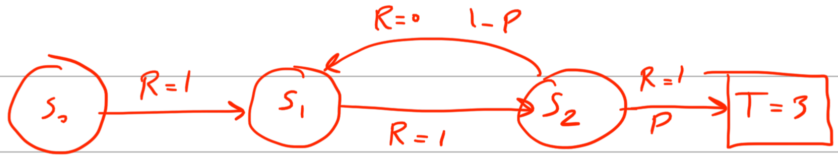
حل سوال بلمن:



بخش الف:

$V_\pi(S_3) = 0 \qquad V_\pi(S_2) = P(1+\gamma V_\pi(S_3)) + (1-P)(0+\gamma V_\pi(S_1))$

$= P + (1-P)V_\pi(S_1) \quad (I)$

$V_\pi(S_1) = 1 + \gamma V_\pi(S_2) = 1 + V_\pi(S_2) \quad (II)$

$(I),(II) \Rightarrow V_\pi(S_2) = P + (1-P)(1+V_\pi(S_2)) = P + 1 - P + V_\pi(S_2) - P\,V_\pi(S_2)$

$= 1 + (1-P)V_\pi(S_2)$

$\Rightarrow V_\pi(S_2) = 1/P \Rightarrow V_\pi(S_1) = 1 + 1/P$

$V_\pi(S_0) = 1 + \gamma V_\pi(S_1) = 1 + 1 + 1/P = 2 + 1/P$

بخش ب) n متغیر تصادفی است. داریم:

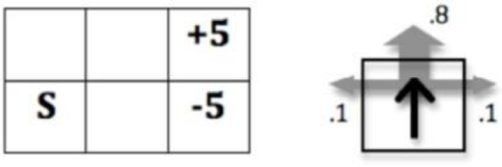
$P(n = 3+2k) = P^k(1-P)^k$

$$E(n) = \sum_{k=0}^{\infty}(3+2k)[(P)(1-P)]^k = \frac{3-(P)(1-P)}{(P(1-P)-1)^2}$$

سوال پیشنهادی دوم

Question 5 – MDPs and Reinforcement Learning – 28 points

This gridworld MDP operates like to the one we saw in class. The states are grid squares, identified by their row and column number (row first). The agent always starts in state (1,1), marked with the letter S. There are two terminal goal states, (2,3) with reward +5 and (1,3) with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability .8. With probability .1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.



(a) (b)

Figure 1: (a) Gridworld MDP. (b) Transition function.

1. Draw the optimal policy for this grid? (5 points)

S =	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
$\pi^*(S)$ =	Up	Left	NA	Right	Right	NA

2. Suppose the agent knows the transition probabilities. Give the first two rounds of value iteration updates for each state, with a discount of 0.9. (Assume V_0 is 0 everywhere and compute V_i for times $i = 1, 2$). (8 points)

Apply the Bellman backups $V_{i+1}(s) = \max_a(\sum_{s'} T(s, a, s')(R(s, a, s') + \gamma V_i(s')))$ twice. I will show the computations for the max actions. Most of the terms will be zero, which are omitted here for compactness.

S =	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
$V_0(S)$ =	0	0	0	0	0	0
$V_1(S)$ =	0	0	0	0	$0.8 \times 5.0 = 4.0$	0
$V_2(S)$ =	0	$0.9 \times 0.8 \times 4$ $+0.1 \times -5 = 2.38$	0	$0.8 \times 0.9 \times 4.0 = 2.88$	$0.8 \times 5.0 = 4.0$	0