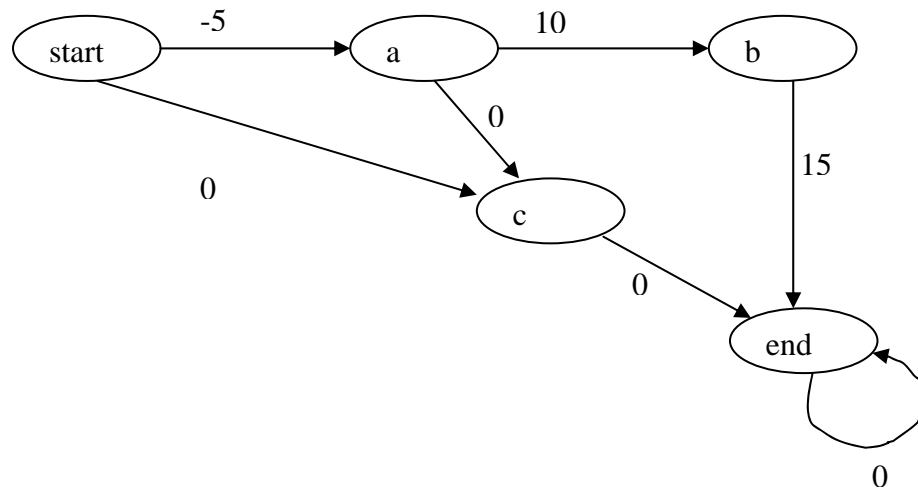


## CS 760 - Homework 4 Solution

1. Assume you wish to use a Q table to represent the Q function. All cells in this table should initially contain zero. Also assume your RL agent uses 1-step Q-learning. Show the state of your Q table after each of the following "episodes" (to represent the Q table, you can simply draw a copy of the above graph, but instead of attaching immediate rewards to arcs, attach the Q values). Be sure to show your work.

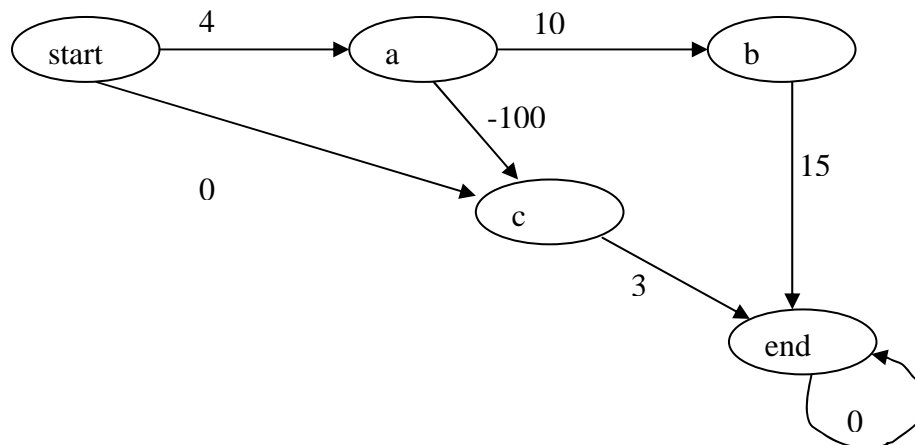
i. start  $\rightarrow$  a  $\rightarrow$  b  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{a}) = -5 + 0.9 \times 0 = -5$
- $Q(\text{a}, \text{a} \rightarrow \text{b}) = 10 + 0.9 \times 0 = 10$
- $Q(\text{b}, \text{b} \rightarrow \text{end}) = 15 + 0.9 \times 0 = 15$



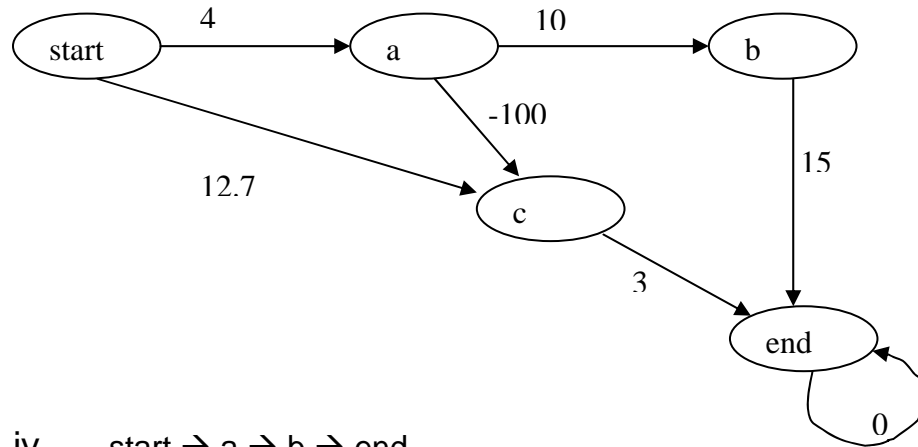
ii. start  $\rightarrow$  a  $\rightarrow$  c  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{a}) = -5 + 0.9 \times 10 = 4$
- $Q(\text{a}, \text{a} \rightarrow \text{c}) = -100 + 0.9 \times 0 = -100$
- $Q(\text{c}, \text{c} \rightarrow \text{end}) = 3 + 0.9 \times 0 = 3$



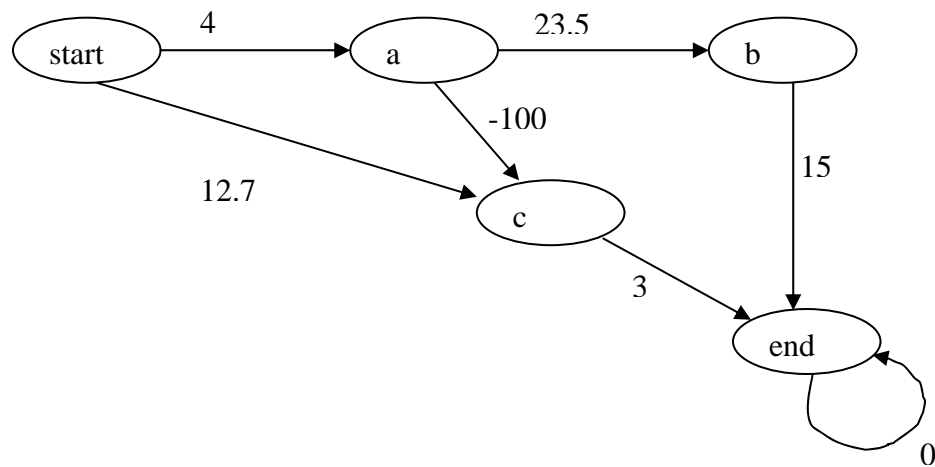
iii. start  $\rightarrow$  c  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{c}) = 10 + 0.9 \times 3 = 12.7$
- $Q(\text{c}, \text{c} \rightarrow \text{end}) = 3 + 0.9 \times 0 = 3$



iv. start  $\rightarrow$  a  $\rightarrow$  b  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{a}) = -5 + 0.9 \times 10 = 4$
- $Q(\text{a}, \text{a} \rightarrow \text{b}) = 10 + 0.9 \times 15 = 23.5$
- $Q(\text{b}, \text{b} \rightarrow \text{end}) = 15 + 0.9 \times 0 = 15$

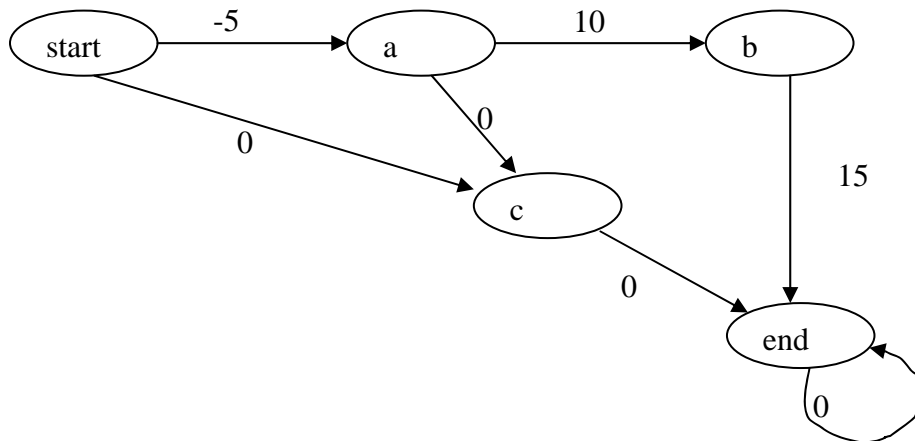


2) Repeat Part 1, using a fresh Q table (i.e, all cells filled with a zero), but this time use SARSA. For SARSA do you need to use  $\alpha$  (a "learning rate" - see Equation 13.10 of Mitchell)? If so, set  $\alpha$  as described in Lecture 24, slides 9-11. Explain your answer.

Solution: The "learning rate"  $\alpha$  should be used. The value of  $\alpha$  for a state-action edge should decay with the increasing number of visits to that edge: this ensures that the Q table learnt by SARSA will eventually converge.

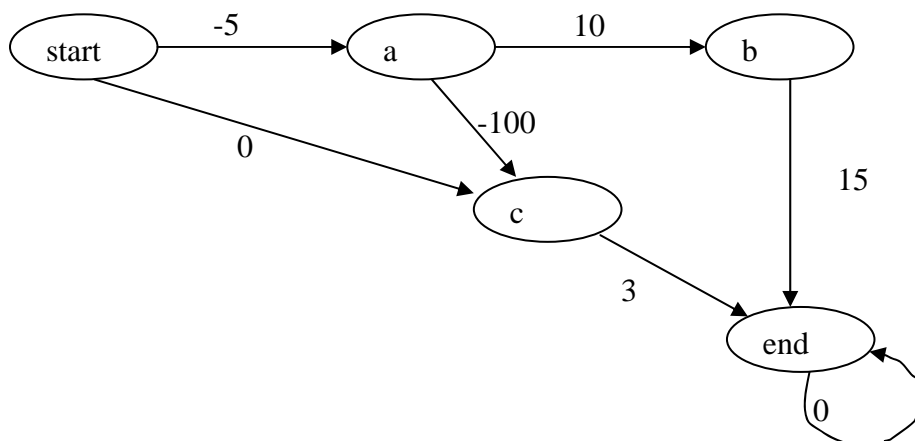
i. Start  $\rightarrow$  a  $\rightarrow$  b  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{a}) = Q(\text{start}, \text{start} \rightarrow \text{a}) + \alpha \times (-5 + 0.9 \times Q(\text{a}, \text{a} \rightarrow \text{b}) - Q(\text{start}, \text{start} \rightarrow \text{a}))$   
Where  $\alpha = 1$   
 $= 0 + 1 \times (-5 + 0 - 0) = -5$
- $Q(\text{a}, \text{a} \rightarrow \text{b}) = Q(\text{a}, \text{a} \rightarrow \text{b}) + \alpha \times (10 + 0.9 \times Q(\text{b}, \text{b} \rightarrow \text{end}) - Q(\text{a}, \text{a} \rightarrow \text{b}))$   
 $= 0 + 1 \times (10 + 0.9 \times 0 - 0) = 10$
- $Q(\text{b}, \text{b} \rightarrow \text{end}) = Q(\text{b}, \text{b} \rightarrow \text{end}) + \alpha \times (15 + 0.9 \times Q(\text{end}, \text{end} \rightarrow \text{end}) - Q(\text{b}, \text{b} \rightarrow \text{end}))$   
 $= 0 + 1 \times (15 + 0.9 \times 0 - 0) = 15$



2. start  $\rightarrow$  a  $\rightarrow$  c  $\rightarrow$  end

- $Q(\text{start}, \text{start} \rightarrow \text{a}) = Q(\text{start}, \text{start} \rightarrow \text{a}) + \alpha \times (-5 + 0.9 \times Q(\text{a}, \text{a} \rightarrow \text{c}) - Q(\text{start}, \text{start} \rightarrow \text{a}))$   
Where  $\alpha = 1/2$   
 $= -5 + \frac{1}{2} \times (-5 + 0.9 \times 0 - -5) = -5$
- $Q(\text{a}, \text{a} \rightarrow \text{c}) = Q(\text{a}, \text{a} \rightarrow \text{c}) + \alpha \times (-100 + 0.9 \times Q(\text{c}, \text{c} \rightarrow \text{end}) - Q(\text{a}, \text{a} \rightarrow \text{c}))$   
Where  $\alpha = 1$   
 $= 0 + 1 \times (-100 + 0.9 \times 0 - 0) = -100$
- $Q(\text{c}, \text{c} \rightarrow \text{end}) = Q(\text{c}, \text{c} \rightarrow \text{end}) + \alpha \times (15 + 0.9 \times Q(\text{end}, \text{end} \rightarrow \text{end}) - Q(\text{c}, \text{c} \rightarrow \text{end}))$   
Where  $\alpha = 1$   
 $= 0 + 1 \times (3 + 0.9 \times 0 - 0) = 3$



3. start  $\rightarrow$  c  $\rightarrow$  end

$$\square Q(\text{start}, \text{start} \rightarrow \text{c}) = Q(\text{start}, \text{start} \rightarrow \text{c}) + \alpha \times (10 + 0.9 \times Q(\text{c}, \text{c} \rightarrow \text{end}) - Q(\text{start}, \text{start} \rightarrow \text{c}))$$

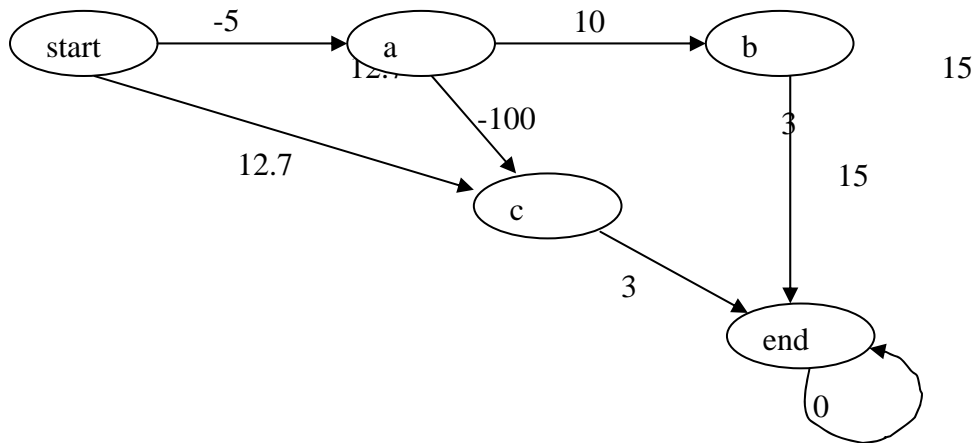
Where  $\alpha = 1$

$$= 0 + 1 \times (10 + 0.9 \times 3 - 0) = 12.7$$

$$\square Q(\text{c}, \text{c} \rightarrow \text{end}) = Q(\text{c}, \text{c} \rightarrow \text{end}) + \alpha \times (3 + 0.9 \times Q(\text{end}, \text{end} \rightarrow \text{end}) - Q(\text{c}, \text{c} \rightarrow \text{end}))$$

Where  $\alpha = 1/2$

$$= 3 + 1/2 \times (3 + 0.9 \times 0 - 3) = 3$$



4. Start  $\rightarrow$  a  $\rightarrow$  b  $\rightarrow$  end

$$\square Q(\text{start}, \text{start} \rightarrow \text{a}) = Q(\text{start}, \text{start} \rightarrow \text{a}) + \alpha \times (-5 + 0.9 \times Q(\text{a}, \text{a} \rightarrow \text{b}) - Q(\text{start}, \text{start} \rightarrow \text{a}))$$

Where  $\alpha = 1/3$

$$= -5 + 1/3 \times (-5 + 0.9 \times 10 - -5) = -2$$

$$\square Q(\text{a}, \text{a} \rightarrow \text{b}) = Q(\text{a}, \text{a} \rightarrow \text{b}) + \alpha \times (10 + 0.9 \times Q(\text{b}, \text{b} \rightarrow \text{end}) - Q(\text{a}, \text{a} \rightarrow \text{b}))$$

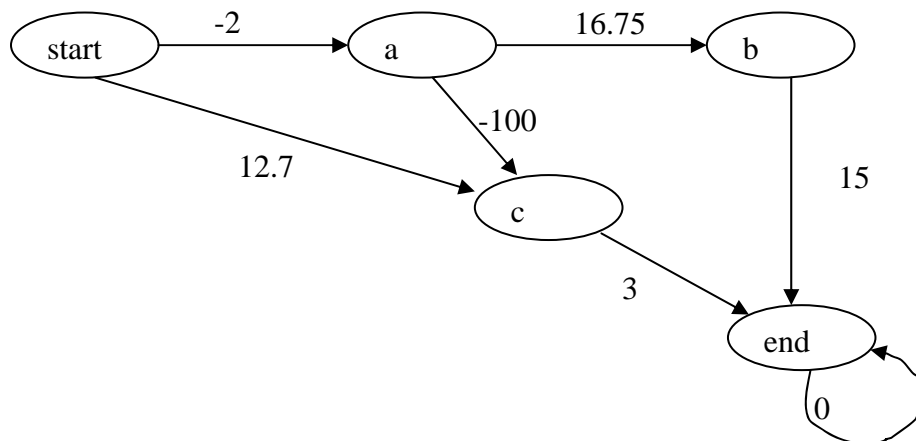
Where  $\alpha = 1/2$

$$= 10 + 1/2 \times (10 + 0.9 \times 15 - 10) = 16.75$$

$$\square Q(\text{b}, \text{b} \rightarrow \text{end}) = Q(\text{b}, \text{b} \rightarrow \text{end}) + \alpha \times (15 + 0.9 \times Q(\text{end}, \text{end} \rightarrow \text{end}) - Q(\text{b}, \text{b} \rightarrow \text{end}))$$

Where  $\alpha = 1/2$

$$= 15 + 1/2 \times (15 + 0.9 \times 0 - 15) = 15$$



3) Repeat Part 1, again using a fresh Q table, but this time use  $TD(\lambda), \lambda = \frac{1}{2}$ . Since  $TD(\lambda)$  involves a good deal of calculation, you only need to process the first two episodes of Part 1. Do you need to use  $\alpha$  here? Explain.

(Solution)  $\alpha$  is needed here because the errors introduced by the exploration steps (during look ahead) of  $TD(\lambda)$  should decay with the increasing number of visits (e.g.  $1/\alpha$ ) to ensure final convergences of Q table.

**Solution I** Based on the Defn of  $TD(\lambda)$  in Lec. 25 Slide 18/ Sutton Chap 7.2  
i. start  $\rightarrow$  a  $\rightarrow$  b  $\rightarrow$  end

$$\begin{aligned} \square \quad Q_t^\lambda(\text{start}, \text{start} \rightarrow \text{a}) &= Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow \text{a}) + \alpha \times [ \\ &\quad (1-\lambda) Q^{(1)}(\text{start}, \text{start} \rightarrow \text{a}) + \\ &\quad (1-\lambda)\lambda Q^{(2)}(\text{start}, \text{start} \rightarrow \text{a}) + \\ &\quad \lambda^2 Q^{(3)}(\text{start}, \text{start} \rightarrow \text{a}) - Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow \text{a})] \\ &= 0.5 \times [-5 + 0.5 \times 4] + 0.5^2 \times 16.15 = 2.54 \end{aligned}$$

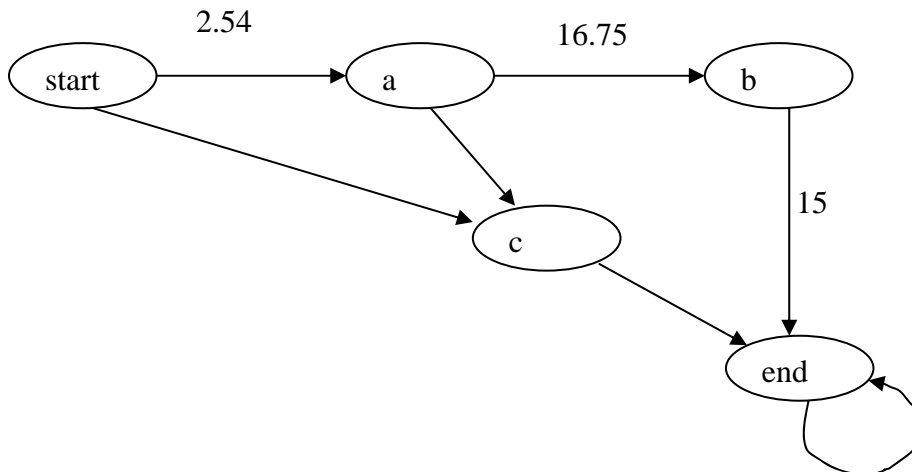
$$\begin{aligned} \text{where } Q^{(1)}(\text{start}, \text{start} \rightarrow \text{a}) &= -5 + 0.9 \times 0 = -5 \\ Q^{(2)}(\text{start}, \text{start} \rightarrow \text{a}) &= -5 + 0.9 \times 10 + 0.81 \times 0 = 4 \\ Q^{(3)}(\text{start}, \text{start} \rightarrow \text{a}) &= -5 + 0.9 \times 10 + 0.81 \times 15 + 0.9^3 \times 0 = 16.15 \\ \alpha &= 1 \text{ and } Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow \text{a}) = 0 \end{aligned}$$

$$\begin{aligned} \square \quad Q_t^\lambda(\text{a}, \text{a} \rightarrow \text{b}) &= Q_{t-1}^\lambda(\text{a}, \text{a} \rightarrow \text{b}) + \\ &\quad \alpha \times [(1-\lambda) Q^{(1)}(\text{a}, \text{a} \rightarrow \text{b}) + \lambda Q^{(2)}(\text{a}, \text{a} \rightarrow \text{b}) - Q_{t-1}^\lambda(\text{a}, \text{a} \rightarrow \text{b})] \\ &= 0.5 \times 10 + 0.5 \times 23.5 = 16.75 \end{aligned}$$

$$\begin{aligned} \text{where } Q^{(1)}(\text{a}, \text{a} \rightarrow \text{b}) &= 10 + 0.9 \times 0 = 10 \\ Q^{(2)}(\text{a}, \text{a} \rightarrow \text{b}) &= 10 + 0.9 \times 15 + 0.81 \times 0 = 23.5 \\ \alpha &= 1 \text{ and } Q_{t-1}^\lambda(\text{a}, \text{a} \rightarrow \text{b}) = 0 \end{aligned}$$

$$\begin{aligned} \square \quad Q_t^\lambda(\text{b}, \text{b} \rightarrow \text{end}) &= Q_{t-1}^\lambda(\text{b}, \text{b} \rightarrow \text{end}) + \alpha \times [ Q^{(1)}(\text{b}, \text{b} \rightarrow \text{end}) - Q_{t-1}^\lambda(\text{b}, \text{b} \rightarrow \text{end})] \\ &= 15 \end{aligned}$$

$$\begin{aligned} \text{where } Q^{(1)}(\text{b}, \text{b} \rightarrow \text{end}) &= 15 + 0.9 \times 0 = 15 = Q^{(n)}(\text{b}, \text{b} \rightarrow \text{end}) \\ \alpha &= 1 \text{ and } Q_{t-1}^\lambda(\text{b}, \text{b} \rightarrow \text{end}) = 0 \end{aligned}$$



i. start  $\rightarrow$  a  $\rightarrow$  c  $\rightarrow$  end

$$\begin{aligned} \square Q_t^\lambda(\text{start}, \text{start} \rightarrow a) &= Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a) + \\ &\quad \alpha \times [(1-\lambda) Q^{(1)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad (1-\lambda)\lambda Q^{(2)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad \lambda^2 Q^{(3)}(\text{start}, \text{start} \rightarrow a) \\ &\quad - Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a)] \\ &= 2.54 + 0.5 \times (-41.86 - 2.54) = -19.66 \end{aligned}$$

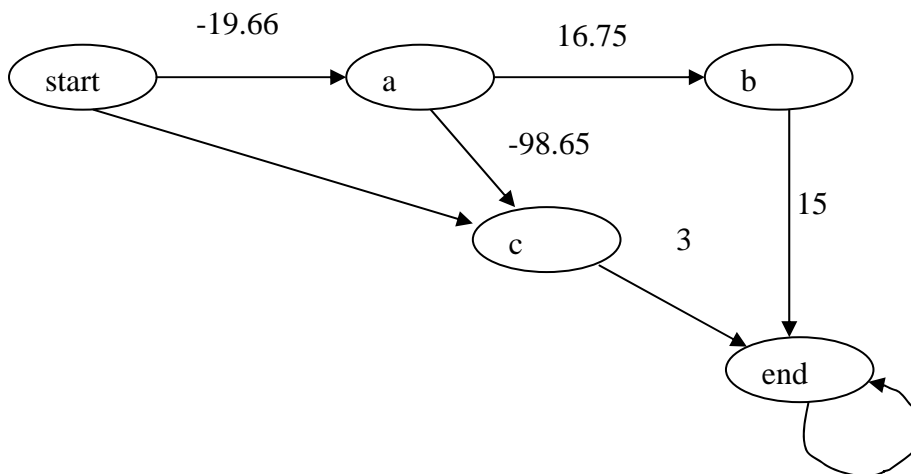
where  $Q^{(1)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times 10.88 = 4.79$   
 $Q^{(2)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times -100 + 0.81 \times 0 = -95$   
 $Q^{(3)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times -100 + 0.81 \times 3 + 0.9^3 \times 0 = -92.57$   
 $\alpha = \frac{1}{2}$  and  $Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a) = 2.54$

$$\begin{aligned} \square Q_t^\lambda(a, a \rightarrow c) &= Q_{t-1}^\lambda(a, a \rightarrow c) + \\ &\quad \alpha \times [(1-\lambda) Q^{(1)}(a, a \rightarrow c) + \lambda Q^{(2)}(a, a \rightarrow c) - Q_{t-1}^\lambda(a, a \rightarrow c)] \\ &= 0.5 \times -100 + 0.5 \times -97.3 = -98.65 \end{aligned}$$

where  $Q^{(1)}(a, a \rightarrow c) = -100 + 0.9 \times 0 = -100$   
 $Q^{(2)}(a, a \rightarrow c) = -100 + 0.9 \times 3 + 0.81 \times 0 = -97.3$

$$\square Q_t^\lambda(c, c \rightarrow \text{end}) = Q_{t-1}^\lambda(c, c \rightarrow \text{end}) + \alpha \times [Q^{(1)}(c, c \rightarrow \text{end}) - Q_{t-1}^\lambda(c, c \rightarrow \text{end})] = 3$$

where  $Q^{(1)}(c, c \rightarrow \text{end}) = 3 + 0.9 \times 0 = 3$



**Solution II** Some of the answers assume the end state has self loop and thus make the Q value of an edge as a sum of infinite power series. We accept the answers, but notice that the provided trajectories stopped at the 'end' state (the HW should have said 'end' was a terminating state, rather than using the self-loop with zero immediate reward).

i.  $\text{start} \rightarrow a \rightarrow b \rightarrow \text{end}$

$$\begin{aligned} \square \quad Q_t^\lambda(\text{start}, \text{start} \rightarrow a) &= (1 - \lambda) [Q^{(1)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad \lambda Q^{(2)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad \lambda^2 Q^{(3)}(\text{start}, \text{start} \rightarrow a) + \dots \\ &\quad \lambda^{n-1} Q^{(n)}(\text{start}, \text{start} \rightarrow a)] \\ &= 0.5 \times [-5 + 0.5 \times 4 + 0.5 \times 16.15] = 2.54 \end{aligned}$$

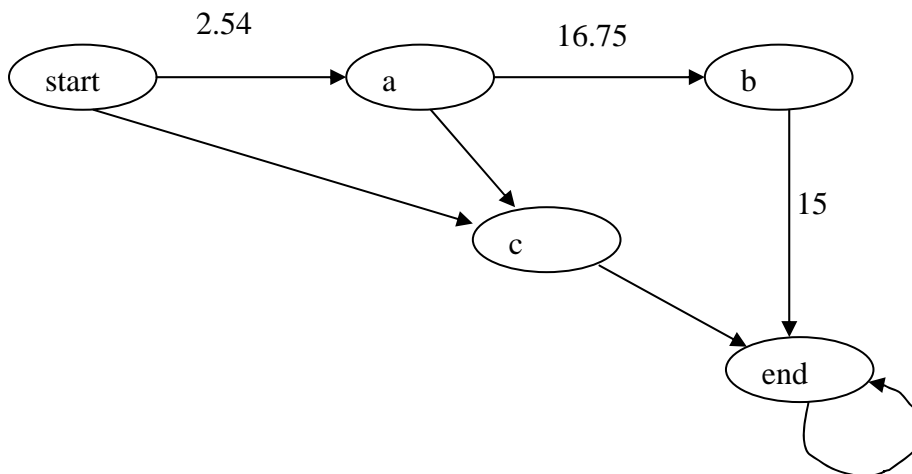
$$\begin{aligned} \text{where } Q^{(1)}(\text{start}, \text{start} \rightarrow a) &= -5 + 0.9 \times 0 = -5 \\ Q^{(2)}(\text{start}, \text{start} \rightarrow a) &= -5 + 0.9 \times 10 + 0.81 \times 0 = 4 \\ Q^{(3)}(\text{start}, \text{start} \rightarrow a) &= -5 + 0.9 \times 10 + 0.81 \times 15 + 0.9^3 \times 0 = 16.15 \\ &\dots \\ Q^{(n)}(\text{start}, \text{start} \rightarrow a) &= Q^{(3)}(\text{start}, \text{start} \rightarrow a) \end{aligned}$$

$$\begin{aligned} \square \quad Q_t^\lambda(a, a \rightarrow b) &= (1 - \lambda) [Q^{(1)}(a, a \rightarrow b) + \lambda Q^{(2)}(a, a \rightarrow b) + \dots + \lambda^{n-1} Q^{(n)}(a, a \rightarrow b)] \\ &= 0.5 \times [10 + 23.5] = 16.75 \end{aligned}$$

$$\begin{aligned} \text{where } Q^{(1)}(a, a \rightarrow b) &= 10 + 0.9 \times 0 = 10 \\ Q^{(2)}(a, a \rightarrow b) &= 10 + 0.9 \times 15 + 0.81 \times 0 = 23.5 = Q^{(n)}(a, a \rightarrow b) \end{aligned}$$

$$\begin{aligned} \square \quad Q_t^\lambda(b, b \rightarrow \text{end}) &= (1 - \lambda) [Q^{(1)}(b, b \rightarrow \text{end}) + \dots + \lambda^{n-1} Q^{(n)}(b, b \rightarrow \text{end})] \\ &= 0.5 \times [15/2] = 15 \end{aligned}$$

$$\text{where } Q^{(1)}(b, b \rightarrow \text{end}) = 15 + 0.9 \times 0 = 15 = Q^{(n)}(b, b \rightarrow \text{end})$$



i. start  $\rightarrow$  a  $\rightarrow$  c  $\rightarrow$  end

$$\begin{aligned} \square Q_t^\lambda(\text{start}, \text{start} \rightarrow a) &= Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a) + \alpha \times [(1-\lambda) [Q^{(1)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad \lambda Q^{(2)}(\text{start}, \text{start} \rightarrow a) + \\ &\quad \lambda^2 Q^{(3)}(\text{start}, \text{start} \rightarrow a) + \dots + \\ &\quad \lambda^{n-1} Q^{(n)}(\text{start}, \text{start} \rightarrow a) \\ &\quad - Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a)] \\ &= -19.66 \end{aligned}$$

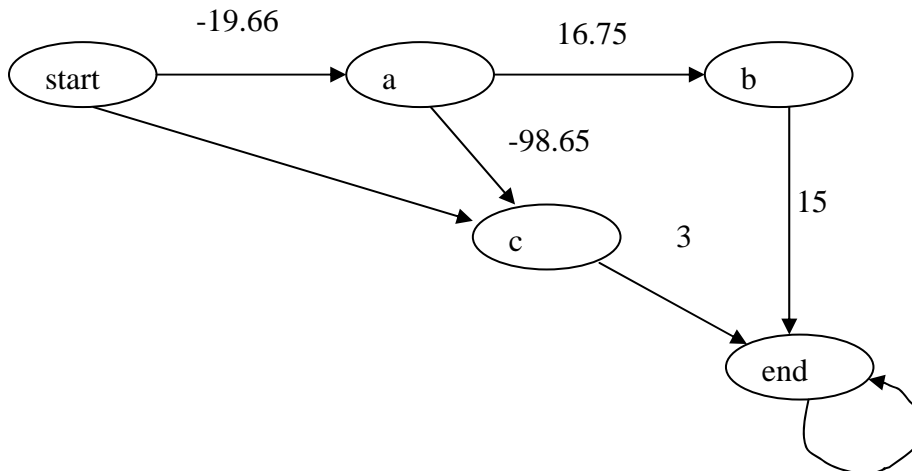
where  $Q^{(1)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times 16.75 = 10.08$   
 $Q^{(2)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times -100 + 0.81 \times 0 = -95$   
 $Q^{(3)}(\text{start}, \text{start} \rightarrow a) = -5 + 0.9 \times -100 + 0.81 \times 3 + 0.9^3 \times 0 = -92.57$   
 $\dots$   
 $Q^{(n)}(\text{start}, \text{start} \rightarrow a) = Q^{(3)}(\text{start}, \text{start} \rightarrow a)$   
 $\alpha = 0.5$  and  $Q_{t-1}^\lambda(\text{start}, \text{start} \rightarrow a) = 2.54$

$$\begin{aligned} \square Q_t^\lambda(a, a \rightarrow c) &= (1-\lambda) [Q^{(1)}(a, a \rightarrow c) + \lambda Q^{(2)}(a, a \rightarrow c) + \dots + \lambda^{n-1} Q^{(n)}(a, a \rightarrow c)] \\ &= 0.5 \times [-100 + -97.3] = -98.65 \end{aligned}$$

where  $Q^{(1)}(a, a \rightarrow c) = -100 + 0.9 \times 0 = -100$   
 $Q^{(2)}(a, a \rightarrow c) = -100 + 0.9 \times 3 + 0.81 \times 0 = -97.3 = Q^{(n)}(a, a \rightarrow c)$

$$\begin{aligned} \square Q_t^\lambda(c, c \rightarrow \text{end}) &= (1-\lambda) [Q^{(1)}(c, c \rightarrow \text{end}) + \dots + \lambda^{n-1} Q^{(n)}(c, c \rightarrow \text{end})] \\ &= 0.5 \times [3 \times 2] = 3 \end{aligned}$$

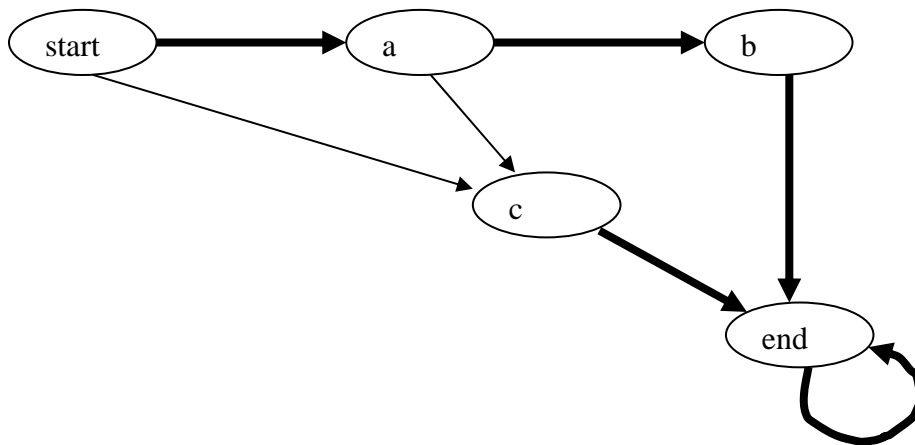
where  $Q^{(1)}(c, c \rightarrow \text{end}) = 3 + 0.9 \times 0 = 3 = Q^{(n)}(c, c \rightarrow \text{end})$



4. If you performed RL for a large number of episodes, what policy would Q learning produce? Indicate this policy by copying the above graph and using thick arrows to represent the policy. Briefly explain your answer.

(Solution) Q learning will produce a policy which maximizes the total gain from a given state to the end state. The main reason is that Q learning always chooses an action with the maximum current Q-value. When Q learning converges, each state-action edge will have the maximum total gain.

Note: Even c is a state cannot be reached from a start state in the optimal policy, we still should compute the best action on that state (in case the agent could be 'dropped' into any state at the start).



5. Repeat Part 4 using SARSA. Show this policy on a fresh copy of the graph.

(Solution) Q learning will produce a policy which maximizes the *expected* total gain (over the exploration/exploitation probability) from a given state to the end state. In our example, the policy happens to be the same as (4). The reason is that the "danger" of going from  $a \rightarrow c$  doesn't outweigh the benefit of going from  $a \rightarrow b$  due to the low exploration probability of 0.05. The situation will change if we, say, make the reward of  $a \rightarrow c$  as  $-1,000,000$ . In that case,  $start \rightarrow c$  will be chosen.

