Policy Evaluation

Easwar Subramanian

TCS Innovation Labs, Hyderabad

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

August 16, 2022

Overview

1. Review
2. Policy
3. Policy Evaluation

Review

Markov Reward Process

Markov Reward Process

A Markov reward process is a tuple $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ is a Markov chain with values

- $\mathcal{S}$: (Finite) set of states
- $\mathcal{P}$: State transition probablity
- $\mathcal{R}$: Reward for being in state s_t is given by a deterministic function $\mathcal{R}$

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

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

- In general, the reward function can also be an expectation $\mathcal{R}(s_t = s) = \mathbb{E}[r_{t+1} | s_t = s]$

Value Function

The value function $V(s)$ gives the long-term value of state $s \in \mathcal{S}$

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

- Value function $V(s)$ determines the value of being in state s
- $V(s)$ measures the potential future rewards we may get from being in state s
- Observe that $V(s)$ is independent of t

Decomposition of Value Function

Let s and s' be successor states at time steps t and $t+1$, the value function can be decomposed into sum of two parts

- Immediate reward r_{t+1}
- Discounted value of next state s' (i.e. $\gamma V(s')$)

$$
\begin{aligned}
V(s) = \mathbb{E}\left(G_t | s_t = s\right) &= \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right) \\
&= \mathbb{E}\left(r_{t+1} + \gamma V(s_{t+1}) | s_t = s\right)
\end{aligned}
$$

Bellman equation for value functions

$$
\boxed{V(s) = \mathbb{E}(r_{t+1} | s_t = s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')}
$$

Bellman Equation in Matrix Form

We have $\mathcal{S} = \{1, 2, \cdots, n\}$ and let $\mathcal{P}$, $\mathcal{R}$ be known. Then one can write the Bellman equation can as,

$$V = \mathcal{R} + \gamma \mathcal{P} V$$

where

$$\begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}(1) \\ \mathcal{R}(2) \\ \vdots \\ \mathcal{R}(n) \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \cdots & \mathcal{P}_{1n} \\ \mathcal{P}_{21} & \mathcal{P}_{22} & \cdots & \mathcal{P}_{2n} \\ \vdots & & & \\ \mathcal{P}_{n1} & \mathcal{P}_{n2} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \times \begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(n) \end{bmatrix}$$

Solving for V, we get,

$$\boxed{V = (I - \gamma \mathcal{P})^{-1} \mathcal{R}}$$

The discount factor should be $\gamma < 1$ for the inverse to exist

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

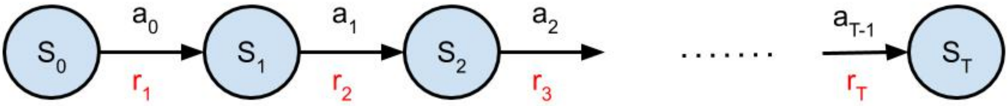
$$\mathcal{P}^a_{ss'} = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a), a_t \in \mathcal{A}$$

- $\mathcal{R}$: 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 $\gamma \in [0, 1]$

Finite and Infinite Horizon MDPs



Depending on time horizon, a Markov decision process can be

- Finite horizon
- Infinite horizon
- Indefinite horizon (SSP)

For finite and (certain) indefinite MDPs with at least absorbing state, we can take the discount factor to be 1

Policy

Policy

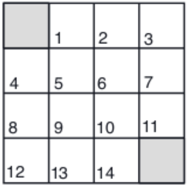
Let π denote a policy that maps state space $\mathcal{S}$ to action space $\mathcal{A}$

Policy

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

Grid World : Revisited

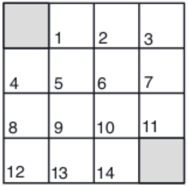
Consider a 4×4 grid world problem



- $\mathcal{S} : \{1, 2, \cdots, 14\}$ (non-terminal) + 2 terminal states (shaded)
- $\mathcal{A} : \{\text{Right, Left, Up, Down}\}$

Figure Source: David Silver's RL course

Grid World : Deterministic Policy



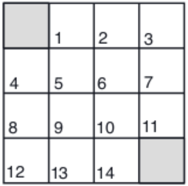
- $\mathcal{S} : \{1, 2, \cdots, 14\}$ (non-terminal) + 2 terminal states (shaded)
- $\mathcal{A} : \{\text{Right, Left, Up, Down}\}$
- **Deterministic policy** :

$$\pi(s) = \left\{ \begin{array}{ll} \text{Down,} & \text{if } s = \{3, 7, 11\} \\ \text{Right,} & \quad\quad \text{Otherwise} \end{array} \right\}$$

- Example sequences : $\{\{8, 9, 10, 11, G\}, \{2, 3, 7, 11, G\}\}$

Figure Source: David Silver's RL course

Grid World : Stochastic Policy



- $\mathcal{S}$: $\{1, 2, \cdots, 14\}$ (non-terminal) + 2 terminal states (shaded)
- $\mathcal{A}$: {Right, Left, Up, Down}
- **Stochastic policy** : $\pi(a|s)$ could be a uniform random action between all available actions at state s
- Example sequences : $\{\{8, 4, 8, 9, 13, \cdots, \}, \{2, 6, 5, 9, 13, \cdots, \}\}$

Figure Source: David Silver's RL course

Stochastic Policy : Rock Scissors Paper

- Two player game of rock-paper-scissors
  - ★ Scissors beats paper
  - ★ Rock beats scissors
  - ★ Paper beats rock
- Consider policies for iterated rock-paper-scissors
  - ★ A deterministic policy is easily exploited
  - ★ A uniform random policy is optimal (i.e. Nash equilibrium)

Policy Evaluation

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

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

Decomposition of State Value Function

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

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

Expanding the expectation, with $\mathcal{R}^a_{ss'} = \mathcal{R}(s, a, s')$ we get,

$$\mathbb{E}_\pi[r_{t+1} | s_t = s] = \sum_a \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'}$$

and

$$\mathbb{E}_\pi[\gamma V^\pi(s_{t+1}) | s_t = s] = \sum_a \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'} \gamma V^\pi(s')$$

Hence,

$$\boxed{V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma V^\pi(s')\right]}$$

The above equation is called the Bellman Evaluation operator

Matrix Formulation of Bellman Evaluation Equation

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

Denote,

$$\begin{aligned} \mathcal{P}^\pi(s'|s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \mathcal{P}^a_{ss'} \\ \mathcal{R}^\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s'} \mathcal{P}^a_{ss'} \mathcal{R}^a_{ss'} = \mathbb{E}(r_{t+1}|s_t = s) \end{aligned}$$

Using $\mathcal{P}^\pi$ and $\mathcal{R}^\pi$, for finite MDP, one can rewrite the Bellman evaluation equation as

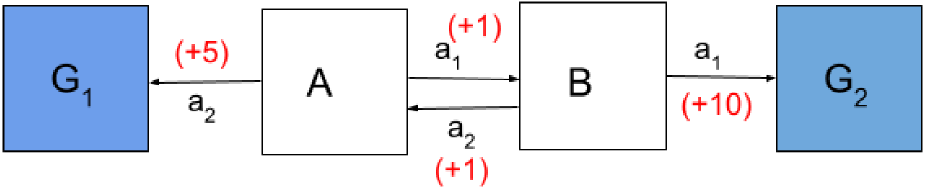
$$V^\pi = \mathcal{R}^\pi + \gamma \mathcal{P}^\pi V^\pi \implies V^\pi = (I - \gamma \mathcal{P}^\pi)^{-1} \mathcal{R}^\pi$$

Remark : Bellman Evaluation Equation for $V^\pi(s)$ is a system of $n = |\mathcal{S}|$ (linear) equations with n variables and can be solved if the model is known

MDP + Policy = MRP

- MDP + policy = Markov Reward Process.
- Given a MDP $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ and a policy π
- The MRP is given by $(\mathcal{S}, \mathcal{P}^{\pi}, \mathcal{R}^{\pi}, \gamma)$

Value Function Computation : Example



- States $\mathcal{S} = \{A, B, G_1, G_2\}$; States G_1 and G_2 are terminal states
- Two actions $\mathcal{A} = \{a_1, a_2\}$
- Value of states $\{A, B\}$ using forward policy π_f is given by, $V_{\pi_f}(A) = 11$, $V_{\pi_f}(B) = 10$
- Value of states $\{A, B\}$ using forward policy π_b is given by, $V_{\pi_b}(B) = 6$, $V_{\pi_b}(A) = 5$

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

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

Expanding the expectation we have $Q^{\pi}(s, a)$ to be

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

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

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

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