Value Functions and Markov Decision Process

Easwar Subramanian

TCS Innovation Labs, Hyderabad

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

August 26, 2021

Overview

1. Review
2. Value Function
3. Markov Decision Process

Review

Markov Property

Markov Property

A state s_t of a stochastic process $\{s_t\}_{t \in T}$ is said to have Markov property if

$$P(s_{t+1}|s_t) = P(s_{t+1}|s_1, \cdots, s_t)$$

The state s_t at time t captures all relevant information from history and is a sufficient statistic of the future

State Transition Matrix

State Transition Probability

For a Markov state s and a successor state s', the state transition probability is defined by

$$\mathcal{P}_{ss'} = P(s_{t+1} = s' | s_t = s)$$

State transition matrix $\mathcal{P}$ then denotes the transition probabilities from all states s to all successor states s' (with each row summing to 1)

$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \cdots & \mathcal{P}_{1n} \\ \vdots & & & \\ \mathcal{P}_{n1} & \mathcal{P}_{n2} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$$

Markov Chain

A stochastic process $\{s_t\}_{t \in T}$ is a **Markov process** or **Markov Chain** if it satisfies Markov property for every state s_t. It is represented by tuple $< \mathcal{S}, \mathcal{P} >$ where $\mathcal{S}$ denote the set of states and $\mathcal{P}$ denote the state transition probablity

No notion of <u>reward</u> or <u>action</u>

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 $\gamma \in [0, 1]$

No notion of <u>action</u>

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 $\mathcal{R}(s_t = s) = \mathbb{E}[r_{t+1} | s_t = s]$
- $\mathcal{R}(\cdot)$ can also be a function of current state s and next state s' , i.e. $r_{t+1} = \mathcal{R}(s, s')$

Value Function

Total Discounted Return

Recall the expression for total discounted return,

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- Note that the entitty G_t is a random variable
  - ★ We do not know what future states we will visit as the environment is stochastic; therefore future rewards are uncertain
  - ★ In a more general setting the function $\mathcal{R}$ can be stochastic function of state(s)
- Hence it is more meaningful to define a function that describes the long term value of the state in expected sense.

Value Function

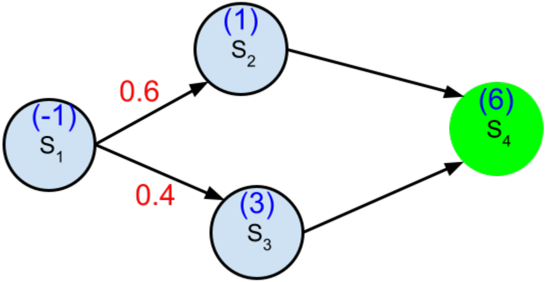
The value function $V(s)$ gives the long-term value of state $s \in \mathcal{S}$

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

- 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

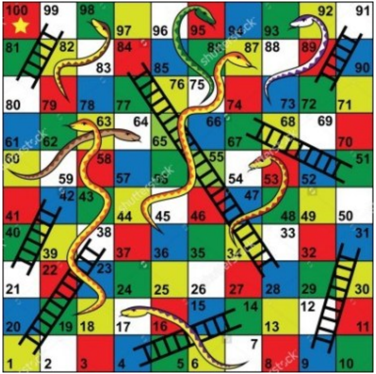
Value Function Computation : Example

Consider the following MRP. Assume $\gamma = 1$



- $V(s_4) = 6$
- $V(s_3) = 3 + \gamma * 6 = 9$
- $V(s_2) = 1 + \gamma * 6 = 7$
- $V(s_1) = -1 + \gamma * (0.6 * 7 + 0.4 * 9) = 6.8$

Example : Snakes and Ladders



Question : What could be a suitable interpretation to the value function for any state $s \in \mathcal{S}$ for snakes and ladders (if we define $\mathcal{R}(s) = -1$)?

Answer : Expected number of plays to reach the goal state s_{100}

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. $\gamma V(s')$)

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

Decomposition of Value Function

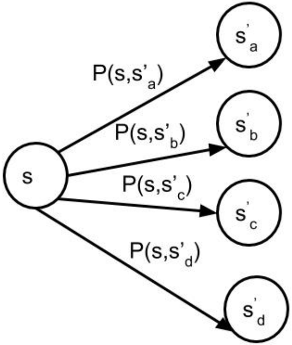
$$
\begin{aligned}
V(s) &= \mathbb{E}\left(G_t | s_t = s\right) = \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right) \\
&= \mathbb{E}\left(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s\right) \\
&= \mathbb{E}(r_{t+1} | s_t = s) + \gamma\, \mathbb{E}\left(G_{t+1} | s_t = s\right) \\
&= \mathbb{E}(r_{t+1} | s_t = s) + \gamma\, \mathbb{E}\left(\mathbb{E}(G_{t+1} | s_{t+1}) | s_t = s\right) \\
&= \mathbb{E}(r_{t+1} | s_t = s) + \gamma\, \mathbb{E}\left(V(s_{t+1}) | s_t = s\right) \\
&= \mathbb{E}(r_{t+1} + \gamma V(s_{t+1}) | s_t = s)
\end{aligned}
$$

We used <u>tower property of conditional expectations</u> in fourth step

Value Function : Evaluation

We have

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



$$V(s) = \mathcal{R}(s) + \gamma \left[\mathcal{P}_{ss_a'} V(s_a') + \mathcal{P}_{ss_b'} V(s_b') + \mathcal{P}_{ss_c'} V(s_c') + \mathcal{P}_{ss_d'} V(s_d') \right]$$

Bellman Equation for Markov Reward Process

TCS Research & Innovation

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

For any $s' \in \mathcal{S}$ a successor state of s with transition probability $\mathcal{P}_{ss'}$, we can rewrite the above equation as (using definition of Expectation)

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

This is the **Bellman Equation** for value functions

TCS TATA CONSULTANCY SERVICES

Bellman Equation in Matrix Form

Let $\mathcal{S} = \{1, 2, \cdots, n\}$ and $\mathcal{P}$ 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

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 $\gamma \in [0, 1]$

On Transition Matrices

Recall given an MDP $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, we have the state transition probability $\mathcal{P}$ defined as

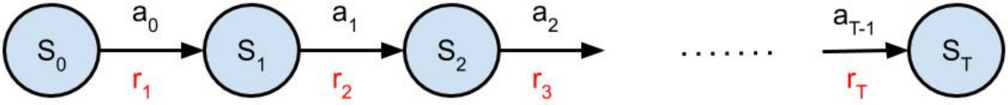
$$\mathcal{P}^{a}_{ss'} = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a), a_t \in \mathcal{A}$$

- In general, note that even after choosing action a at state s (as prescribed by the policy) the next state s' need not be a fixed state

Example : A slight modification of the grid world example can be as follows

- *Having chosen an action* a *at state* s
  - ★ The agent trip off to another neighbouring state (square) with small probability (say 0.1)

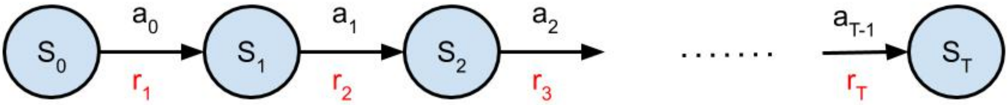
Flow Diagram



- The set of states belongs to $\mathcal{S}$
- The set of actions belongs to $\mathcal{A}$
- State transitions are governed by transition matrices $\mathcal{P}^a_{ss'}$
- Rewards are governed by the deterministic function $\mathcal{R}$ given by $r_{t+1} = \mathcal{R}(s_t, a_t, s_{t+1})$
- The goal is to choose a sequence of actions such that the total discounted future reward $\mathbb{E}(G_t | s_t = s)$ is maximized where

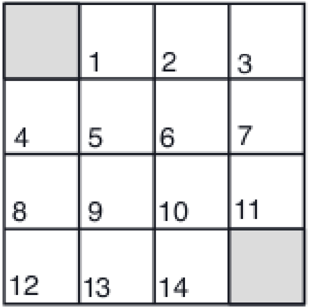
$$G_t = \sum_{k=0}^{\infty} \left(\gamma^k r_{t+k+1}\right)$$

Wealth Management Problem



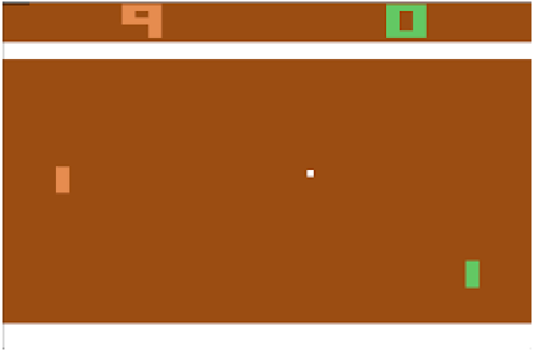
- States $\mathcal{S}$: Current value of the portfolio and current valuation of instruments in the portfolio
- Actions $\mathcal{A}$: Buy / Sell instruments of the portfolio
- Reward $\mathcal{R}$: Return on portfolio compared to previous decision epoch

Navigation Problem



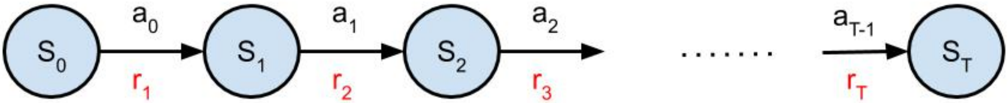
- States $\mathcal{S}$: Squares of the grid
- Actions $\mathcal{A}$: Any of the four directions possible
- Reward $\mathcal{R}$: -1 for every move made until reaching goal state

Example : Atari Games



- States $\mathcal{S}$: Possible set of all (Atari) images
- Actions $\mathcal{A}$: Move the paddle up or down
- Reward $\mathcal{R}$: +1 for making the opponent miss the ball; -1 if the agent miss the ball; 0 otherwise;

Finite and Infinite Horizon MDPs



- If T is fixed and finite, the resultant MDP is a finite horizon MDP
  - ★ Wealth management problem
- If N is infinite, the resultant MDP is infinite horizon MDP
  - ★ Certain Atari games
- When $|\mathcal{S}|$ is finite, the MDP is called finite state MDPs

Grid World Example

	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

Question : Is Grid world finite / infinite horizon problem ? Why ?

(Stochastic Shortest Path MDPs)

- For finite horizon MDPs and stochastic shortest path MDPs, one can use $\gamma = 1$