CSE 5525 Artificial Intelligence II

Quiz #3: MDPs and Reinforcement Learning

Wei Xu, Ohio State University

Your Name: _____ OSU Username: _____

1 Blackjack

In this question, you will play a simplified version of blackjack where the deck is infinite and the dealer always has a fixed count of 15. The deck contains cards 2 through 10, J, Q, K, and A, each of which is equally likely to appear when a card is drawn. Each number card is worth the number of points shown on it, the cards J, Q, and K are worth 10 points, and A is worth 11. At each turn, you may either *hit* or *stay*. If you choose to hit, you receive no immediate reward and are dealt an additional card. If you stay, you receive a reward of 0 if your current point total is exactly 15, +10 if it is higher than 15 but not higher than 21, and −10 otherwise (i.e. lower than 15 or larger than 21). After taking the *stay* action, the game enters a terminal state *end* and ends. A total of 22 or higher is refered to as a *bust*; from a bust, you can only choose the action *stay*. As your state space you take the set $\{0, 2, \ldots, 21, \text{bust}, \text{end}\}$ indicating point totals, "bust" if your point total exceeds 21, and "end" for the end of the game.

Questions:

1) Suppose you have performed k iterations of value iteration. Compute $V_{k+1}(12)$ given the partial table below for $V_k(s)$. Give your answer in terms of the discount γ as a variable. Note: do not worry about whether the listed V_k values could actually result from this MDP!

$$V_{k+1}(s) = \max_a \sum_{s'} T(s,a,s')\left(R(s,a,s') + \gamma V_k(s')\right)$$

a: {hit, stay}

s': 2, 3, 4, ..., 9, 10, J, Q, K, A

s	$V_k(s)$
13	2
14	10
15	10
16	10
17	10
18	10
19	10
20	10
21	10
bust	−10
end	0

$V_k(s') = 10$ $\quad$ $V_k(s') = -10$

$V_{k+1}(12) = ?$ $\frac{1}{13}(8 \cdot \gamma \cdot 10 + 5 \cdot \gamma \cdot (-10)) = \frac{30}{13}\gamma$

2) You suspect that the cards do not actually appear with equal probability and decide to use Q-learning instead of value iteration. Given the partial table of initial Q-values below, fill in the partial table of Q-values on the right after the following episode occurred. Assume a learning rate of 0.5 and a discount factor of 1. The initial portion of the episode has been omitted. Leave blank any values which Q-learning does not update.

Initial values

s	a	$Q(s,a)$
19	hit	−2
19	stay	5
20	hit	−4
20	stay	7
21	hit	−6
21	stay	8
bust	stay	−8

Episode

s	a	r	s'	a'	r'	s''	a''	r''
19	hit	0	21	hit	0	bust	stay	−10

Updated values ?

s	a	$Q(s,a)$
19	hit	3
19	stay	
20	hit	
20	stay	
21	hit	−7
21	stay	
bust	stay	−9

$$Q(19,\text{hit}) \leftarrow (1-\alpha)Q(19,\text{hit}) + \alpha \cdot \left(R(s,a,r) + \gamma \max_{a'} Q(s',a')\right)$$

$\alpha = 0.5$

$$(1-0.5) * -2 + 0.5 * (0 + 1 * \max(-6, 8)) = 3$$