Policy Evaluation

Easwar Subramanian

TCS Innovation Labs, Hyderabad

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

August 30, 2021

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. $\gamma V(s')$)

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

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 $\gamma \in [0, 1]$

Policy

Policy

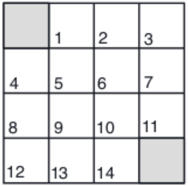
Let π denote a policy that maps state space $\mathcal{S}$ to action space $\mathcal{A}$

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

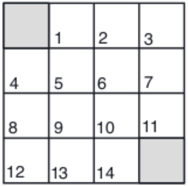
Consider a 4×4 grid world problem



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

Figure Source: David Silver's RL course

Grid World : Deterministic Policy



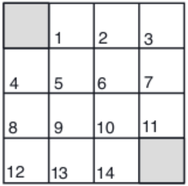
- $\mathcal{S}$: $\{1, 2, \cdots, 14\}$ (non-terminal) + 2 terminal states (shaded)
- $\mathcal{A}$: {Right, Left, Up, Down}
- **Deterministic policy** :

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

- 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)
- $\mathcal{A} : \{\text{Right, Left, Up, Down}\}$
- **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

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,

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

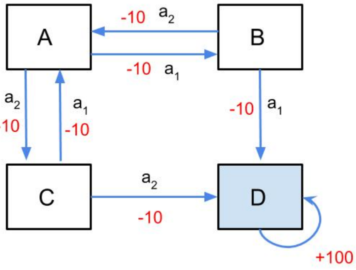
- MDP + policy = Markov Reward Process.
- The MRP is given by $(\mathcal{S}, \mathcal{P}^\pi, \mathcal{R}^\pi, \gamma)$

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

Value Function Computation : Example



- States $\mathcal{S} = \{A, B, C, D\}$; State D is terminal state
- Two actions $\mathcal{A} = \{a_1, a_2\}$
- Stochastic Environment with action chosen succeeding 90% and failing 10%
- Upon failure, agent moves in the direction suggested by the other action

Value Function Computation : Example

- Consider a deterministic policy (π_1) that chooses action a_1 in all states
- Transition matrix corresponding to policy π_1 is given by

$$
\begin{array}{c|cccc}
 & A & B & C & D \\
A & 0 & 0.9 & 0.1 & 0 \\
B & 0.1 & 0 & 0 & 0.9 \\
C & 0.9 & 0 & 0 & 0.1 \\
D & 0 & 0 & 0 & 1
\end{array}
$$

- Value of the states under the policy π_1 is given by,
  - ★ $V^{\pi_1}(D) = 100$
  - ★ $V^{\pi_1}(A) = 0.9 * [-10 + V^{\pi_1}(B)] + 0.1 * [-10 + V^{\pi_1}(C)]$
  - ★ $V^{\pi_1}(B) = 0.9 * [-10 + V^{\pi_1}(D)] + 0.1 * [-10 + V^{\pi_1}(A)]$
  - ★ $V^{\pi_1}(C) = 0.9 * [-10 + V^{\pi_1}(A)] + 0.1 * [-10 + V^{\pi_1}(D)]$
- $V^{\pi_1} = \{75.61, 87.56, 68.05, 100\};$

Value Function Computation : Example

- Consider a deterministic policy (π_2) that chooses action a_2 in all states
- Transition matrix corresponding to policy π_2 is given by

$$\begin{array}{c|cccc} & A & B & C & D \\ \hline A & 0 & 0.1 & 0.9 & 0 \\ B & 0.9 & 0 & 0 & 0.1 \\ C & 0.1 & 0 & 0 & 0.9 \\ D & 0 & 0 & 0 & 1 \end{array}$$

- Value of the states under the policy π_2 is given by,
  - $V^{\pi_2}(D) = 100$
  - $V^{\pi_2}(A) = 0.9 * [-10 + V^{\pi_2}(C)] + 0.1 * [-10 + V^{\pi_2}(D)]$
  - $V^{\pi_2}(B) = 0.9 * [-10 + V^{\pi_2}(A)] + 0.1 * [-10 + V^{\pi_2}(D)]$
  - $V^{\pi_2}(C) = 0.9 * [-10 + V^{\pi_2}(D)] + 0.1 * [-10 + V^{\pi_2}(A)]$
- $V^{\pi_2} = \{75.61, 68.05, 87.56, 100\};$