

Value Functions and Markov Decision Process

Easwar Subramanian

TCS Innovation Labs, Hyderabad

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

August 12, 2022

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

Review

Markov Property

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

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

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

State Transition 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)

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

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 $\langle \mathcal{S}, \mathcal{P} \rangle$ where \mathcal{S} denote the set of states and \mathcal{P} denote the state transition probability

No notion of reward or action

Markov Reward Process

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

- ▶ \mathcal{S} : (Finite) set of states
- ▶ \mathcal{P} : State transition probability
- ▶ \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 action

Markov Reward Process

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

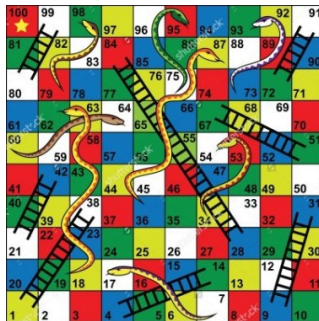
- ▶ \mathcal{S} : (Finite) set of states
- ▶ \mathcal{P} : State transition probability
- ▶ \mathcal{R} : Reward for being in state s_t is given by a deterministic function \mathcal{R}

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

- ▶ γ : Discount factor such that $\gamma \in [0, 1]$
- ▶ In general, the reward function can also be an expectation $\mathcal{R}(s_t = s) = \mathbb{E}[r_{t+1} | s_t = s]$

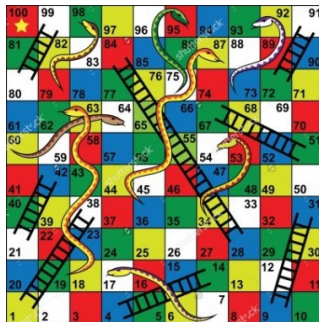
Value Function

Snakes and Ladders : Revisited



- **Reward \mathcal{R}** : $\mathcal{R}(s) = -1$ for $s \in s_1, \dots, s_{99}$ and for $\mathcal{R}(s_{100}) = 0$
- **Discount Factor $\gamma = 1$**

Snakes and Ladders : Revisited



Question : Are all intermediate states equally 'valuable' just because they have equal reward ?

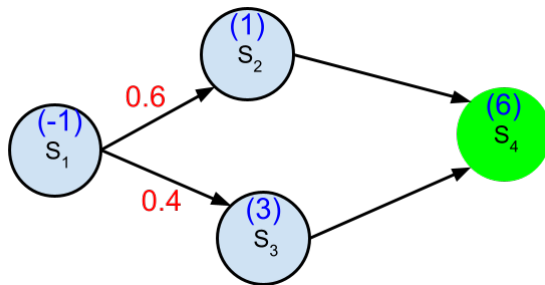
The value function $V(s)$ gives the long-term value of state $s \in \mathcal{S}$

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

- ▶ 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
- ▶ $V(s)$ is independent of t

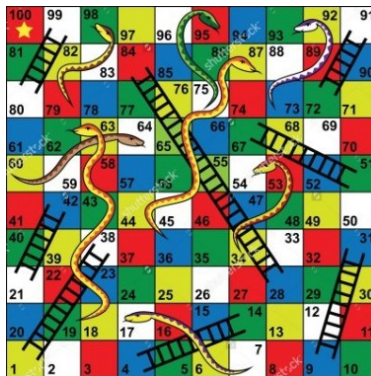
Value Function Computation : Example

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



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

Example : Snakes and Ladders



Question : How can we evaluate the value of each state in a large MRP such as 'Snakes and Ladders' ?

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. $\gamma V(s')$)

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

Decomposition of Value Function

Recall that,

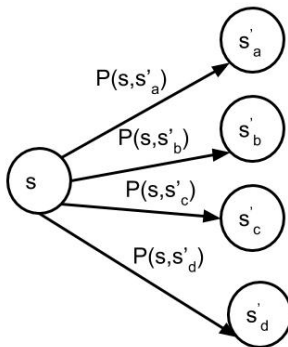
$$G_t = (r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots) = \sum_{k=0}^{\infty} (\gamma^k r_{t+k+1})$$

$$\begin{aligned} V(s) &= \mathbb{E}(G_t | s_t = s) = \mathbb{E}\left(\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right) \\ &= \mathbb{E}(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s_t = s) \\ &= \mathbb{E}(r_{t+1} | s_t = s) + \sum_{k=1}^{\infty} \gamma^k \mathbb{E}(r_{t+k+1} | s_t = s) \\ &= \mathbb{E}(r_{t+1} | s_t = s) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s) \sum_{k=0}^{\infty} \gamma^k \mathbb{E}(r_{t+k+1} | s_t = s, s_{t+1} = s') \\ &= \mathbb{E}(r_{t+1} | s_t = s) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s) \sum_{k=0}^{\infty} \gamma^k \mathbb{E}(r_{t+k+1} | s_{t+1} = s') \quad (\text{Markov property}) \\ &= \mathbb{E}(r_{t+1} + \gamma V(s_{t+1}) | s_t = s) \end{aligned}$$

Value Function : Evaluation

We have

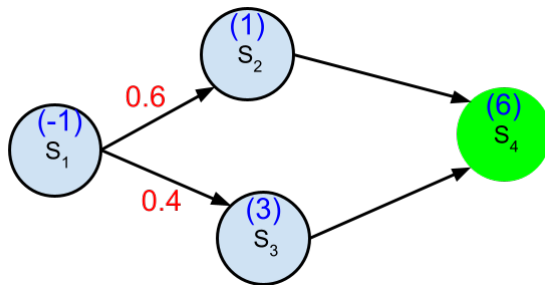
$$V(s) = \mathbb{E}(r_{t+1} + \gamma V(s_{t+1}) | s_t = s)$$



$$V(s) = \mathcal{R}(s) + \gamma \left[\mathcal{P}_{ss'_a} V(s'_a) + \mathcal{P}_{ss'_b} V(s'_b) + \mathcal{P}_{ss'_c} V(s'_c) + \mathcal{P}_{ss'_d} V(s'_d) \right]$$

Value Function Computation : Example

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



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

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

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)

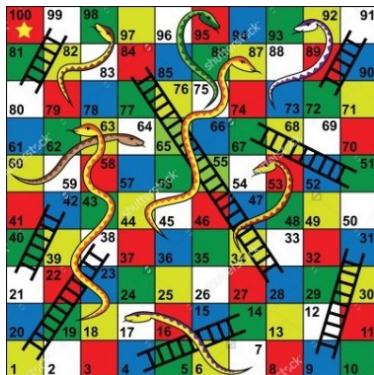
$$V(s) = \mathbb{E}(r_{t+1} | s_t = s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')$$

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

Snakes and Ladders

Question : How can we evaluate the value of (all) states using the value function decomposition ?

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



Let $\mathcal{S} = \{1, 2, \dots, n\}$ and \mathcal{P} be known. Then one can write the Bellman equation as,

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

where

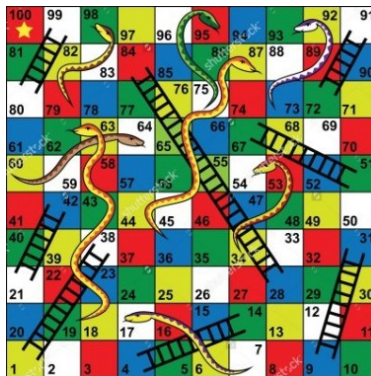
$$\begin{bmatrix} V(1) \\ V(2) \\ \vdots \\ V(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}(1) \\ \mathcal{R}(2) \\ \vdots \\ \mathcal{R}(n) \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \mathcal{P}_{12} & \cdots & \mathcal{P}_{1n} \\ \mathcal{P}_{21} & \mathcal{P}_{22} & \cdots & \mathcal{P}_{2n}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

Solving for V , we get,

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

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

Example : Snakes and Ladders



- ▶ We can now compute the value of states in such 'large' MRP using the matrix form of Bellman equation
- ▶ Value function computed for a particular state provides the **expected** number of plays to reach the goal state s_{100} from that state

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

- ▶ Mathematically convenient to discount rewards
- ▶ Avoids infinite returns in cyclic and infinite horizon setting
- ▶ Discount rate determines the present value of future reward
- ▶ Offers trade-off between being 'myopic' and 'far-sighted' reward
- ▶ In certain class of MDPs, it is **sometimes** possible to use undiscounted reward (i.e. $\gamma = 1$), for example, if all sequences terminate

Markov Decision Process

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

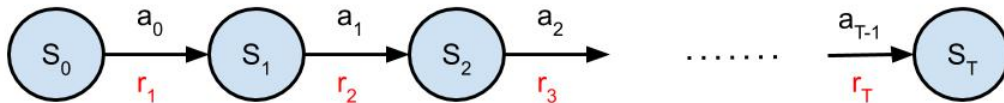
- ▶ \mathcal{S} : (Finite) set of states
- ▶ \mathcal{A} : (Finite) set of actions
- ▶ \mathcal{P} : State transition probability

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

- ▶ \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 $\gamma \in [0, 1]$

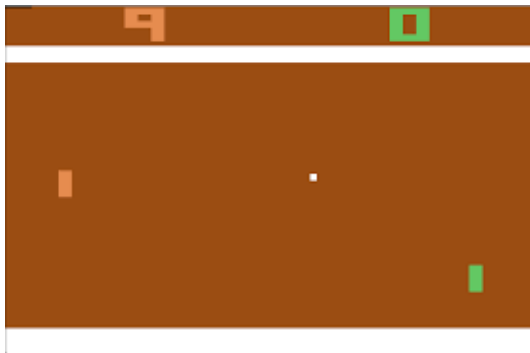


- States \mathcal{S} : Current value of the portfolio and current valuation of instruments in the portfolio
- Actions \mathcal{A} : Buy / Sell instruments of the portfolio
- Reward \mathcal{R} : Return on portfolio compared to previous decision epoch

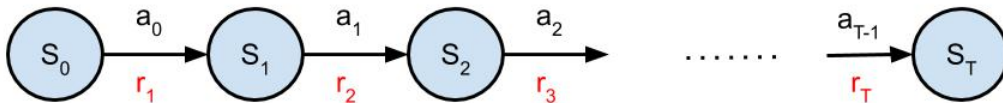
	1	2	3
4	5	6	7
8	9	10	11
12	13	14	

- 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;



- The goal is to choose a sequence of actions such that the expected total discounted future reward $\mathbb{E}(G_t | s_t = s)$ is maximized where

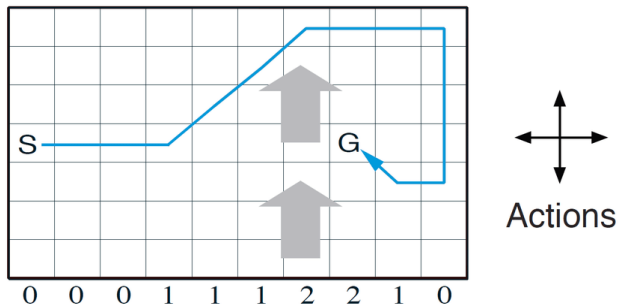
$$G_t = \sum_{k=0}^{\infty} (\gamma^k r_{t+k+1})$$

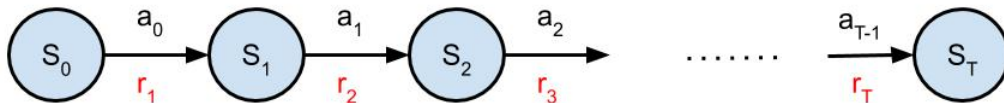
Windy Grid World : Stochastic Environment

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

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

- 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





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

	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 $\gamma = 1$