Optimality in Policies

Easwar Subramanian

TCS Innovation Labs, Hyderabad

Email : easwar.subramanian@tcs.com / cs5500.2020@iith.ac.in

August 19, 2022

Overview

Review

Markov Decision Process

Markov decision process is a tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ where

- $\mathcal{S}$: (Finite) set of states
- $\mathcal{A}$: (Finite) set of actions
- $\mathcal{P}$: State transition probability

$$\mathcal{P}^a_{ss'} = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a), a_t \in \mathcal{A}$$

- $\mathcal{R}$: Reward for taking action a_t at state s_t and transitioning to state s_{t+1} is given by the deterministic function $\mathcal{R}$

$$r_{t+1} = \mathcal{R}(s_t, a_t, s_{t+1})$$

- γ : Discount factor such that $\gamma \in [0, 1]$

Policy

Let π denote a policy that maps state space $\mathcal{S}$ to action space $\mathcal{A}$

Policy

- Deterministic policy: $a = \pi(s), s \in \mathcal{S}, a \in \mathcal{A}$
- Stochastic policy $\pi(a|s) = P[a_t = a|s_t = s]$

Value Functions with Policy

Given a MDP and a policy π, we define the value of a policy as follows :

State-value function

The value function $V^\pi(s)$ in state s is the expected (discounted) total return starting from state s and then following the policy π

$$V^\pi(s) = \mathbb{E}_\pi\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right)$$

The state-value function can be decomposed into immediate reward plus discounted value of successor state

$$V^\pi(s) = \mathbb{E}_\pi(r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s)$$

Action Value Function

Action-value function

The action-value function $Q(s,a)$ under policy π is the expected return starting from state s and taking action a and then following the policy π

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\right)$$

The action-value function can similarly be decomposed as

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}(r_{t+1} + \gamma Q^{\pi}(s_{t+1}, a_{t+1}) | s_t = s, a_t = a)$$

Relationship between $V^\pi(\cdot)$ and $Q^\pi(\cdot)$

Using definitions of $V^\pi(s)$ and $Q^\pi(s,a)$, we can arrive at the following relationships

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s,a)$$

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

Optimality in Policies

Solution to an MDP

Solving an MDP <u>means</u> finding a policy π_* as follows

$$\pi_* = \arg\max_{\pi} \left[\mathbb{E}_{\pi} \left(\sum_{t=0}^{\infty} \gamma^t r_{t+1} \right) \right]$$

is **maximum**

- Denote optimal value function $V_*(s) = V^{\pi_*}(s)$
- Denote optimal action value function $Q_*(s,a) = Q^{\pi_*}(s,a)$
- The main goal in RL or solving an MDP means finding an **optimal value function** V_* or **optimal action value function** Q_* or **optimal policy** π_*

Optimal Policy

Define a partial ordering over policies

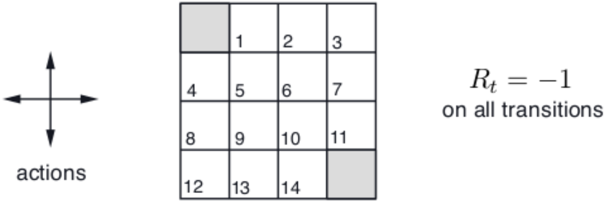
$$\pi \geq \pi', \quad \text{if} \quad V^{\pi}(s) \geq V^{\pi'}(s), \quad \forall s \in \mathcal{S}$$

Theorem

- There exists an optimal policy π_* that is better than or equal to all other policies.
- All optimal policies achieve the optimal value function, $V_*(s) = V^{\pi_*}(s)$
- All optimal policies achieve the optimal action-value function, $Q_*(s, a) = Q^{\pi_*}(s, a)$

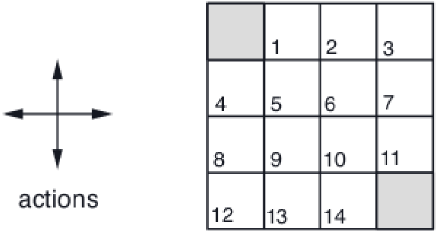
Grid World Problem

Consider a 4×4 grid world problem



- $\mathcal{S}$: $\{1, 2, \cdots, 14\}$ (non-terminal) + 2 terminal states (shaded)
- $\mathcal{A}$: {East, West, North, South}
- $\mathcal{P}$: Upon choosing an action from $\mathcal{A}$, state transitions are deterministic; except the actions that would take the agent off the grid in fact leave the state unchanged
- $\mathcal{R}$: Reward is -1 on all transitions until the terminal state is reached

Grid World Problem



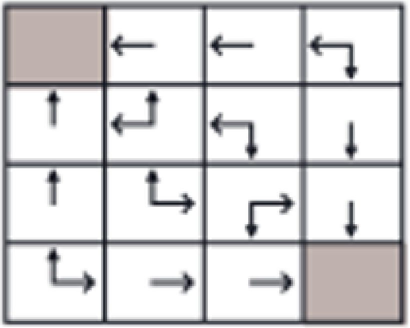
$$R_t = -1$$

on all transitions

Goal : Reach any of the goal state in as minimum plays as possible

Question : What could be an optimal policy to achieve the above objective ?

Grid World Problem : Optimal Policies



Question : How many optimal policies are there ?

Answer : There are infinite optimal policies (including some deterministic ones)

Finding an Optimal Policy

Question : Suppose we are given $Q_*(s,a)$. Can we find an optimal policy ?

Answer : An optimal policy can be found by maximising over $Q_*(s,a)$

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} Q_*(s,a) \\ 0 & \text{Otherwise} \end{cases}$$

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

Greedy (Optimal) Policy

Greedy policy with respect to optimal (action) value function is an optimal policy

An optimal policy can be found by maximising over $Q_*(s,a)$

$$\pi_*(s) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} Q_*(s,a) \\ 0 & \text{Otherwise} \end{cases}$$

Greedy Policy

For a given $Q^\pi(\cdot,\cdot)$, define $\pi'(s)$ as follows

$$\pi'(s) = \text{greedy}(Q) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} Q^\pi(s,a) \\ 0 & \text{Otherwise} \end{cases}$$

For a given $V^\pi(\cdot)$, define $\pi'(s)$ as follows

$$\pi'(s) = \text{greedy}(V) = \begin{cases} 1 & \text{if } a = \arg\max_{a \in \mathcal{A}} \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right)\right] \\ 0 & \text{Otherwise} \end{cases}$$

Relationship between $V_*(\cdot)$ and $Q_*(\cdot,\cdot)$

Question : Suppose we are given $Q_*(s,a), \forall s \in \mathcal{S}$. Can we find $V_*(s)$?

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

Question : Suppose we are given $V_*(s), \forall s \in \mathcal{S}$. Can we find $Q_*(s,a)$?

$$Q_*(s,a) = \left[\sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \left(\mathcal{R}^a_{ss'} + \gamma V_*(s')\right)\right]$$

Policy Iteration

Policy Iteration

Question : Is there a way to arrive at π_* starting from an arbitrary policy π ?

Answer : Policy Iteration

- Evaluate the policy π
  - ★ Compute $V^\pi(s) = \mathbb{E}_\pi(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s)$
- Improve the policy π

$$\pi'(s) = \text{greedy}(V^\pi(s))$$

$$\pi_0 \xrightarrow{\text{E}} V^{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} V^{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi^* \xrightarrow{\text{E}} V^*,$$

Policy Evaluation

- **Problem** : Evaluate a given policy π
- Compute $V^{\pi}(s) = \mathbb{E}_{\pi}(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s)$
- **Solution 1** : Solve a system of linear equations using any solver
- **Solution 2** : Iterative application of Bellman Evaluation Equation
- Iterative update rule :

$$V^{\pi}_{k+1}(s) \leftarrow \sum_{a} \pi(a|s) \sum_{s'} \mathcal{P}^{a}_{ss'} \left[\mathcal{R}^{a}_{ss'} + \gamma V^{\pi}_{k}(s')\right]$$

- The sequence of value functions $\{V^{\pi}_1, V^{\pi}_2, \cdots, \}$ converge to V^{π}

Policy Improvement

Suppose we know V^π. How to improve policy π ?

The answer lies in the definition of action value function $Q^\pi(s,a)$. Recall that,

$$
\begin{aligned}
Q^\pi(s,a) &= \mathbb{E}_\pi\left(\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a\right) \\
&= \mathbb{E}(r_{t+1} + \gamma V^\pi(s_{t+1})|s_t = s, a_t = a) \\
&= \sum_{s'\in\mathcal{S}} \mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right]
\end{aligned}
$$

- If $Q^\pi(s,a) > V^\pi(s) \implies$ Better to select action a in state s and thereafter follow the policy π
- This is a special case of the policy improvement theorem

Policy Improvement Theorem

Theorem

Let π and π' be any pair of deterministic policies such that, for all $s \in \mathcal{S}$,

$$Q^\pi(s, \pi'(s)) \geq V^\pi(s).$$

Then $V^{\pi'}(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$

Proof.

$$\begin{aligned}
V^\pi(s) &\leq Q^\pi(s, \pi'(s)) = \mathbb{E}_{\pi'}(r_{t+1} + \gamma V^\pi(s_{t+1}) | s_t = s) \\
&\leq \mathbb{E}_{\pi'}(r_{t+1} + \gamma Q^\pi(s_{t+1}, \pi'(s_{t+1})) | s_t = s) \\
&= \mathbb{E}_{\pi'}(r_{t+1} + \gamma r_{t+2} + \gamma^2 V^\pi(s_{t+2}) | s_t = s) \\
&\leq \mathbb{E}_{\pi'}(r_{t+1} + \gamma r_{t+2} + \gamma^2 Q^\pi(s_{t+2}, \pi'(s_{t+2})) | s_t = s) \\
&\leq \mathbb{E}_{\pi'}(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s) = V^{\pi'}(s)
\end{aligned}$$

□