CSCI 4150: Introduction to Artificial Intelligence
MDP, RL
Total points: 60.

We only accept electronic submission at Submitty. Please try to ask questions on Piazza. If Piazza is not helpful, please contact the TAs.

Problem 1 (30pt). Consider the MDP with the transition model, reward function, and V^1 as given in the Tables 1, 2, and 3. The set of states is $\{A, B\}$, and the set of actions is $\{1, 2, 3\}$. Assume the discount factor $\gamma = 1$, i.e., no discounting. Do two-step value iteration taught in the class by answering the questions below.

Table 1: Starting from A

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
A	1	A	0	0
A	1	B	1	0
A	2	A	1	2
A	2	B	0	0
A	3	A	0.5	0
A	3	B	0.5	0

Table 2: Starting from B

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
B	1	A	0.5	12
B	1	B	0.5	0
B	2	A	1	0
B	2	B	0	0
B	3	A	0.5	2
B	3	B	0.5	4

Table 3: V^1

V^1(A)	0
V^1(B)	0

1. Fill in the values for Q^1, Q^2 in the table below.

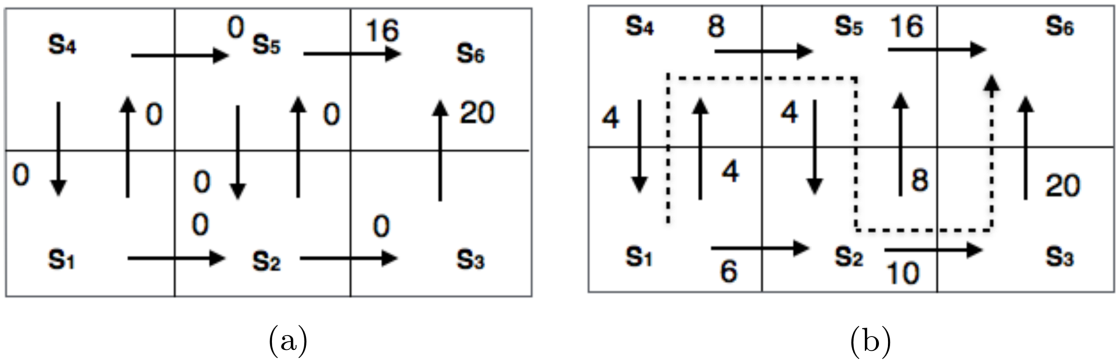
Q^1(A,1)	Q^1(A,2)	Q^1(A,3)	Q^1(B,1)	Q^1(B,2)	Q^1(B,3)
Q^2(A,1)	Q^2(A,2)	Q^2(A,3)	Q^2(B,1)	Q^2(B,2)	Q^2(B,3)

2. Let $\pi^{i+1}(s)$ be the optimal action in state s after i-th iteration of the algorithm (i.e., after you computed Q^i). What are $\pi^2(A)$, $\pi^2(B)$, $\pi^3(A)$, and $\pi^3(B)$? Show your calculations.

Problem 2 (30pt). Consider the deterministic world in Figure 1. Allowable actions (including $\uparrow, \rightarrow, \downarrow$, **not including** $\leftarrow$) are shown by arrows, and the numbers on the arrows indicate the reward for performing each action. For example, the agent receives 0 when moving from S_5 to S_2 by taking $\downarrow$; the agent receives 16 when moving from S_5 to S_6 by taking $\rightarrow$.

Suppose a Q-learning algorithm has run for a while, and the current Q values are shown in Fig. 1 (b). For example, $Q(S_1, \rightarrow) = 6$, $Q(S_2, \uparrow) = 8$. S_6 is a terminal state with exist reward 0, i.e., $Q(S_6, a) = 0$ for all actions a. Show the updates to the Q values after each step, when the agent takes the path shown by the dotted line $(S_1, S_4, S_5, S_2, S_3, S_6)$ with the discount factor $\gamma = 0.3$ and the learning rate $\alpha = 0.8$. Show all your calculations.

Hint: There are five steps, and after each step only one Q value is updated. The reward should be read from Fig. 1 (a). For example, the first step is $(S_1, \uparrow, S_4)$ with reward 0. No state abstraction is needed in this problem.



(a)

(b)

Figure 1: Problem 2