Markov Decision Process

Easwar Subramanian

TCS Innovation Labs, Hyderabad

Email : cs5500.2020@iith.ac.in

August 12, 2023

Administrivia

- Please consult Prof. Vineeth, for all queries related to registration and other administrative issues
- If need be, register for CS 5500 instead of AI 3000 (relevant for CS, PhD students)
- The Piazza course page is ready; Enrollments are to be done
- Tentative schedule for assignments and exams are in Google sheet

Overview

Review

Types of Learning : Summary

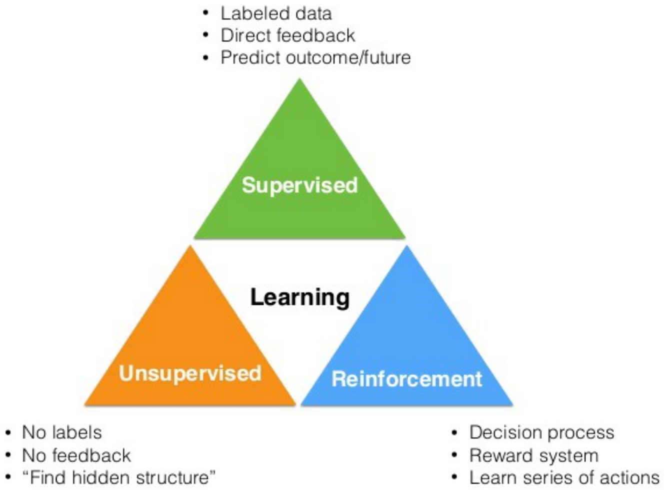
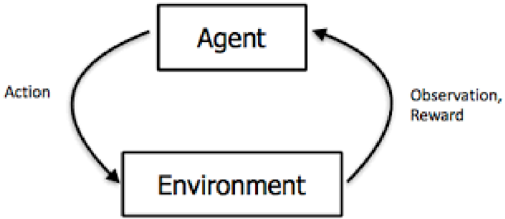


Figure Source: Saggie

Characteristics of Reinforcement Learning



- Observations are non i.i.d and are sequential in nature
- Agent's action (may) affect the subsequent observation seen
- There is no supervisor; Only reward signal (feedback)
- Reward or feedback can be delayed

Reinforcement Learning : History

Learning By Trial and Error (1900's)

Thondrike, Pavlov, Skinner, Klopf, Turing, Minsky and others

Advances in Markov Decision Process and Dynamic Programming (1950's)

Bellman, Howard, Bertsekas and others

↓

Modern Reinforcement Learning (1980's)

Sutton, Barto, Watkins, Tesauro, Bertsekas, Tsitsiklis and others

↓

Deep Reinforcement Learning (2000's)

Andrew Ng, P Abeel, S Levine, J Schulman and others

Course Setup

Markov Decision Process and Dynamic Programming

↓

Modern Reinforcement Learning

↓

Deep Reinforcement Learning

Mathematical Framework for Decision Making

RL Framework : Notations

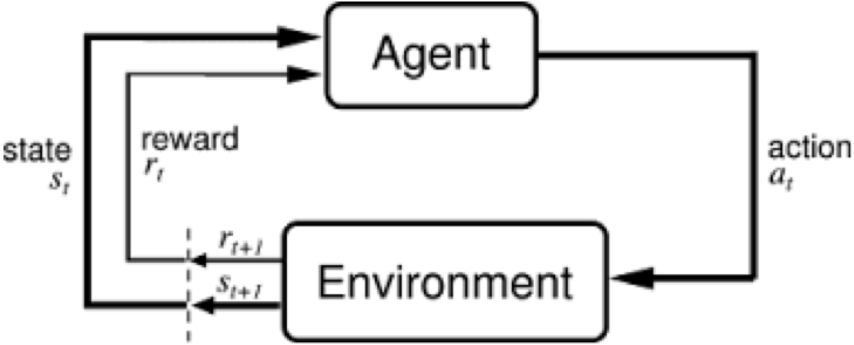


Figure Source: Sutton and Barto

Markov Decision Process

- Markov Decision Process (MDP) provides a mathematical framework for modeling decision making process
- Can formally describe the working of the environment and agent in the RL setting
- Can handle huge variety of interesting settings
  - Multi-arm Bandits - Single state MDPs
  - Optimal Control - Continuous MDPs
- Core problem in solving an MDP is to find an 'optimal' policy (or behaviour) for the decision maker (agent) in order to maximize the total future reward

Markov Chains

Random Variables and Stochastic Process

Random Variable (Non-mathematical definition)

A random variable is a variable whose value depend on the outcome of a random phenomenon

- Outcome of a coin toss
- Outcome of roll of a dice

Stochastic Process

A stochastic or random process, denoted by $\{s_t\}_{t \in T}$, can be defined as a collection of random variables that is indexed by some mathematical set T

- Index set has the interpretation of time
- The set T is, typically, $\mathbb{N}$ or $\mathbb{R}$

Notations

- Typically, in optimal control problems, the index set is continuous (say $\mathbb{R}$)
- Throughout this course (RL), the index set is always discrete (say $\mathbb{N}$)
- Let $\{s_t\}_{t \in T}$ be a stochastic process
- Let s_t be the state at time t of the stochastic process $\{s_t\}_{t \in T}$

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

Transition Probability

State Transition Probability

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

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

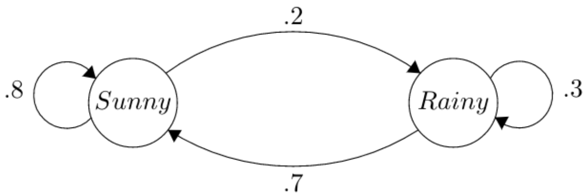
State transition matrix $\mathcal{P}$ then denotes the transition probabilities from all states s to all successor states s' (with each row summing to 1)

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

Markov Chain

A stochastic process $\{s_t\}_{t \in T}$ is a **Markov process** or **Markov Chain** if the sequence of random states satisfy the Markov property. 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

Example 1 : Simple Two State Markov Chain

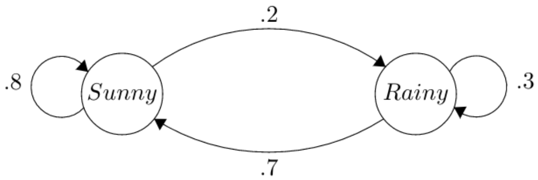


- **State** $\mathcal{S} = \{\text{Sunny}, \text{Rainy}\}$
- **Transition Probability Matrix**

$$\mathcal{P} = \begin{bmatrix} .8 & .2 \\ .7 & .3 \end{bmatrix}$$

Figure Source: https://bookdown.org/probability

Markov Chain : Example Revisited



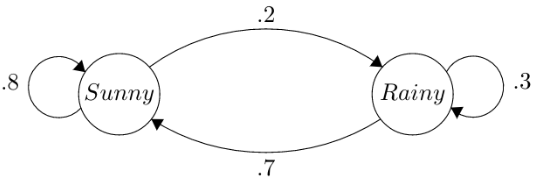
State $\mathcal{S} = \{\text{Sunny}, \text{Rainy}\}$ and **Transition Probability Matrix**

$$\mathcal{P} = \begin{bmatrix} .8 & .2 \\ .7 & .3 \end{bmatrix}$$

- Probability that tomorrow will be *'Rainy'* given today is *'Sunny'* $= 0.2$

Figure Source: https://bookdown.org/probability

Multi-Step Transitions



- Probability that day-after-tomorrow will be *'Rainy'* given today is *'Sunny'* is given by $0.2 * 0.3 + 0.8 * 0.2 = 0.22$

In general, if one step transition matrix is given by,

$$\mathcal{P} = \begin{bmatrix} P_{ss} & P_{sr} \\ P_{rs} & P_{rr} \end{bmatrix}$$

then the two step transition matrix is given by,

$$\mathcal{P}_{(2)} = \begin{bmatrix} P_{ss} * P_{ss} + P_{sr} * P_{rs} & P_{ss} * P_{sr} + P_{sr} * P_{rr} \\ P_{rr} * P_{rs} + P_{rs} * P_{ss} & P_{rr} * P_{rr} + P_{rs} * P_{sr} \end{bmatrix} = P^2$$

Figure Source: https://bookdown.org/probability

Multi-Step Transitions

In general, n-step transition matrix is given by,

$$P_{(n)} = P^n$$

Assumption

We made an important assumption in arriving at the above expression. That the one-step transition matrix stays constant through time or <u>independent</u> of time

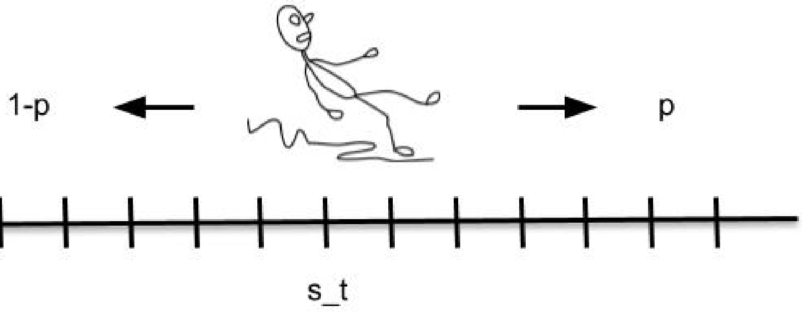
- Markov chains generated using such transition matrices are called <u>homogeneous</u> Markov chains
- For much of this course, we will consider homogeneous Markov chains, for which the transition probabilities depend on the length of time interval $[t_1, t_2]$ but not on the exact time instants

Markov Chains : Examples

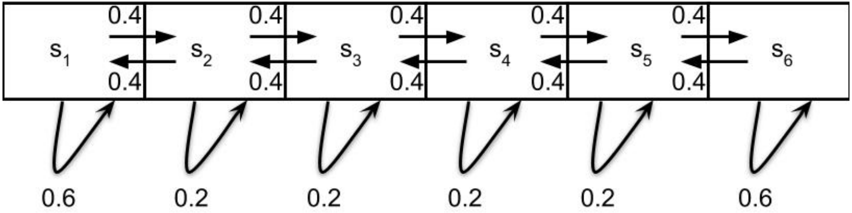
Example 2 : One dimensional random walk

A walker flips a coin every time slot to decide which 'way' to go.

$$s_{t+1} = \begin{cases} s_t + 1 & \text{with probability } p \\ s_t - 1 & \text{with probability } 1 - p \end{cases}$$



Example 3 : Simple Grid World



- $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5, s_6, s_6\}$
- $\mathcal{P}$ as shown above
- Example Markov Chains with s_2 as start state
  - ★ $\{s_2, s_3, s_2, s_1, s_2, \cdots\}$
  - ★ $\{s_2, s_2, s_3, s_4, s_3, \cdots\}$

Example 4 : Dice roll experiment

Let $\{s_t\}_{t \in T}$ model the stochastic process representing the cumulative sum of a fair six-sided die rolls

Example 5 : Natural Language Processing

Let $\{s_t\}_{t \in T}$ model the stochastic process that keeps track of the chain of letters in a sentence. Consider an example

`Tomorrow is a sunny day`

- We normally don't ask the question what is probability of character '`a`' appearing given previous character is '`d`'
- Sentence formation is typically **non-Markovian**

Notion of Absorbing State

Absorbing State

A state $s \in \mathcal{S}$ is called **absorbing** state if it is impossible to leave the state. That is,

$$P_{ss'} = \left\{ \begin{array}{ll} 1, & \text{if } s = s' \\ 0, & \text{otherwise} \end{array} \right\}$$

Markov Reward Process

Markov Reward Process

Markov Reward Process

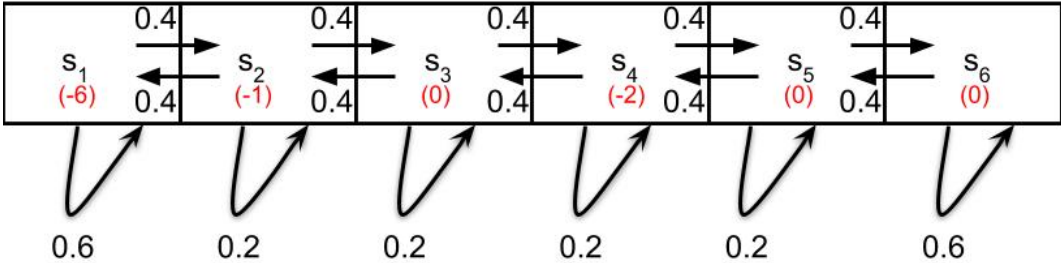
A Markov reward process is a tuple $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ is a Markov chain with values

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

Simple Grid World : Revisited



- For the Markov chain $\{s_2, s_3, s_2, s_1, s_2, \cdots\}$ the corresponding reward sequence is $\{-1, 0, -1, -6, -1, \cdots\}$

No notion of action

Slide Credit: Emma Brunskill
CS234 Stanford

Example : Snakes and Ladders

Example : Snakes and Ladders

- **States** $\mathcal{S}$: $\{s_1, s_2, \cdots, s_{100}\}$
- **Transition Probability** $\mathcal{P}$:
  - ★ What is the probability to move from state 2 to 6 in one step ?
  - ★ What are the states that can be visited in one-step from state 2 ?
  - ★ What is the probability to move from state 2 to 4 ?
  - ★ Can we transition from state 15 to 7 in one step ?

  Question : Is transition matrix independent of time ?

Question : Can we formulate the game of Snake and Ladders as a MRP ?

Need to define suitable reward function and discounting factor

On Rewards : Total Return

- At each time step t, there is a reward r_{t+1} associated with being in state s_t
- Ideally, we would like the agent to pick such trajectories in which the cumulative reward accumulated by traversing such a path is high

Question : How can we formalize this ?

Answer : If the reward sequence is given by $\{r_{t+1}, r_{t+2}, r_{t+3}, \cdots\}$, then, we want to maximize the sum

$$r_{t+1} + r_{t+2} + r_{t+3} + \cdots$$

Define G_t to be

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

$$\boxed{\text{The goal of the agent is to pick such paths that maximize } G_t}$$

Total (Discounted) Return

Recall that,

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

- In the case that the underlying stochastic process has infinite terms the above summation could be divergent

Therefore, we introduce discount factor $\gamma \in [0, 1]$ and redefine G_t as

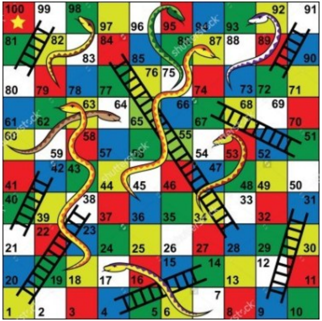
$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

- G_t is the total discounted return starting from time t
- If $\gamma < 1$ then the infinite sum has a finite value if the reward sequence is bounded
- γ close to 0 the agent is concerned only with immediate reward(s) (myopic)
- γ close to 1 the agent considers future reward more strongly (far-sighted)

Few Remarks on Discounting

- Mathematically convienient 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 finite MDPs, it is sometimes possible to use undiscounted reward (i.e. $\gamma = 1$), for example, if all sequences terminate

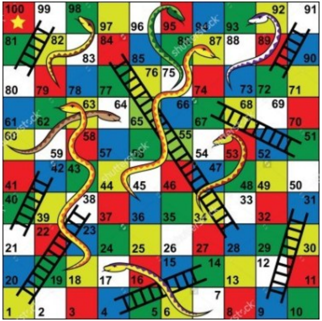
Snakes and Ladders : Revisited



Question : What can be a suitable reward function and discount factor to describe 'Snake and Ladders' as a Markov reward process ?

- **Goal** : From any given state reach s_{100} in as few steps as possible
- **Reward** $\mathcal{R}$: $\mathcal{R}(s) = -1$ for $s \in s_1, \cdots, s_{99}$ and for $R(s_{100}) = 0$
- **Discount Factor** $\gamma = 1$

Snakes and Ladders : Revisited



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

Value Function

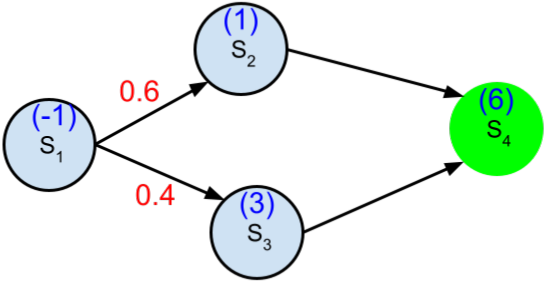
The value function $V(s)$ gives the long-term value of state $s \in \mathcal{S}$

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

- 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
- $V(s)$ is independent of t

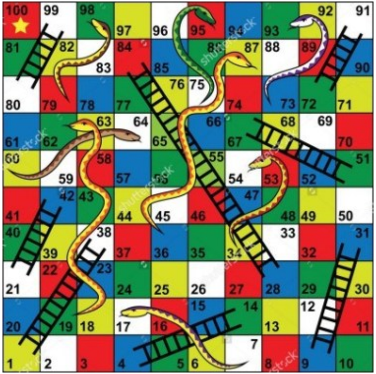
Value Function Computation : Example

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 ' ?

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

Recall that,

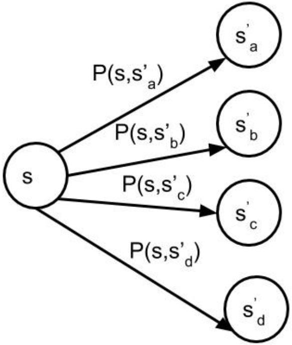
$$G_t = \left(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots\right) = \sum_{k=0}^{\infty} \left(\gamma^k r_{t+k+1}\right)$$

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

Value Function : Evaluation

We have

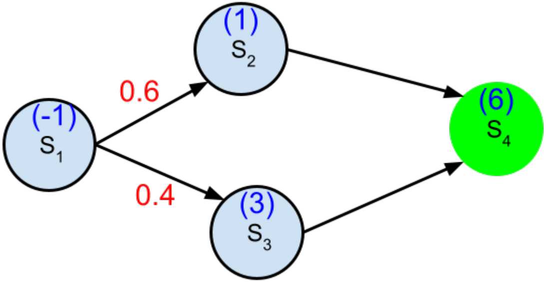
$$V(s) = \mathbb{E}(r_{t+1} + \gamma V(s_{t+1}) | s_t = s)$$



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

Value Function Computation : Example

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



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

Bellman Equation for Markov Reward Process

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

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)

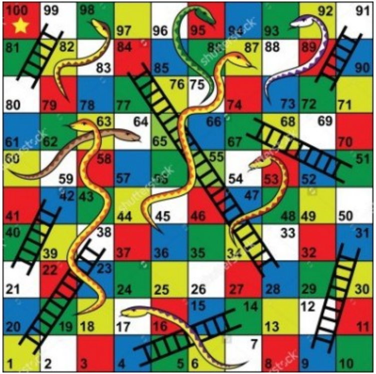
$$\boxed{V(s) = \mathbb{E}(r_{t+1} | s_t = s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')}$$

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

Snakes and Ladders

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

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



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,

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

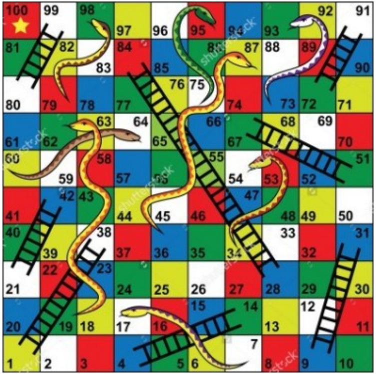
where

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

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



- 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

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

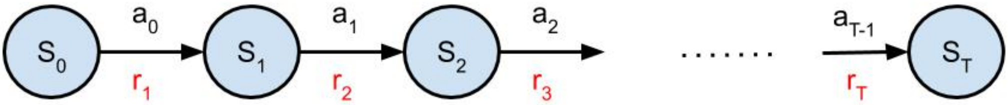
$$\mathcal{P}^{a}_{ss'} = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a), a_t \in \mathcal{A}$$

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

Wealth Management Problem



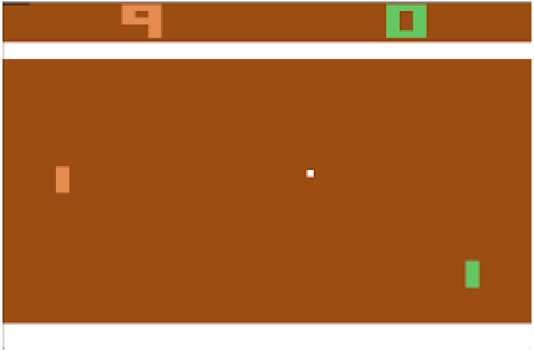
- States $\mathcal{S}$: Current value of the portfolio and current valuation of instruments in the portfolio
- Actions $\mathcal{A}$: Buy / Sell instruments of the portfolio
- Reward $\mathcal{R}$: Return on portfolio compared to previous decision epoch

Navigation Problem

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

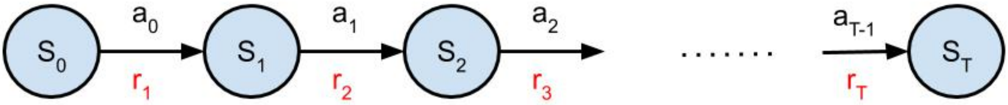
- States $\mathcal{S}$: Squares of the grid
- Actions $\mathcal{A}$: Any of the four directions possible
- Reward $\mathcal{R}$: -1 for every move made until reaching goal state

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;

Flow Diagram



- 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

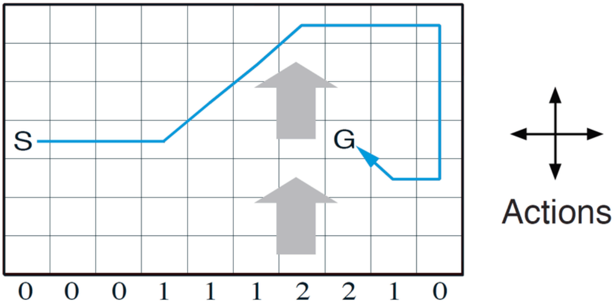
$$G_t = \sum_{k=0}^{\infty} \left(\gamma^k r_{t+k+1} \right)$$

Windy Grid World : Stochastic Environment

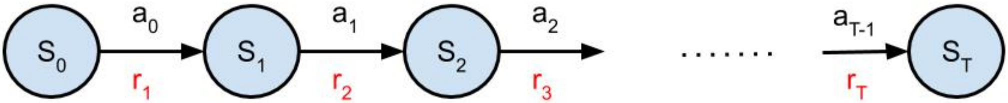
Recall given an MDP $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, we have the state transition probability $\mathcal{P}$ defined as

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

- 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

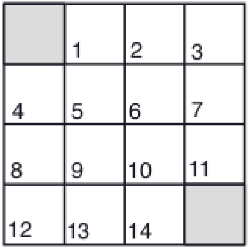


Finite and Infinite Horizon MDPs



- 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

Grid World Example



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 $\gamma = 1$