COMS W4701 HW3

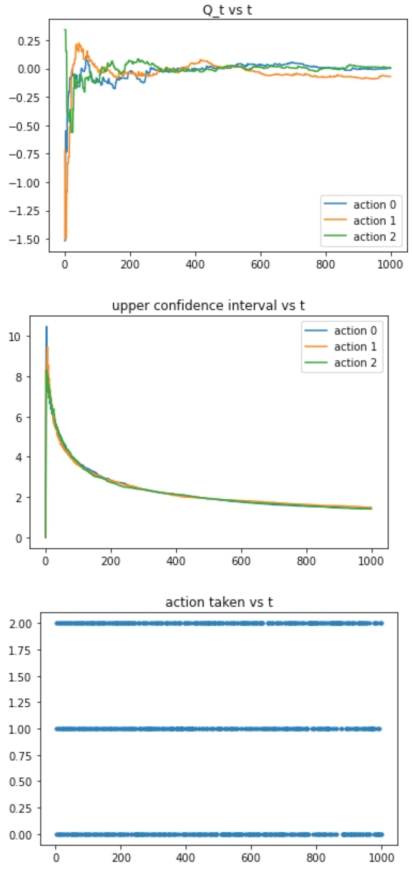
Dawei He (dh3027)

March 9, 2022

Problem 1

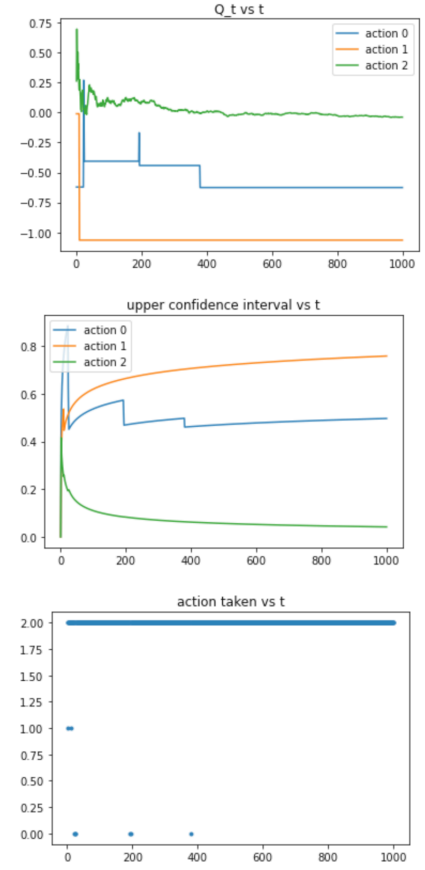
(a)

```
means = [0, 0, 0]
variances = [1, 1, 1]
UCB_bandit(means, variances,10)  means: [0, 0, 0]
```



(a) Experiment 1

```
means = [0, 0, 0]
variances = [1, 1, 1]
UCB_bandit(means, variances,0.5)  means: [0, 0, 0
```



(b) Experiment 2

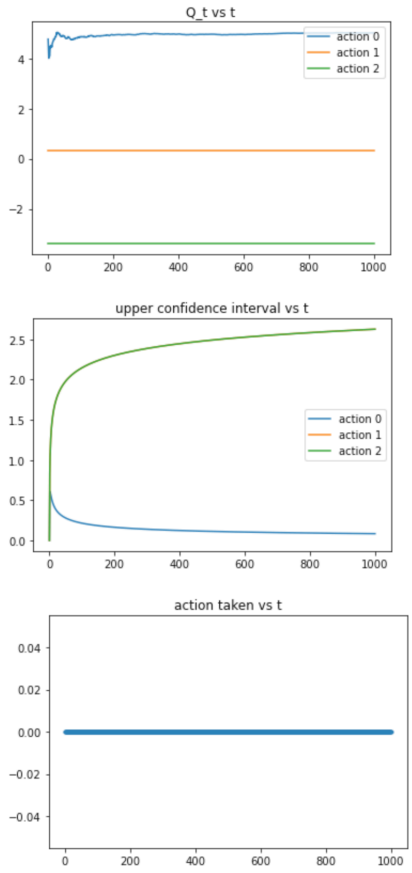
The c value is set to be 10 in experiment 1, which encourage more exploration. However c value is set to be 0.5 in experiment 2, which will lead to exploitation after only few iterations.

All confidence intervals are keep decreasing in the experiment 1, but two confidence intervals are keep increasing in the experiment 2. That means two Q values are not converge in experiment 2.

The distribution of three actions taken are equally the same in experiment 1. But for the experiment 2, action 2 is taken for the most of time.

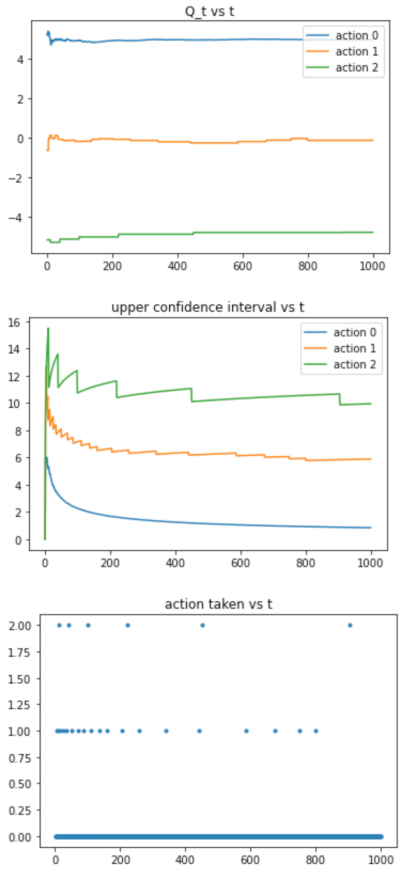
(b)

```
means = [5, 0, -5]
variances = [1, 1, 1]
UCB_bandit(means, variances,1)  means: [5, 0, -5]
```



(a) $c = 1$

```
means = [5, 0, -5]
variances = [1, 1, 1]
UCB_bandit(means, variances,10)  means: [5, 0, -5]
```



(b) $c = 10$

For $c = 1$ scenario, there is some fluctuation of Q value of all action 0, 1 and 2. They converged to the true value eventually. All confidence intervals of three actions keep decreasing. The confidence interval of action 0 decreased most to 0 during iterations and then the action 1 and two. The distribution plot shows that action 1 and 2 were taken from time to time during iterations, but they are much less than action 0. The action 0 was taken most of time during iterations.

For $c = 10$ scenario, there is some fluctuation of Q value of action 0 and it converged to the true value eventually. And for action 1 and action 2, their Q values stay almost the same with no variation and they did not converge to their true values. The confidence intervals of action 1 and 2 keep increasing with iterations, and only the confidence interval of action 0 decrease. The distribution plot shows that there is barely no action 1 and 2 were taken during iterations. The action 0 dominate the distribution.

(c)

```
means = [1, 0, -1]
variances = [10, 10, 10]
UCB_bandit(means, variances,1)  means: [1, 0, -1]
```

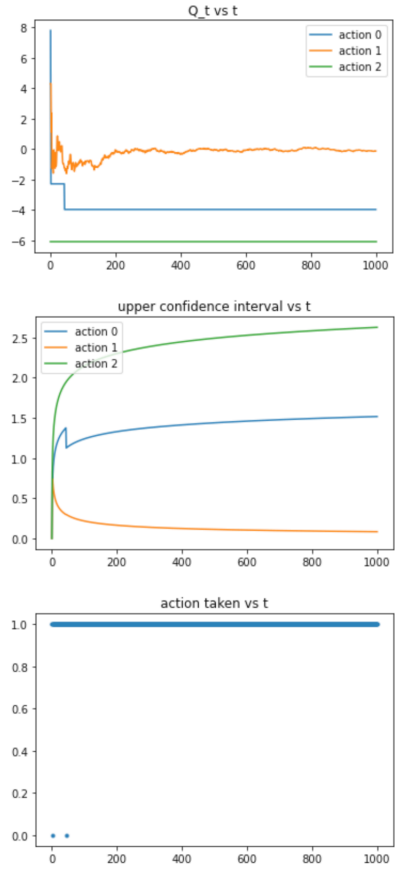


Figure 3: $c = 1$

Because when $c = 1$, there is no much exploration for the agent. And because all three actions have very close mean and large variation, the agent can find out accurately which action has the largest outcome in only very few tries. So there is a large probability that action 0 was not giving the largest value in these first few tries, but action 1 did gave the largest values due to the similar mean and large variance. So the agent will stop exploring due to the small c value and start to exploit action 1 instead of action 0. But if c is large, the agent can have a lot of explorations, so even the variance of bandits are large, the agent can still accurately approximate the true mean of bandit from a lot of explorations. So c being large is important when bandits have large variances.

Problem 2

(a)

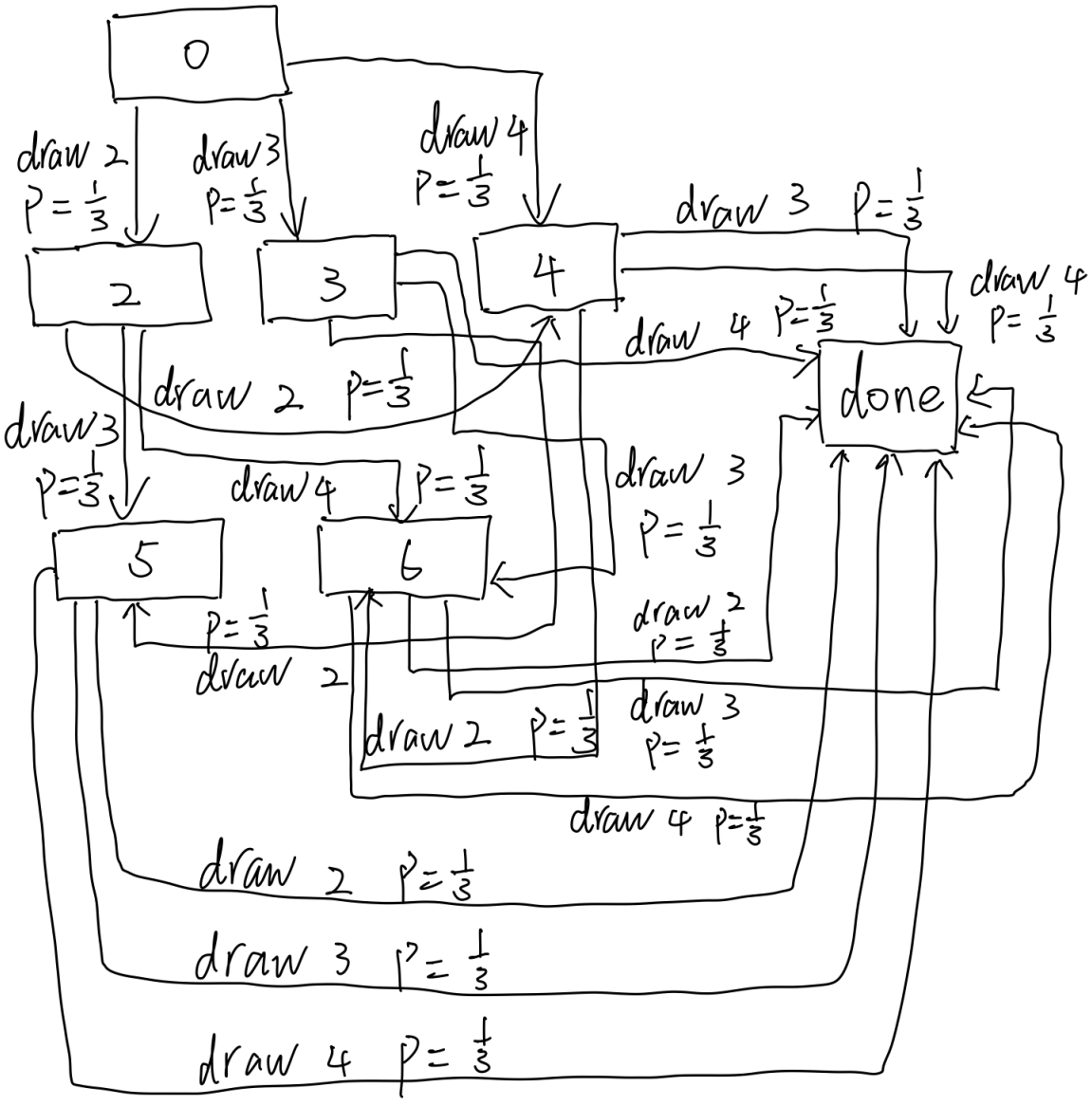


Figure 4: State transition diagram

(b)

The optimal action of state 5 and 6 is to stop. The values of state 5 and state 6 are $V^*(5) = 5$ and $V^*(6) = 6$.

$$V^{draw}(4) = \frac{1}{3}[0 + V^*(6)] + \frac{1}{3}[0 + 0] + \frac{1}{3}[0 + 0] = \frac{1}{3}[0 + 6] = 2$$

$$V^{stop}(4) = 4$$

$$\Downarrow$$

$$\pi^*(4) = stop$$

$$V^*(4) = 4$$

$$V^{draw}(3) = \frac{1}{3}[0 + V^*(5)] + \frac{1}{3}[0 + V^*(6)] + \frac{1}{3}[0 + 0] = \frac{1}{3}[0 + 5] + \frac{1}{3}[0 + 6] = \frac{11}{3}$$

$$V^{stop}(3) = 3$$

$$\Downarrow$$

$$\pi^*(3) = draw$$

$$V^*(3) = \frac{11}{3}$$

$$V^{draw}(2) = \frac{1}{3}[0 + V^*(4)] + \frac{1}{3}[0 + V^*(5)] + \frac{1}{3}[0 + V^*(6)] = \frac{1}{3}[0 + 4] + \frac{1}{3}[0 + 5] + \frac{1}{3}[0 + 6] = 5$$

$$V^{stop}(2) = 2$$

$$\Downarrow$$

$$\pi^*(2) = draw$$

$$V^*(2) = 5$$

$$V^{draw}(0) = \frac{1}{3}[0 + V^*(2)] + \frac{1}{3}[0 + V^*(3)] + \frac{1}{3}[0 + V^*(4)] = \frac{1}{3}[0 + 5] + \frac{1}{3}[0 + \frac{11}{3}] + \frac{1}{3}[0 + 4] = \frac{38}{9}$$

$$V^{stop}(0) = 0$$

$$\Downarrow$$

$$\pi^*(0) = draw$$

$$V^*(0) = \frac{38}{9}$$

Because there is only very limit states, so we can do the recursive call without worry about having not enough memory. So DP is not required in this problem.

(c)

$$\frac{1}{3}[0+\gamma\times 5]+\frac{1}{3}[0+\gamma\times 6]\leq V^{stop}(3)=3 \tag{1}$$

$$\frac{1}{3}[0+\gamma\times 4]+\frac{1}{3}[0+\gamma\times 5]+\frac{1}{3}[0+\gamma\times 6]\leq V^{stop}(2)=2 \tag{2}$$

(1) $\Rightarrow \gamma \leq \frac{9}{11}$ and (2) $\Rightarrow \gamma \leq \frac{2}{5}$
$\gamma = \frac{2}{5}$ would possibly lead to different optimal actions in both state 2 and 3.
When $\gamma = \frac{2}{5}$:

$\pi^*(3) = stop$

$V^*(3) = 3$

$\pi^*(2) = draw$ or $stop$

$V^*(2) = 2$

$$\begin{aligned} V^{draw}(0) &= \frac{1}{3}[0+\gamma\times V^*(2)]+\frac{1}{3}[0+\gamma\times V^*(3)]+\frac{1}{3}[0+\gamma\times V^*(4)] \\ &= \frac{1}{3}[0+\frac{2}{5}\times 2]+\frac{1}{3}[0+\frac{2}{5}\times 3]+\frac{1}{3}[0+\frac{2}{5}\times 4] \\ &= \frac{6}{5} \end{aligned}$$

$V^{stop}(0) = 0$

$\Downarrow$

$\pi^*(0) = draw$

$V^*(0) = \frac{6}{5}$

When $\gamma = 1$, the optimal action for state 0, 2 and 3 is to draw. When γ decrease, V^{draw}(state 0, 2 or 3)decrease as well. When γ is small enough, the optimal of state 2 and 3 become to stop, so their values become the V^{stop}. But for state 4, 5 and 6, their optimal action is always to stop. Their V^{draw} always smaller than V^{stop}, so even V^{draw} keep decreasing, it would not change the V^* of them.

Problem 3

(a)

$V_1(s) = s$

The state value should be the reward received when stopping action is taken at each state, which is the number of s of each state. Because if we choose to draw, there is no immediate reward which is 0. And the value of successor state is also 0 because of the initialization in in V_0. So the max value of each state should be the reward received when stopping action is taken.

(b)

$\pi_1(s) = stop$

The new policies π_1 for all s are to stop. Because the values of s when stopping is taken are greater or equal than 0. And the values of s when drawing is taken are all 0. So values of stopping are all greater or equal than values of drawing for all states.

(c)

$$V_2(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_1(s')]$$

$$V_2(done) = \max_a\{0\} = 0$$

$$V_2(0) = \max_a\{\frac{1}{3}[0 + V_1(2)] + \frac{1}{3}[0 + V_1(3)] + \frac{1}{3}[0 + V_1(4)], 0\} = \frac{9}{3}$$

$$V_2(2) = \max_a\{\frac{1}{3}[0 + V_1(4)] + \frac{1}{3}[0 + V_1(5)] + \frac{1}{3}[0 + V_1(6)], 2\} = 5$$

$$V_2(3) = \max_a\{\frac{1}{3}[0 + V_1(5)] + \frac{1}{3}[0 + V_1(6)] + \frac{1}{3}[0 + V_1(done)], 3\} = \frac{11}{3}$$

$$V_2(4) = \max_a\{\frac{1}{3}[0 + V_1(6)] + \frac{1}{3}[0 + V_1(done)] + \frac{1}{3}[0 + V_1(done)], 4\} = 4$$

$$V_2(5) = \max_a\{\frac{1}{3}[0 + V_1(done)] + \frac{1}{3}[0 + V_1(done)] + \frac{1}{3}[0 + V_1(done)], 5\} = 5$$

$$V_2(6) = \max_a\{\frac{1}{3}[0 + V_1(done)] + \frac{1}{3}[0 + V_1(done)] + \frac{1}{3}[0 + V_1(done)], 6\} = 6$$

Value of state 0 have not converged.

(d)

$$\pi^*(s) = \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_1(s')]$$

According the value calculated in (c)

$\pi_2^*(0) = draw$

$\pi_2^*(2) = draw$

$\pi_2^*(3) = draw$

$\pi_2^*(4) = stop$

$\pi_2^*(5) = stop$

$\pi_2^*(6) = stop$

All policies have converged.

Problem 4

(a)

$G_1(0) = 0$

$G_2(0) = 0$

$G_3(0) = 5$

$G_4(0) = 5$

$G_5(0) = 6$

$V^\pi(0) = \frac{1}{5}(0 + 0 + 5 + 5 + 6) = \frac{16}{5}$

$G_1(2) = 0$

$G_3(2) = 5$

$V^\pi(2) = \frac{1}{2}(0 + 5) = \frac{5}{2}$

$G_2(3) = 0$

$G_4(3) = 5$

$V^\pi(3) = \frac{1}{2}(0 + 5) = \frac{5}{2}$

$G_1(4) = 0$

$G_5(4) = 6$

$V^\pi(4) = \frac{1}{2}(0 + 6) = 3$

$G_3(5) = 5$

$G_4(5) = 5$

$V^\pi(5) = \frac{1}{2}(5 + 5) = 5$

$G_5(6) = 6$

$V^\pi(6) = \frac{1}{1}(6) = 6$

The order does not affect the estimated state value. Because we will average utilities over multiple episodes, changing the order of them dose not change the sum up value, so it will not change the average values.

(b)

Episode 1:

Transition	(0,+0)	(2,+0)	(4,+0)
$V^\pi(0)$	0	0	0
$V^\pi(2)$	0	0	0
$V^\pi(4)$	0	0	0

Episode 2:

Transition	(0,+0)	(3,+0)
$V^\pi(0)$	0	0
$V^\pi(3)$	0	0

Episode 3:

Transition	(0,+0)	(2,+0)	(5,+5)
$V^\pi(0)$	0	0	0
$V^\pi(2)$	0	0	0
$V^\pi(5)$	0	0	2.5

Episode 4:

Transition	(0,+0)	(3,+0)	(5,+5)
$V^\pi(0)$	0	0	0
$V^\pi(3)$	0	1.25	1.25
$V^\pi(5)$	2.5	2.5	3.75

Episode 5:

Transition	(0,+0)	(4,+0)	(6,+6)
$V^\pi(0)$	0	0	0
$V^\pi(4)$	0	0	0
$V^\pi(6)$	0	0	3

The order of episodes will affects the estimated state values, because when using TD(0) to update $V^\pi(s)$, it will take in to account of the successor state value $V^\pi(s')$. If we change the order, $V^\pi(s')$ will also change, so $V^\pi(s)$ will change accordingly. For example if episode 2 is placed after 3 and 4, $V^\pi(0)$ and $V^\pi(3)$ will have different values.

(c)

I think for state 2, 3 and 4 will converge differently by two methods. Because when action drawing is taken at these states, the next states of them could be state 5 and 6. The optimal action for state 5 and 6 are to stop. When Q learning is used, the agent will assume stopping action will always be taken at state 5 and 6 to get the largest Q value. But for SARSA, exploratory action will still be occasionally taken, which means drawing action could still be taken at state 5 and 6. So Q value of state 2, 3 and 4 by using SARSA method should be smaller than the values by using Q learning.

Problem 5

5.3

(a)

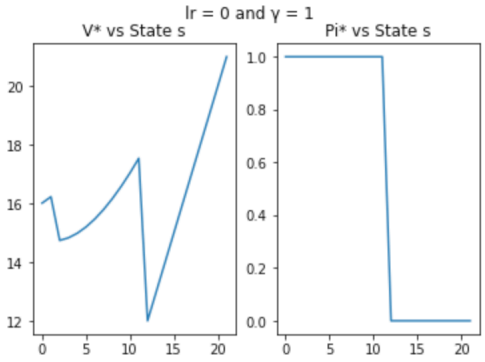


Figure 5: Plot of values V^* and policy π^* for living reward $lr = 0$ and $\gamma = 1$

When state s is less or equal than 11, drawing is always the best choice for agent, because the sum will not exceed 21. However, when state is larger or equal than 12, there is a chance that drawing another card will exceed 21, so the optimal choice become stopping. Therefore, state value will have suddenly drop between 10 and 11. In this case, the living reward is 0 and γ is 1, so the state value following drawing policy is fully dependent on next state's value. State 1 will depend on 2 to 11, state 2 will depend on 2 to 12. Because there is a drop between 11 and 12, such a drop will be also reflected on values of state 1 and 2.

(b)

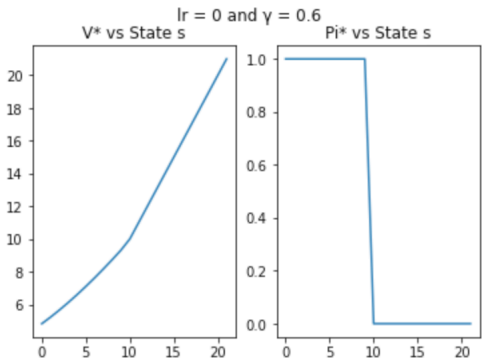
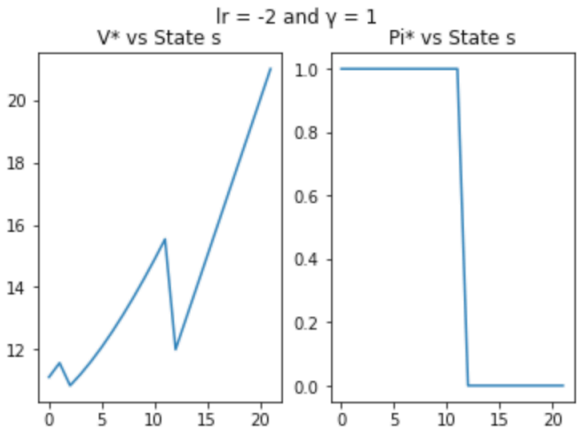
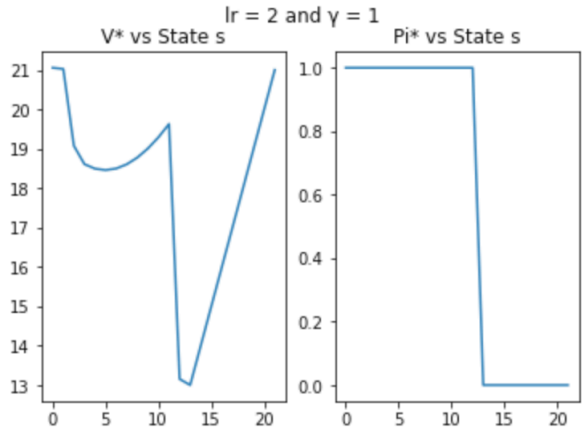


Figure 6: Plot of values V^* and policy π^* for living reward $lr = 0$ and $\gamma = 0.6$

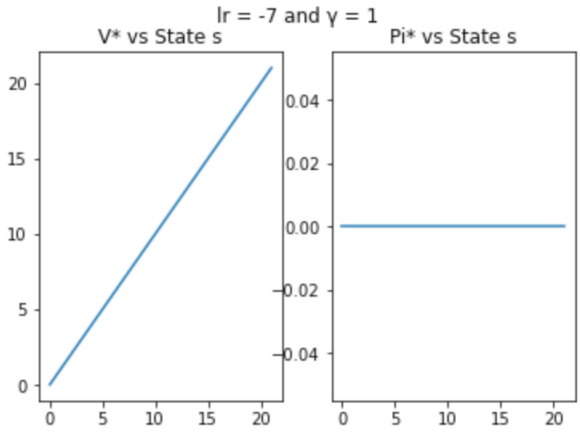
In this case, the living reward is 0 and γ is 0.6, so the state value following drawing policy is fully dependent on next state's discount value. Because of the low discount rate, the value of state 10 following drawing policy is relatively low compare to (a), which is low enough to the value of stop policy so it wipe out the sudden drop in (a). So the action of state 10 also become stopping instead of drawing. Such phenomenal is also reflected on state 0 and 1, so there is only two segment now.

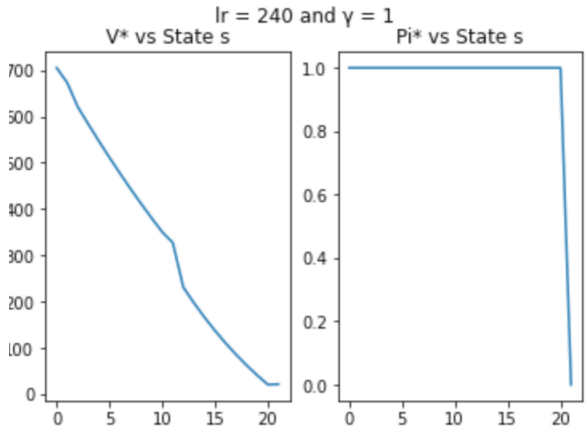
(c)





We can clearly see that segment from v^* 0 to 11 shift for negative living rewards or slightly positive living rewards. In this case when lr change from -2 to 2, the segment from v^* 0 to 11 shift upward. If state value shift upward, π^* will change from stop to draw. If state value shift downward, π^* will change from draw to stop.





The approximate thresholds of the living reward in which π^* becomes stop in all states is -7.
The approximate thresholds of the living reward in which π^* becomes draw in all states is 240.

5.5

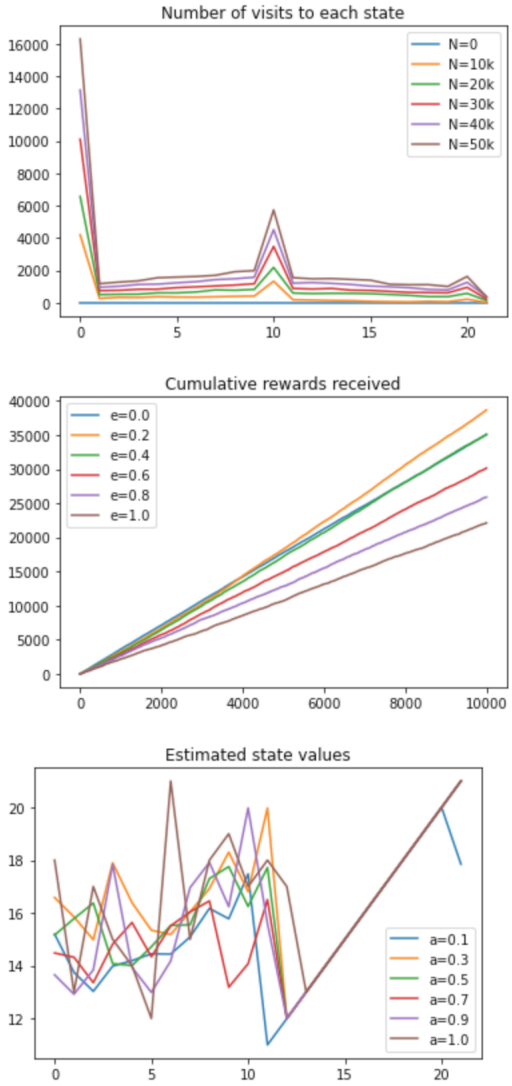


Figure 9: Result plots of the RL_analysis function

(a)

Because most of visits is given to state 0 and state 10, there are comparatively fewer visits are given to other states. If there is no enough total visit times, which is N here, values of states other than 0 and 10 can not converge due to the limit visit times. So N has to be large to ensure values of all states can converge.

Because every time a stop action is taken or the sum of all cards exceed 21, the agent will go back to state 0. So the state 0 is the most visited state. The state 10 is the second visited state because drawing card with value 10 has probability of $\frac{4}{13}$ which is larger than other state with probability of $\frac{1}{13}$.

(b)

The position of the curve of $\epsilon = 0$ is lower than the curve with ϵ, which means the agent will get lower reward if there is no exploration. But if we keep increase ϵ, the reward will decrease due to too much exploration. That means the reward vs ϵ should be a concave function. Because if ϵ is large, even the agent know what

is the optimal action at a given state, the agent will still do the bad action with high probability to explore. So high ϵ will decrease the reward.

(c)

When α is too low, the agent will learn very slowly so the value of each state could be lower than the true value. We can see from the plot that the curve with $\alpha = 0.1$ has the lowest value for most of states compare to other curves, that means the values have not get to the true value and need more iterations to learn. Also this curve is smoother, which means the computed values do not reflect the difference of values of all states.

When α is too high, the agent will learn very quickly and the computed values may be overshoot by each update, which will result a large deviation to the true values. The true values for two nearby states should be closed for most of the cases. But we can see from the plot that the curve with $\alpha = 1$ has most fluctuations compare to other curves, that means the values is overshoot and thus not converge to the true values.