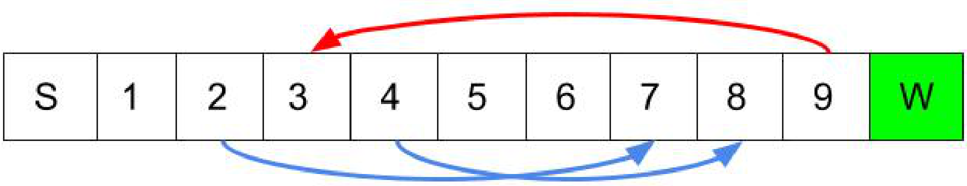
AI 3000 / CS 5500 : Reinforcement Learning
Assignment № 1

Due Date : 01/09/2022

Course Instructor : Easwar Subramanian 21/08/2022

Problem 1 : Markov Reward Process

Consider the following snake and ladders game as depicted in the figure below.



- Initial state is S and a fair four sided die is used to decide the next state at each time
- Player must land exactly on state W to win
- Die throws that take you further than state W leave the state unchanged

(a) Identify the states, transition matrix of this Markov process. (1 point)

(b) Construct a suitable reward function, discount factor and use the Bellman equation for the Markov reward process to compute how long does it take "on average" (the expected number of die throws) to reach the state W from any other state. (4 points)

(a) The states of the Markov process is given by, $\mathcal{S} = \{S, 1, 3, 5, 6, 7, 8, W\}$. Positions 2 and 4 of the grid are same as positions 7 and 8 respectively.

The transition matrix is given by,

$$P = \begin{pmatrix} 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0.25 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0 & 0.25 & 0.25 & 0.25 & 0.25 & 0 \\ 0 & 0.25 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0 & 0.25 & 0 & 0 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(b) The state W is an absorbing state since if the Markov process reach the state W there are no further state transitions possible apart from staying at state W.

(c) Following are the suitable reward functions and discount factor γ.

- The suitable discount factor is $\gamma = 1$ as we are estimating "average" number of steps to reach W.
- The reward for any state could be $R(s) = -1$ for $s \neq W$ and $R(s) = 0$ for $s = W$. Then $V(s) = 0$ for $s = W$.
- The Bellman evaluation equation for an MRP given by $V = (I - \gamma P)^{-1} R$ which when solved for $V(s)$ would give the "average number" of die throws required to reach state W from state s. The matrix $(I - \gamma P)$ becomes invertible if we set $V(s) = 0$ for $s = W$. One may find the inverse of the matrix $(I - \gamma P_{7\times 7})$ to compute the average die throws from other seven states. Upon solving, the vector $V(s)$ is given by,

$$V(s) = \{7.0833, 7, 6.6667, 6.6667, 5.3333, 5.3333, 5.3333\}$$

Problem 2 : Markov Decision Process

A production facility has N machines. If a machine starts up correctly in the morning, it renders a daily revenue of $1\$$. A machine that does not start up correctly, needs to be repaired. A visit by a repair man costs $\frac{N}{2}\$$ per day and he repairs all broken machines on the same day. The repair cost is a lump-sum amount and does not depend on the number of machines that is repaired. A machine that has been repaired always starts up correctly the next day. The number of machines that start up correctly the next day depends on the number of properly working machines at present day and is governed by the probability distribution given in the table below, where m stands for the number of (presently) working machines and n stands for the number of ones that would start up correctly the next day. The goal for the facility manager is to maximize the profits (revenue - costs) earned.

m	$n=0$	$n=1$	$n=2$	$n=3$	$\cdots$	$n=N-1$	$n=N$
$m=1$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0
$m=2$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0
$m=3$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m=N-1$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	0
$m=N$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$

(a) Formulate the above problem as a Markov decision process by ennumerating the state space, action space, rewards and transition probabilities. (3 Points)

- State space : Number of working machines on any given day $\mathcal{S} = \{0, 1, \cdots, N\}$
- Action space : $\{\text{repair}, \text{no-repair}\}$

- Reward : $\mathcal{R}(s, \text{no-repair}, s') = s \quad \mathcal{R}(s, \text{repair}, s') = s - \frac{N}{2}$;
- Transition probablities, for action **no-repair** action

m	$n=0$	$n=1$	$n=2$	$n=3$	$\cdots$	$n=N-1$	$n=N$
$m=0$	1	0	0	0	0	0	0
$m=1$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0	0
$m=2$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0	0
$m=3$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m=N-1$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	0
$m=N$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$

Transition probablities, for action **repair** action

m	$n=0$	$n=1$	$n=2$	$n=3$	$\cdots$	$n=N-1$	$n=N$
$m=0$	0	0	0	0	0	0	1
$m=1$	0	0	0	0	0	$\frac{1}{2}$	$\frac{1}{2}$
$m=2$	0	0	0	0	0	$\frac{1}{3}$	$\frac{1}{3}$
$m=3$	0	0	0	0	0	$\frac{1}{4}$	$\frac{1}{4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$m=N-1$	0	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$	$\frac{1}{N}$
$m=N$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$	$\frac{1}{N+1}$

(b) Would you use discounted or undiscounted setting for the above MDP formulation ? Justify the answer. (1 Point)

It is preferred to use the discounted setting as the problem as formulated is an infinite horizon case.

(c) Suppose the facility manager adopts the policy to never call the repair man. Calculate the value of the policy. For this sub-problem assume that the number of machines in the facility to be five. (3 Points)

For this sub-problem and the next, let us assume $\gamma = 0.9$. One can use the Bellman evaluation equation in the following form to the value of the policy (π) 'no-repair'. We let $|\mathcal{S}|$ to be five.

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

With the identity matrix of size 6 $\times 6$, we have the value for the 6 states as

$$V^{\pi} = \{0, 1.818, 3.63, 5.45, 7.29, 9.09\}$$

(d) Perform one iteration the of policy iteration algorithm on the no-repair policy adopted by the facility manager to get an improved policy for the five machine scenario. (3 Points)

Improvement is achieved by being greedy w.r.t to the no-repair policy π at every state s At state $s = 0$, we have,

$$\pi_1(0) = \arg\max_a(0 + \gamma V^\pi(0), -2.5 + 0.9 * V^\pi(5)) = \arg\max(0, -2.5 + 0.9 * 9.09) = \text{repair}$$

$$\pi_1(1) = \arg\max_a(1.818, -5/2 + 0/9 * (0.5 * V^\pi(5) + 0.5 * V^\pi(4))) = \text{repair}$$

$$\pi_1(2) = \arg\max_a(3.63, -5/2 + 0/9 * (0.33 * V^\pi(5) + 0.33 * V^\pi(4) + 0.33 * V^\pi(3))) = \text{repair}$$

$$\pi_1(3) = \arg\max_a(5.45, -5/2+0/9*(0.25*V^\pi(5)+0.25*V^\pi(4)+0.25*V^\pi(3)+0.25*V^\pi(2))) = \text{no-repair}$$

Similarly, it si possible to work out for state $s = 4$ and $s = 5$ a better action is **no-repair**

Problem 3 : On Ordering of Policies

Consider the MDP shown in Figure 1. The MDP has 4 states $\mathcal{S} = \{A, B, C, D\}$ and there are two actions a_1 and a_2 possible. The actions determine which direction to move from a given state. We consider a stochastic environment such that action suggested by the policy succeeds 90 % of the times and fails 10 % of the times. Upon failure, the agent moves in the direction suggested by the other action. The state D is a terminal state with reward of 100. One can think that terminal states have only one action (an exit option) which gives the terminal reward 100. We consider three policies to this MDP.

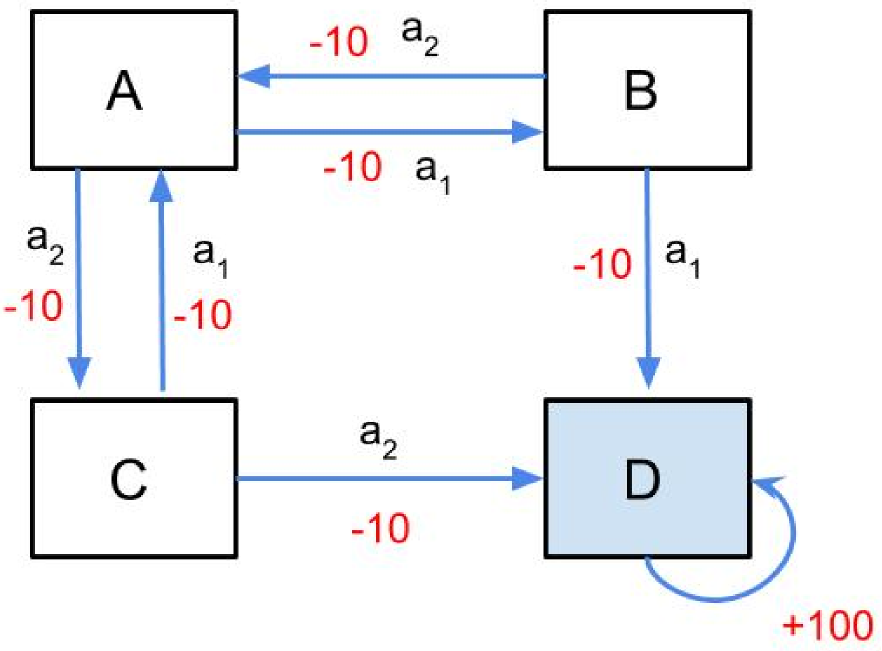


Figure 1: Partial Ordering of Policies

- Policy π_1 is deterministic policy that chooses action a_1 at all states $s \in \mathcal{S}$.

- Policy π_2 is another deterministic policy that chooses action a_2 at all states $s \in \mathcal{S}$.
- Policy π_3 is a stochastic policy described as follows
  - Action a_1 is chosen in states B and D with probability 1.0
  - Action a_2 is chosen in state C with probability 1.0
  - Action a_1 is chosen in state A with probability 0.4 and action a_2 is chosen with probability 0.6

(a) Evaluate $V^{\pi}(s)$ for each policy described above using the Bellman evaluation equation for all states $s \in \mathcal{S}$. (3 Points)

For states $\{A, B, C, D\}$, we have

(a) $V^{\pi_1} = \{75.61, 87.56, 68.05, 100\};$

(b) $V^{\pi_2} = \{75.61, 68.05, 87.56, 100\};$

(c) $V^{\pi_3} = \{77.78, 87.78, 87.78, 100\};$

(b) Which is the best policy among the suggested policies ? Why ? (1 Point)

Policy π_3 is the best policy as its value at each state is greater than or equal to the values obtained by policy π.

(c) Are all policies comparable ? Provide reason for your answer. (1 Point)

Policies π_1 and π_2 are not comparable since for state B, $V^{\pi_1}(B) > V^{\pi_2}(B)$ whereas for state C we have $V^{\pi_2}(C) > V^{\pi_1}(C)$

(d) Let π_1 and π_2 be two deterministic stationary policies of an MDP M . Construct a new policy π that is better than policies π_1 and π_2. Explain the answer. (3 Points)

[**Note** : M in sub-question (d) is any arbitrary MDP]

Here we are given two policies π_1 and π_2 for MDP M. At every state $s \in \mathcal{S}$ of the MDP, construct a new policy π as follows :

$$\pi(s) = \begin{cases} \pi_1(s) & \text{if } V^{\pi_1}(s) \geq V^{\pi_2}(s) \\ \pi_2(s) & \text{Otherwise} \end{cases}$$

This construction will lead to the conclusion that for every state $s \in \mathcal{S}$ of the MDP M, we have $V^{\pi}(s) \geq \max(V^{\pi_1}(s), V^{\pi_2}(s))$, which makes the policy π better than policy π_1 and π_2.

Problem 4 : Effect of Noise and Discounting

Consider the grid world problem shown in Figure 2. The grid has two terminal states with positive payoff (+1 and +10). The bottom row is a cliff where each state is a terminal state with negative payoff (-10). The greyed squares in the grid are walls. The agent starts from the yellow state

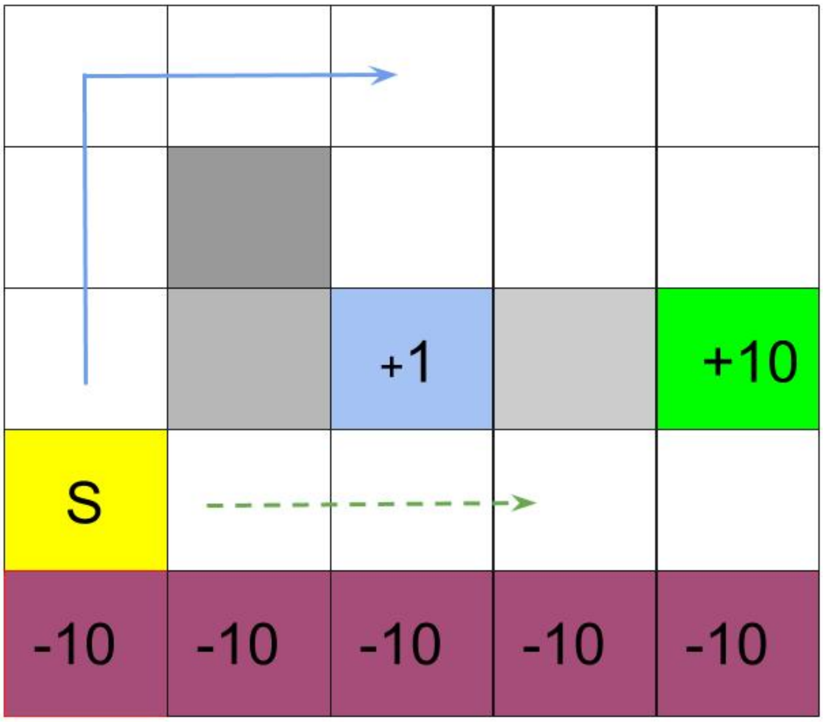


Figure 2: Modified Grid World

S. As usual, the agent has four actions $\mathcal{A} =$ (Left, Right, Up, Down) to choose from any non-terminal state and the actions that take the agent off the grid leaves the state unchanged. Notice that, if agent follows the dashed path, it needs to be careful not to step into any terminal state at the bottom row that has negative payoff. There are four possible (optimal) paths that an agent can take.

- Prefer the close exit (state with reward +1) but risk the cliff (dashed path to +1)
- Prefer the distant exit (state with reward +10) but risk the cliff (dashed path to +10)
- Prefer the close exit (state with reward +1) by avoiding the cliff (solid path to +1)
- Prefer the distant exit (state with reward +10) by avoiding the cliff (solid path to +10)

There are two free parameters to this problem. One is the discount factor γ and the other is the noise factor (η) in the environment. Noise makes the environment stochastic. For example, a noise of 0.2 would mean the action of the agent is successful only 80 % of the times. The rest 20 % of the time, the agent may end up in an unintended state after having chosen an action.

(a) Identify what values of γ and η lead to each of the optimal paths listed above with reasoning. (8 Points)

[**Hint** : For the discount factor, consider high and low γ values like 0.9 and 0.1 respectively. For noise, consider deterministic and stochastic environment with noise level η being 0 or 0.5 respectively]

Answer

1. When γ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is zero in the environment, there is no danger of tripping to the cliff. Therefore, for low γ and low η, the agent would prefer the close exit and risk the cliff.

2. When γ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is high or moderate in the environment, there is danger of tripping to the cliff. Therefore, for low γ and low η, the agent would prefer the close exit and not risk the cliff.

3. When γ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is low or zero in the environment, there is less or no danger of tripping to the cliff. Therefore, for high γ and low η, the agent would prefer the distant exit and risk the cliff.

4. When γ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is high or medium in the environment, there is danger of tripping to the cliff. Therefore, for high γ and high η, the agent would prefer the distant exit and not risk the cliff.

Problem 5 : Value Functions

Let M_1 and M_2 be two identical MDPs with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$ except for reward formulation. That is, $M_1 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_1, \gamma >$ and $M_2 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_2, \gamma >$. Let M_3 be another MDP such that $M_3 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_1 + \mathcal{R}_2, \gamma >$. Assume $\gamma < 1$.

(a) For an arbitrary but fixed policy π, suppose we are given action value functions $Q_1^\pi(s,a)$ and $Q_2^\pi(s,a)$, corresponding to MDPs M_1 and M_2, respectively. Explain whether it is possible to combine these action value functions in a simple manner to calculate $Q_3^\pi(s)$ corresponding to MDP M_3. (2 Points)

Yes, it is possible to combine the two action value functions of the MDP into a single action value function for the composite MDP since the combination involve only expectation operator and it is linear in nature.

(b) Suppose we are given optimal polices π_1^* and π_2^* corresponding to MDPs M_1 and M_2, respectively. Explain whether it is possible to combine these optimal policies in a simple manner to formulate an optimal policy π_3^* corresponding to MDP M_3. (2 Points)

Combining optimal policies is not straightforward as it involves taking care of the max operator which is a non-linear operator. Hence, optimal policies of the two MDPs cannot be combined in a straightforward fashion

(c) Suppose π^* is an optimal policy for both MDPs M_1 and M_2. Will π^* also be an optimal policy for MDP M_3 ? Justify the answer. (2 Points)

Given that π^* is an optimal policy for both M_1 and M_2 we have,

$$\begin{aligned} V^*_{M_1} &= \max_a \left[R_1 + \gamma \sum_{s'} \left(P(s'|s,a) V^*_{M_1}(s') \right) \right] \\ V^*_{M_2} &= \max_a \left[R_2 + \gamma \sum_{s'} \left(P(s'|s,a) V^*_{M_2}(s') \right) \right] \end{aligned}$$

Adding the optimal value function for MDPs M_1 and M_2, with the knowledge that π^* is an optimal policy for both MDPs, is given by,

$$V^* = \max_a \left[(R_1 + R_2) + \gamma \sum_{s'} \left(P(s'|s,a) \left(V^*_{M_1}(s)' + V^*_{M_2}(s') \right) \right) \right]$$

Optimal value function for MDP M_3 is given by

$$W^*_{M_3} = \max_a \left[(R_1 + R_2) + \gamma \sum_{s'} \left(P(s'|s,a) W^*_{M_3}(s') \right) \right]$$

Equating the last two equations, one can recognize that

$$W^*_{M_3} = V^*_{M_1} + V^*_{M_2}$$

Sum of two optimal value functions following the same optimal policy results in optimal value function for M_3.

(d) Let ε be a fixed constant. Assume that the reward functions $\mathcal{R}_1$ and $\mathcal{R}_2$ are related as

$$\mathcal{R}_1(s,a,s') - \mathcal{R}_2(s,a,s') = \varepsilon$$

for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Let π be an arbitrary policy and let $V_1^\pi(s)$ and $V_2^\pi(s)$ be the corresponding value functions of policy π for MDPs M_1 and M_2, respectively. Derive an expression that relates $V_1^\pi(s)$ to $V_2^\pi(s)$ for all $s \in \mathcal{S}$. (3 Points)

Considering the definition of $V^\pi(s)$, the state value function under policy π, we have

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

We assume that each reward has a constant added to it. That is we consider the reward $\hat{r}_{t+k+1}$ in terms of r_{t+k+1} by

$$\hat{r}_{t+k+1} = r_{t+k+1} + \varepsilon$$

Then, the state value function for this new sequence of rewards is given by,

$$\begin{aligned}
\hat{V}^{\pi}(s) &= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^{k} \hat{r}^{t+k+1}\right) \\
&= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^{k}\left(r_{t+k+1}+\varepsilon\right)\right) \\
&= \mathbb{E}_{\pi}\left(\sum_{k=0}^{\infty} \gamma^{k} r_{t+k+1}\right)+\mathbb{E}_{\pi}\left(\gamma^{k} \epsilon\right) \\
&= V^{\pi}(s)+\mathbb{E}_{\pi}\left(\gamma^{k} \varepsilon\right)=V^{\pi}(s)+\varepsilon \sum_{k=0}^{\infty} \gamma \\
&= V^{\pi}(s)+\frac{\varepsilon}{1-\gamma}
\end{aligned}$$

The alternate relation

$$\hat{V}^{\pi}=V^{\pi}(I-\gamma P)^{-1} \varepsilon$$

is dependent on the model of the MDP.

ALL THE BEST