DDA 4230 Assignment 2

Name: Xiang Fei ; ID: 120090414

Problem 1.

1. $Q_1(S_1, a_1) = 8 + 0.2 \times 10 + 0.6 \times 1 = 10.6$

$Q_1(S_1, a_2) = 10 + 0.1 \times 10 + 0.2 \times 1 = 11.2$

$V_1(S_1) = \max(10.6, 11.2) = 11.2$

$\pi_1(S_1) = a_2$

$Q_1(S_2, a_1) = 1 + 0.3 \times 10 + 0.3 \times 1 = 4.3$

$Q_1(S_2, a_2) = -1 + 0.5 \times 10 + 0.3 \times 1 = 4.3$

$V_1(S_2) = \max(4.3, 4.3) = 4.3$

$\pi_1(S_2) = a_1$

$Q_2(S_1, a_1) = 8 + 0.2 \times 11.2 + 0.6 \times 4.3 = 12.82$

$Q_2(S_1, a_2) = 10 + 0.1 \times 11.2 + 0.2 \times 4.3 = 11.98$

$V_2(S_1) = \max(12.82, 11.98) = 12.82$

$\pi_2(S_1) = a_1$

$Q_2(S_2, a_1) = 1 + 0.3 \times 11.2 + 0.3 \times 4.3 = 5.65$

$Q_2(S_2, a_2) = -1 + 0.5 \times 11.2 + 0.3 \times 4.3 = 5.89$

$V_2(S_2) = \max(5.65, 5.89) = 5.89$

$\pi_2(S_2) = a_2$

2. $Q_k(S_1, a_1) - Q_k(S_1, a_2) = 8 - 10 + (0.2 - 0.1) \times V_{k-1}(S_1) + (0.6 - 0.2) \times V_{k-1}(S_2)$

$= -2 + 0.1 \times V_{k-1}(S_1) + 0.4 \times V_{k-1}(S_2) \quad , \forall k \geq 1$

$Q_k(S_2, a_2) - Q_k(S_2, a_1) = -1 - 1 + (0.5 - 0.3) \times V_{k-1}(S_1) + (0.3 - 0.3) \times V_{k-1}(S_2)$

$= -2 + 0.2 \times V_{k-1}(S_1) \quad , \forall k \geq 1$

We have $V_{k-1}(S_1) \geq 12.82$, $V_{k-1}(S_2) \geq 5.89$

$\Rightarrow Q_k(S_1, a_1) - Q_k(S_1, a_2) > 0 \quad , Q_k(S_2, a_2) - Q_k(S_2, a_1) > 0$

$\Rightarrow Q_k(S_1, a_1) > Q_k(S_1, a_2) , \quad Q_k(S_2, a_2) > Q_k(S_2, a_1)$

$\Rightarrow \pi_k(S_1) = a_1 = \pi_2(S_1) , \; \pi_k(S_2) = a_2 = \pi_2(S_2)$

$\Rightarrow \pi_k(S) = \pi_2(S) \quad$, Q.E.D.

Problem 2.

1. see the code file.

2. (a)

```
Value Iteration:
V= [ 49.68634184  55.28325417  61.58053188  65.87810383  48.02738914
  52.31668307  68.14365569  73.25636944  50.22946871  -0.42050045
  77.06735759  81.36387215  66.3637945   76.31487478 100.
  89.90594114   0.        ] ,epsilon= 0.008649071085109483 ,nIterations= 30
```

(b)

```
Policy Iteration:
policy= [3 3 1 1 3 0 1 1 1 3 3 1 3 3 0 2 0] ,V= [ 49.6874368   55.28459708  61.5816329   65.87870774  48.02955063
  52.31901654  68.14444944  73.25667365  50.22994191  -0.4197238
  77.06760787  81.36396041  66.36413579  76.31508189 100.
  89.90596379   0.        ] ,nIterations= 4
```

(c) in partial policy: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

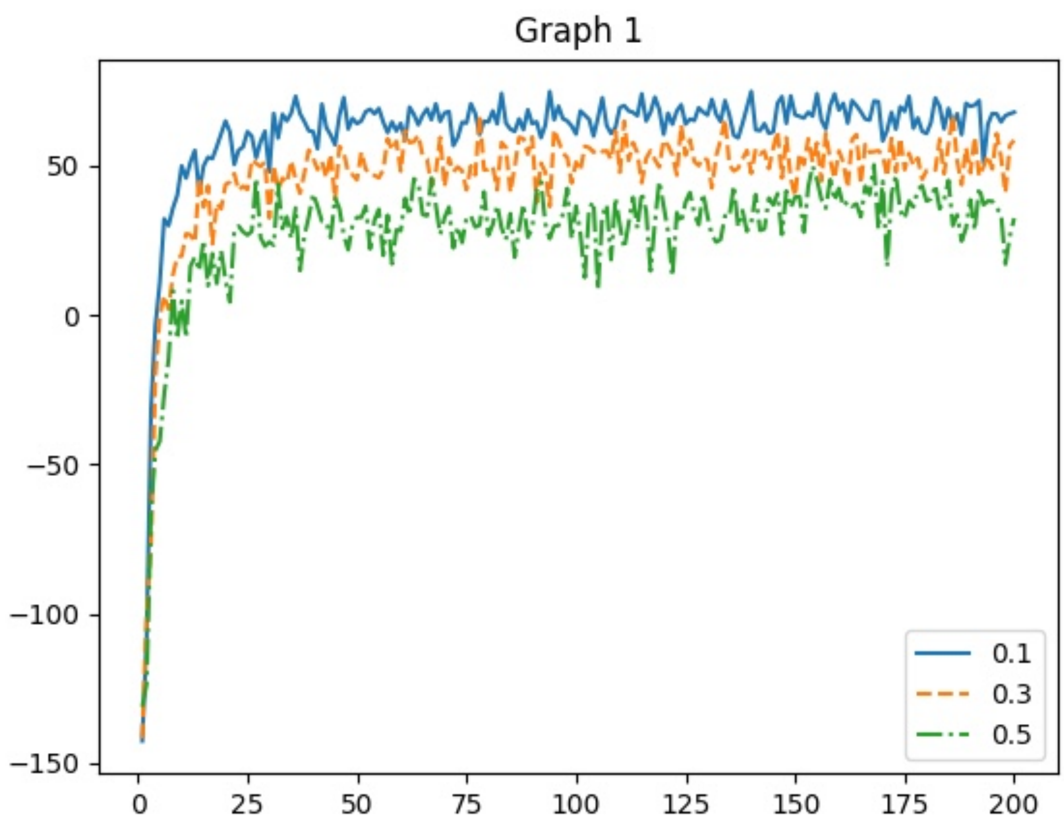
number of iterations: 17, 12, 10, 9, 8, 8, 7, 8, 8, 7

when the number of iterations in partial policy evaluation increases, the result decrease, and finally converge / stable.

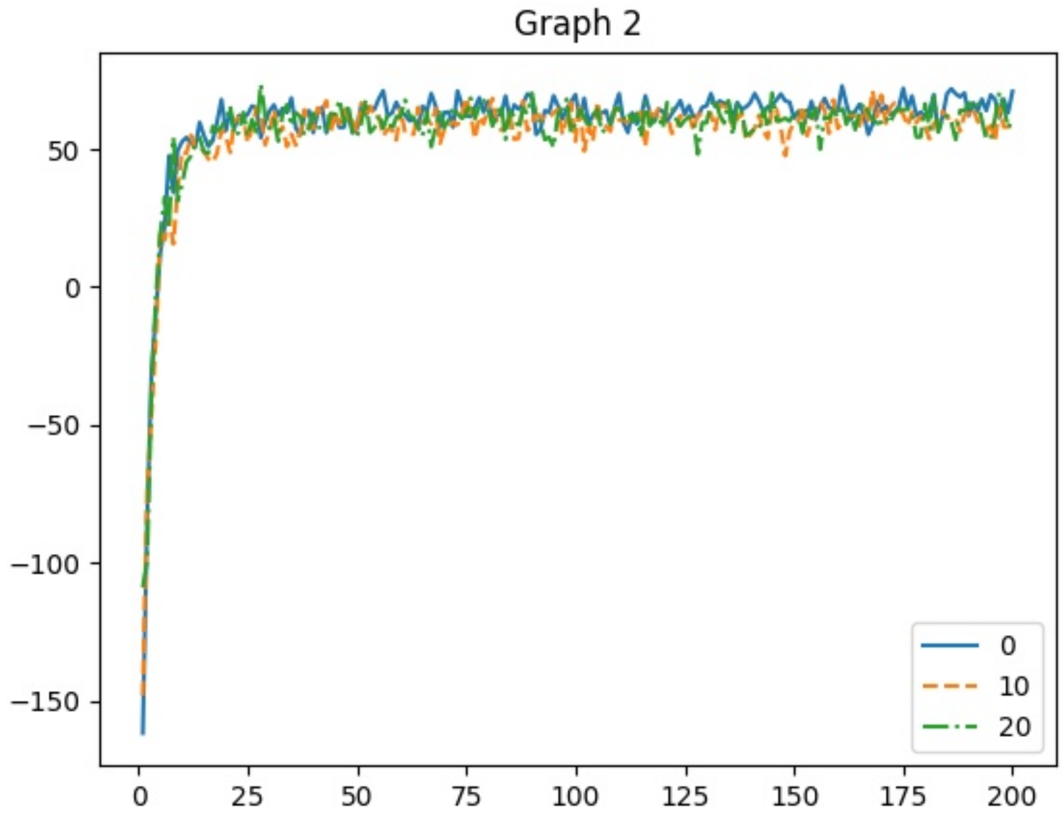
Problem 3.

1. see the code file.

2.



3.



4. when the exploration probability epsilon increases, the cumulative discounted rewards per episode earned decreases. The Boltzmann temperature does not have remarkable impact on the rewards.