{\cdots, S_0, a_0}] = P[S_{t+1}|S_t, a_t]$.
That is, your current state is already a "sufficient statistic", also known as the information state.

Solution Concept: Policy

Question:
To make sure you can come up with an optimal solution, you'd like to know:

- (A) Only your current state
- (B) All the history from the beginning up until now
- (C) Need to know more

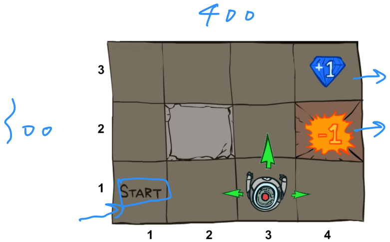
Theorem:
Markov property holds: $P[S_{t+1}|S_t, a_t, \cdots, S_0, a_0] = P[S_{t+1}|S_t, a_t]$.
That is, your current state is already a "sufficient statistic", also known as the information state.

Policy:
A solution is a policy $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$

Follow-up Question: Maze

Question:
Given a large maze, you (with deterministic actions U/D/L/R) are supposed to find a nice way from the entrance to the exit, which agent you'd like to choose

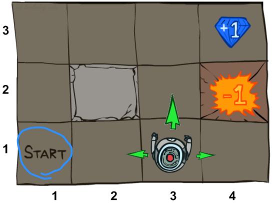
- (A) State machines with infite memory
- (B) Agents that can A^* search
- (C) Agents that can compute policies
- (D) None of them

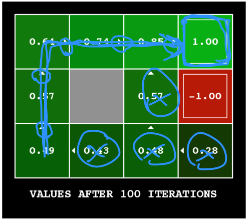


Follow-up Question: Maze

Question:
Given a large maze, you (with deterministic actions U/D/L/R) are supposed to find a nice way from the entrance to the exit, which agent you'd like to choose

- (A) State machines with infite memory
- (B) Agents that can A^* search
- (C) Agents that can compute policies
- (D) None of them





Value Functions

Notations with Time Index

$\downarrow t \times$ $t \downarrow$

- Transition: $T^a_{s,s'} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
- Reward: $R^a_s = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$ $R_{t+1} \sim \mathbb{P}(S, a)$
- Stationary policy: $\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$
- Return: The return G_t is the total discounted reward from time t,

$G_t = R_{t+1} + \gamma G_{t+1}$

Time-dependent

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Policy: $\pi_t(a|s) = \mathbb{P}[A_t = a | S_t = s]$

$v(s)$ and $q(s, a)$

$\pi_1 \rightarrow \pi_2 \rightarrow \cdots \rightarrow \pi_*$

State-value function:
The state-value function $v_\pi(s)$ for an MDP is the expected return starting from state s, and then following policy π,

$$\boxed{v_\pi(s)} = \boxed{\mathbb{E}_\pi}[G_t | S_t = s]$$

$\pi(a_1|s)$ s_1'

s

$\pi(a_2|s)$ s_2'

Action-value function:
The action-value function $q_\pi(s, a)$ for an MDP is the expected return starting from state s, taking action a, and then following policy π,

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$$

π v, q

Bellman Expectation Equation

Adam's Law:
For any random variables X and Y,

$$E[E[Y|X]] = E[Y]$$

Adam's Law with Extra Conditioning:
For any random variables X, Y and Z,

$$E[E[Y|X, Z]|Z] = E[Y|Z]$$

$$\hat{E}[\hat{E}[Y|X]] = \hat{E}[Y] = E[Y|Z]$$

$$\hat{E}(\cdot) = E(\cdot|Z)$$

Bellman Expectation Equation

- For state-value function,

$$v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$$

- For action-value function,

$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

Bellman Expectation Equation for $v^{\pi}(s)$

We first prove the Bellman equation for state-value function.

$$
\begin{aligned}
v_\pi(S_t = s) &= E_\pi[\boxed{G_t}|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots)|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s]
\end{aligned}
$$

Linearity

Since

$Y \quad X \quad Z \quad Z$

$$
\begin{aligned}
E_\pi[G_{t+1}|S_t] &= E_\pi[E_\pi[G_{t+1}|S_{t+1}, S_t]|S_t] \\
&= E_\pi[E_\pi[G_{t+1}|S_{t+1}]|S_t] \\
&= E_\pi[v_\pi(S_{t+1})|S_t]
\end{aligned}
$$

Adam Law.

Markov

Thus,

$$
\begin{aligned}
v(S_t = s) &= E_\pi[G_t|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\
&= E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]
\end{aligned}
$$

$= E_\pi[R_{t+1}] + \gamma E(G_{t+1}|S_t = s)$

Bellman Expectation Equation for $q^\pi(s, a)$

We then prove the Bellman equation for action-state function.

$$\begin{aligned} q_\pi(S_t = s, A_t = a) &= E_\pi[G_t | S_t = s, A_t = a] \\ &= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \end{aligned}$$

Since

$$\begin{aligned} E[G_{t+1} | S_t, A_t] &= E[E[\overset{Y}{G_{t+1}} | (\overset{X}{S_{t+1}, A_{t+1}}), \cancel{(\overset{Z}{S_t, A_t})}] | (\overset{Z}{S_t, A_t})] \quad \text{Adam} \\ &= E[E[G_{t+1} | (S_{t+1}, A_{t+1})] | (S_t, A_t)] \quad \text{Markov} \end{aligned}$$

Under policy π, we have

$$\begin{aligned} E_\pi[G_{t+1} | S_t, A_t] &= E_\pi[E_\pi[G_{t+1} | (S_{t+1}, A_{t+1})] | (S_t, A_t)] \\ &= E_\pi[q_\pi(S_{t+1}, A_{t+1}) | (S_t, A_t)] \end{aligned}$$

Thus,

$$\begin{aligned} q_\pi(S_t = s, A_t = a) &= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \end{aligned}$$

$\pi_0 \rightarrow \underline{v \cdot q} \rightarrow \underline{v \cdot q} \rightarrow \underline{?}$

definition Equation Solve

$\underline{?} \downarrow$

π_1

Bellman Optimality Equation

$\boxed{\underline{? \ \pi_*}} \rightarrow \underline{v \cdot q} \rightarrow \underline{v \cdot q} \rightarrow \underline{?}$

definition Equation solve

Optimal Value Function

?

π_1: S_1, S_2 $\quad$ π_2: S_3, S_4

$V_* = V_{\pi_1}$ $\quad$ $V_* = V_{\pi_2}$

- For state-value function,

$V_*(S_1) = V_{\pi_1}(S_1)$

$V_*(S_2) = V_{\pi_3}(S_3)$

$$v_*(s) = \max_\pi \boxed{v_\pi(s)}$$

- For action-value function,

$$q_*(s, a) = \max_\pi \boxed{q_\pi(s, a)}$$

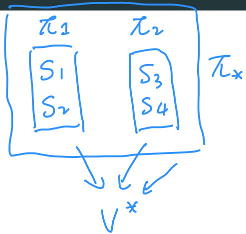
- The optimal value function specifies the best possible performance in the MDP.
- An MDP is “solved” once we know the optimal values.

Optimal Policy

Define a partial ordering over policies:

$$\pi \geq \pi', \text{ if } v_\pi(s) \geq v'_\pi(s), \text{ for all } s.$$

Optimal Policy



Define a partial ordering over policies:

$$\pi \geq \pi', \text{ if } v_\pi(s) \geq v'_\pi(s), \text{ for all } s.$$

For any MDP:

- There exists an optimal policy π_* that is better than or equal to all other policies, $\pi_* \geq \pi$, for all π.
- All optimal policies achieve the optimal state-value function, $v_{\pi_*}(s) = v_*(s)$, for all s.
- All optimal policies achieve the optimal action-value function, $q_{\pi_*}(s, a) = q_*(s, a)$, for all s, a.

Finding Optimal Policy

$V/q_* \rightarrow \pi_* \rightarrow \pi_*^{deterministic}$!

Theorem:

- An optmial policy can be found by maximizing over $q_*(s, a)$,

$$\pi_*(a|s) = \begin{cases} 1 & \text{, if } a = argmax_{a \in A} q_*(s, a) \\ 0 & \text{, otherwise} \end{cases}$$

- There is always a deterministic optimal policy for any MDP.
- If we know $q_*(s, a)$, we immediately have the optimal policy.

$V_*(s) = \max_a q_*(s, a)$

Bellman Optimality Equation

- For state-value function,

$$v_*(s) = \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

- For action-value function,

$$q_*(s, a) = E[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a]$$

$q \sim G_t \mid s, a$

Bellman Optimality Equation for v_* and q_*

Under the optimal policy, we first show the relation of v^* and q^*.

- v_* in terms of q_*,

$$\boxed{v_*(s) = \max_a q_*(s, a)}$$

Bellman Optimality Equation for v_* and q_* (cont'd)

- q_* in terms of v_*,

$$
\begin{aligned}
q_*(s,a) &= \max_{\pi} q_\pi(s,a) \\
&= R_s^a + \gamma \sum_{s'} T_{ss'}^a \max_{\pi} v_\pi(s') \\
&= R_s^a + \gamma \sum_{s'} T_{ss'}^a v_*(s') \\
&= E[R_{t+1}|S_t = s, A_t = a] \\
&+ \gamma \sum_{s'} \{P(S_{t+1} = s'|S_t = s, A_t = a) \\
&\cdot E[v_*(S_{t+1})|S_{t+1} = s', S_t = s, A_t = a]\} \\
&= E[R_{t+1}|S_t = s, A_t = a] + \gamma E[v_*(S_{t+1})|S_t = s, A_t = a] \\
&= E[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a]
\end{aligned}
$$

Bellman Optimality Equation for V_* and Q_*

Then we show the Bellman optimal equation for v^*,

$$
\begin{aligned}
v_*(s) &= \max_a \boxed{q_*(s, a)} \quad \text{obvious} \\
&= \max_a E[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]
\end{aligned}
$$

Finally, we show the Bellman equation for q^*,

$$
\begin{aligned}
q_*(s, a) &= E[R_{t+1} + \gamma \underline{v_*(S_{t+1})} | S_t = s, A_t = a] \\
&= E[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') | S_t = s, A_t = a]
\end{aligned}
$$

$\pi_0 \to$ BEE $\to$ $\underline{V, q}$

solve.

PI ✓

VI ✓

Thanks!

$\pi_* \to$ BOE $\to$ $\underline{V_*, q_*}$

solve