Midterm Review RL

Q1. Markov Decision Processes

Consider a simple MDP with two states, S_1 and S_2, two actions, A and B, a discount factor γ of $1/2$, reward function R given by

$$R(s,a,s') = \begin{cases} 1 & \text{if } s' = S_1; \\ -1 & \text{if } s' = S_2; \end{cases}$$

and a transition function specified by the following table.

s	a	s'	$T(s,a,s')$
S_1	A	S_1	$1/2$
S_1	A	S_2	$1/2$
S_1	B	S_1	$2/3$
S_1	B	S_2	$1/3$
S_2	A	S_1	$1/2$
S_2	A	S_2	$1/2$
S_2	B	S_1	$1/3$
S_2	B	S_2	$2/3$

(a) Perform a single iteration of value iteration, filling in the resultant Q-values and state values in the following tables. Use the specified initial value function V_0, rather than starting from all zero state values. Only compute the entries not labeled "skip".

s	a	$Q_1(s,a)$
S_1	A	
S_1	B	
S_2	A	skip
S_2	B	skip

s	$V_0(s)$	$V_1(s)$
S_1	2	
S_2	3	skip

(b) Suppose that Q-learning with a learning rate α of $1/2$ is being run, and the following episode is observed.

s_1	a_1	r_1	s_2	a_2	r_2	s_3
S_1	A	1	S_1	A	-1	S_2

Using the initial Q-values Q_0, fill in the following table to indicate the resultant progression of Q-values.

s	a	$Q_0(s,a)$	$Q_1(s,a)$	$Q_2(s,a)$
S_1	A	$-1/2$		
S_1	B	0		
S_2	A	-1		
S_2	B	1		

(c) Given an arbitrary MDP with state set S, transition function $T(s, a, s')$, discount factor γ, and reward function $R(s, a, s')$, and given a constant $\beta > 0$, consider a modified MDP (S, T, γ, R') with reward function $R'(s, a, s') = \beta \cdot R(s, a, s')$. Prove that the modified MDP (S, T, γ, R') has the same set of optimal policies as the original MDP (S, T, γ, R).

(d) Although in this class we have defined MDPs as having a reward function $R(s, a, s')$ that can depend on the initial state s and the action a in addition to the destination state s', MDPs are sometimes defined as having a reward function $R(s')$ that depends only on the destination state s'. Given an arbitrary MDP with state set S, transition function $T(s, a, s')$, discount factor γ, and reward function $R(s, a, s')$ that *does depend* on the initial state s and the action a, define an *equivalent* MDP with state set S', transition function $T'(s, a, s')$, discount factor γ', and reward function $R'(s')$ that depends only on the destination state s'.

By *equivalent*, it is meant that there should be a one-to-one mapping between state-action sequences in the original MDP and state-action sequences in the modified MDP (with the same value). **You do not need to give a proof of the equivalence.**

States: $S' =$

Transition function: $T'(s, a, s') =$

Discount factor: $\gamma' =$

Reward function: $R'(s') =$

Q2. Q-learning

Consider the following gridworld (rewards shown on left, state names shown on right).

Rewards

+10	+1

State names

A	B
G1	G2

From state A, the possible actions are right($\rightarrow$) and down($\downarrow$). From state B, the possible actions are left($\leftarrow$) and down($\downarrow$). For a numbered state (G1, G2), the only action is to exit. Upon exiting from a numbered square we collect the reward specified by the number on the square and enter the end-of-game absorbing state X. We also know that the discount factor $\gamma = 1$, and in this MDP all actions are **deterministic** and always succeed.

Consider the following episodes:

Episode 1 ($E1$)

s	a	s'	r
A	$\downarrow$	$G1$	0
$G1$	exit	X	10

Episode 2 ($E2$)

s	a	s'	r
B	$\downarrow$	$G2$	0
$G2$	exit	X	1

Episode 3 ($E3$)

s	a	s'	r
A	$\rightarrow$	B	0
B	$\downarrow$	$G2$	0
$G2$	exit	X	1

Episode 4 ($E4$)

s	a	s'	r
B	$\leftarrow$	A	0
A	$\downarrow$	$G1$	0
$G1$	exit	X	10

(a) Consider using temporal-difference learning to learn $V(s)$. When running TD-learning, all values are initialized to zero.
For which sequences of episodes, if repeated infinitely often, does $V(s)$ converge to $V^*(s)$ for all states s? (Assume appropriate learning rates such that all values converge.)
Write the correct sequence under "Other" if no correct sequences of episodes are listed.

☐ $E1, E2, E3, E4$ ☐ $E1, E2, E1, E2$ ☐ $E1, E2, E3, E1$ ☐ $E4, E4, E4, E4$
☐ $E4, E3, E2, E1$ ☐ $E3, E4, E3, E4$ ☐ $E1, E2, E4, E1$
☐ Other ______________________________

(b) Consider using Q-learning to learn $Q(s, a)$. When running Q-learning, all values are initialized to zero.
For which sequences of episodes, if repeated infinitely often, does $Q(s, a)$ converge to $Q^*(s, a)$ for all state-action pairs (s, a)
(Assume appropriate learning rates such that all Q-values converge.)
Write the correct sequence under "Other" if no correct sequences of episodes are listed.

☐ $E1, E2, E3, E4$ ☐ $E1, E2, E1, E2$ ☐ $E1, E2, E3, E1$ ☐ $E4, E4, E4, E4$
☐ $E4, E3, E2, E1$ ☐ $E3, E4, E3, E4$ ☐ $E1, E2, E4, E1$
☐ Other ______________________________