CS 6300 **HW05: Policy Iteration and TD Learning** **Feb 15**

Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in via gradescope.

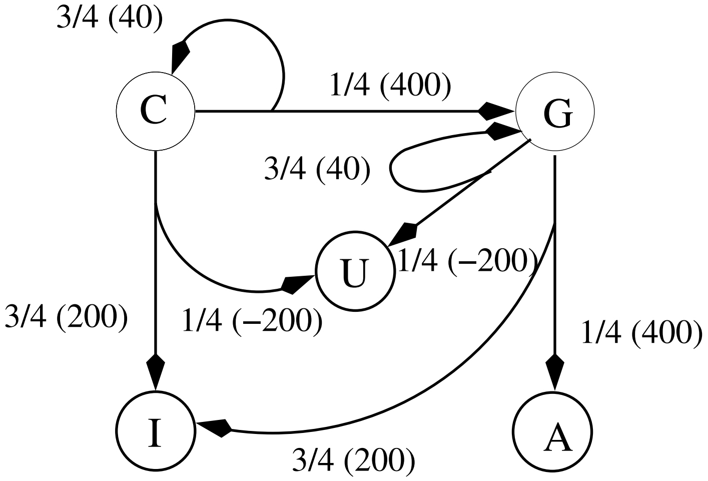
1 Temporal Difference Learning

We meet out beloved MDP again. There are 5 states: C(ollege), G(rad school), I(ndustry), A(cademia), and U(nemployed). States I, A and U are terminal states. The possible actions from states C and G are:

- State C: You may choose stayC, but with probability of 1/4 you end up going to state G.

  You may also choose to goI, but with probability 1/4 you end up in state U.
- State G: You may choose to stayG, but with probability 1/4 you end up in state U.

  You may also choose to goA, but with probability 3/4 you end up in state I.



For the MDP above, you decide to use experience and TD learning to find the values. You experience the following 3 episodes.

Episode 1			Episode 2			Episode 3		
S	A	R	S	A	R	S	A	R
C	stayC	40	C	stayC	40	C	stayC	400
C	stayC	40	C	goI	200	G	stayG	40
C	stayC	400	I			G	goA	400
G	stayG	40				A		
G	stayG	-200						
U								

The learning rate is $\alpha = (1/2)^n$, where n is the episode number. The discount factor is $\gamma = 1$. Perform TD learning to estimate the state values $V^\pi(S)$. All values should be initialized to 0.

2 Q-learning

In this simplied version of blackjack, 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 $\{0, 2, \ldots, 21, bust, end\}$ indicating point totals.

Given the partial table of initial Q-values below left, fill in the partial table of Q-values on the right after the episode center below occurs. Assume $\alpha = 0.5$ and $\gamma = 1$. The initial portion of the episode has been omitted. Show the derivation of the Q values that are updated.

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

s	a	r	s'
19	hit	0	21
21	hit	0	bust
bust	stay	-10	end

s	a	$Q(s,a)$
19	hit	
19	stay	
20	hit	
20	stay	
21	hit	
21	stay	
bust	stay	