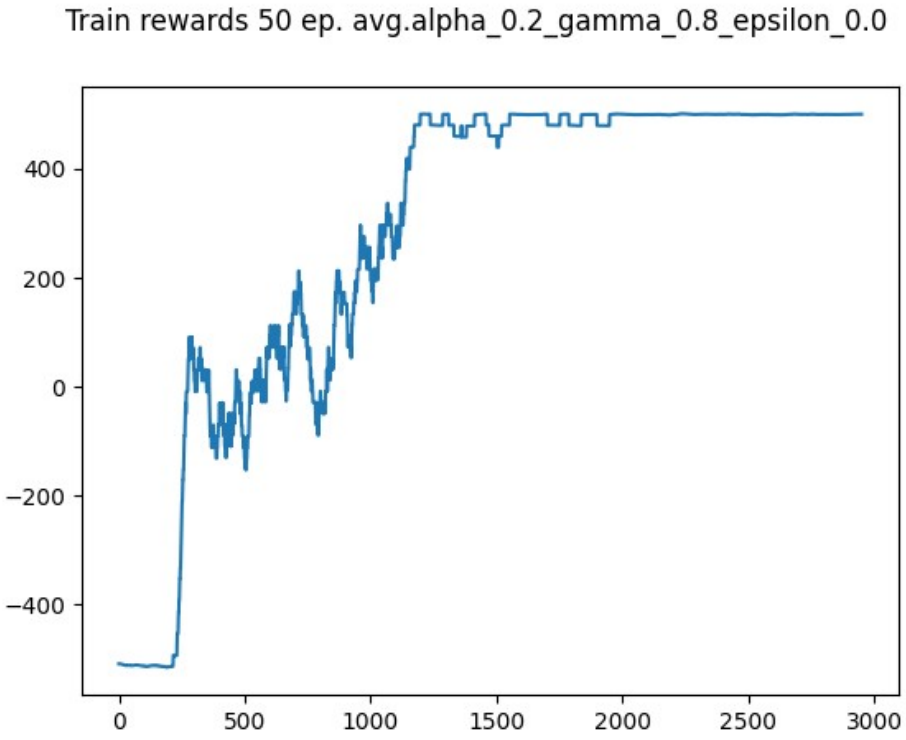
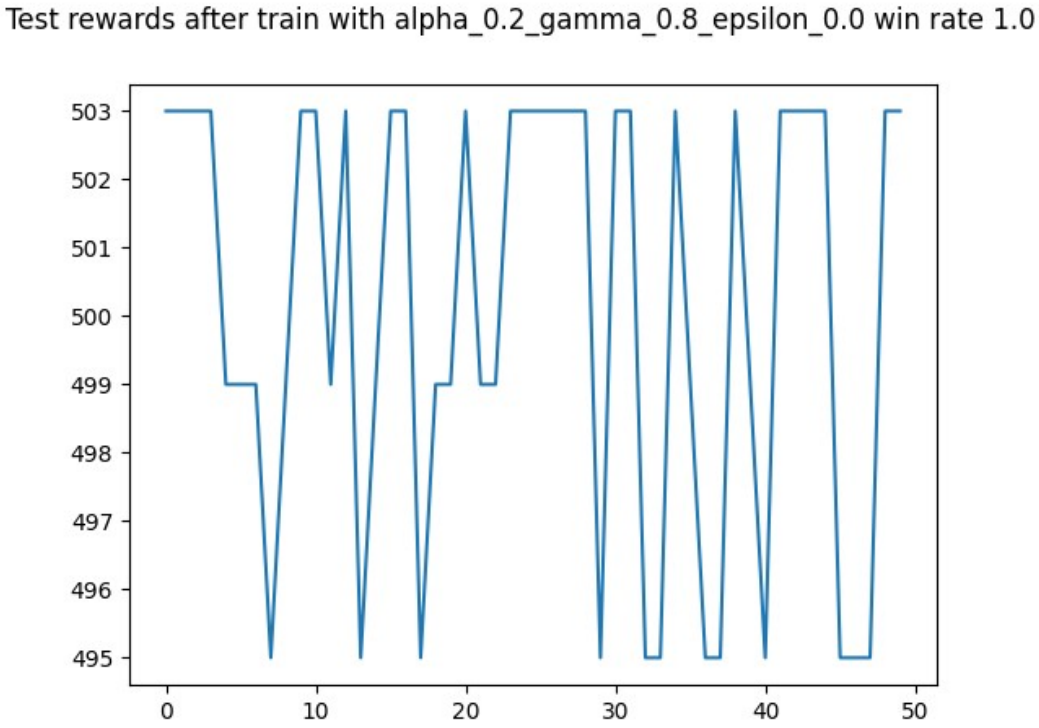
RL Programming assignment 1 – Tabular Q learning – Moshe Beutel 037580792

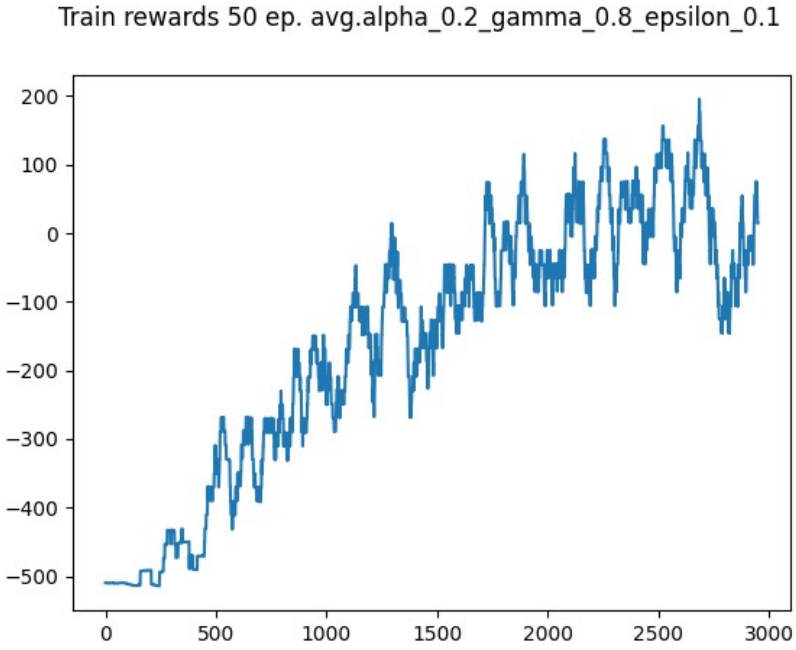
Questions 3+4

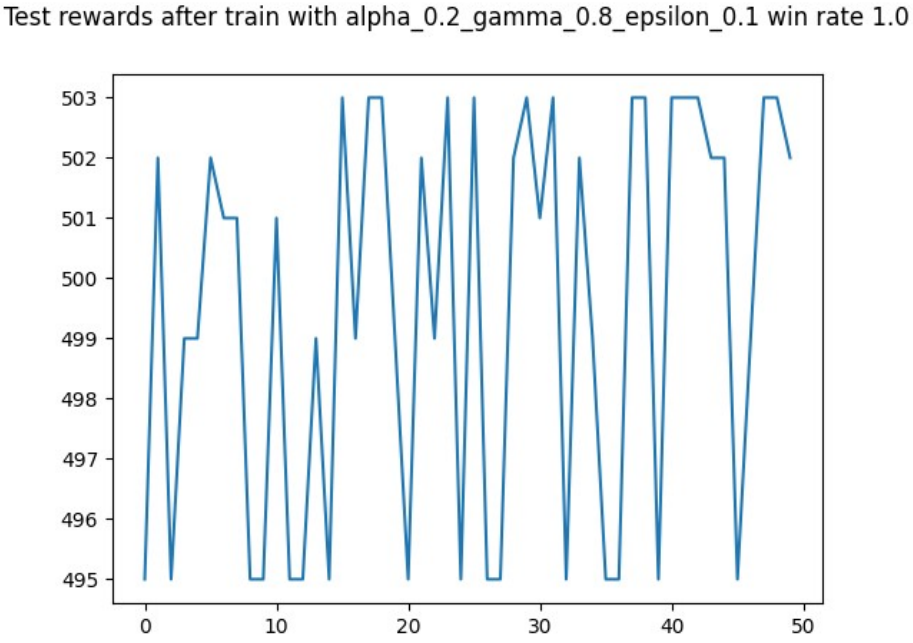
epsilon = 0.0



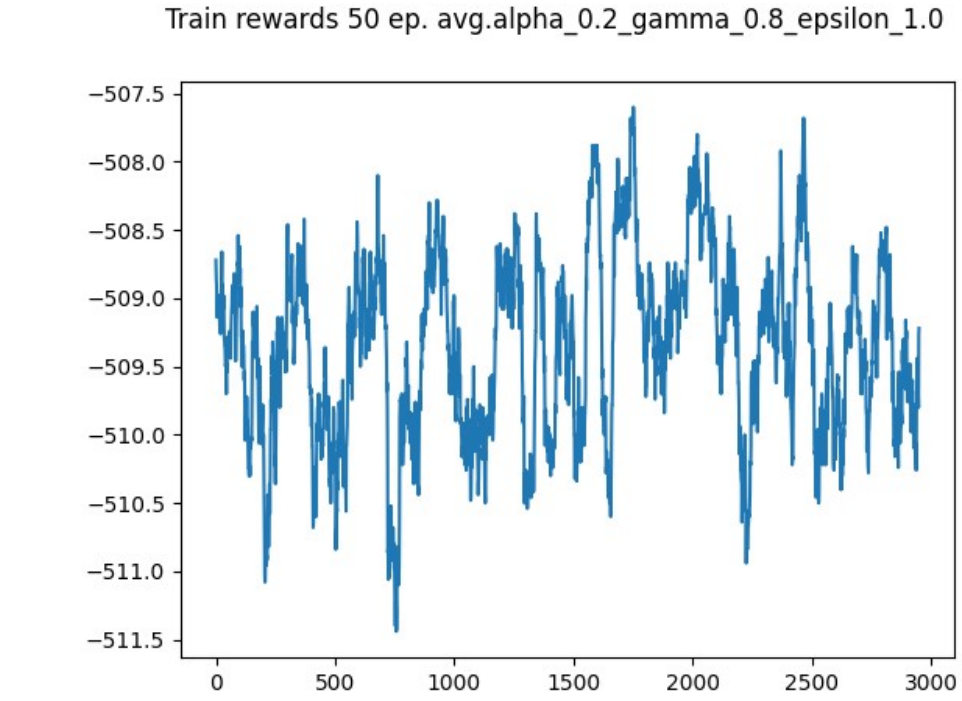


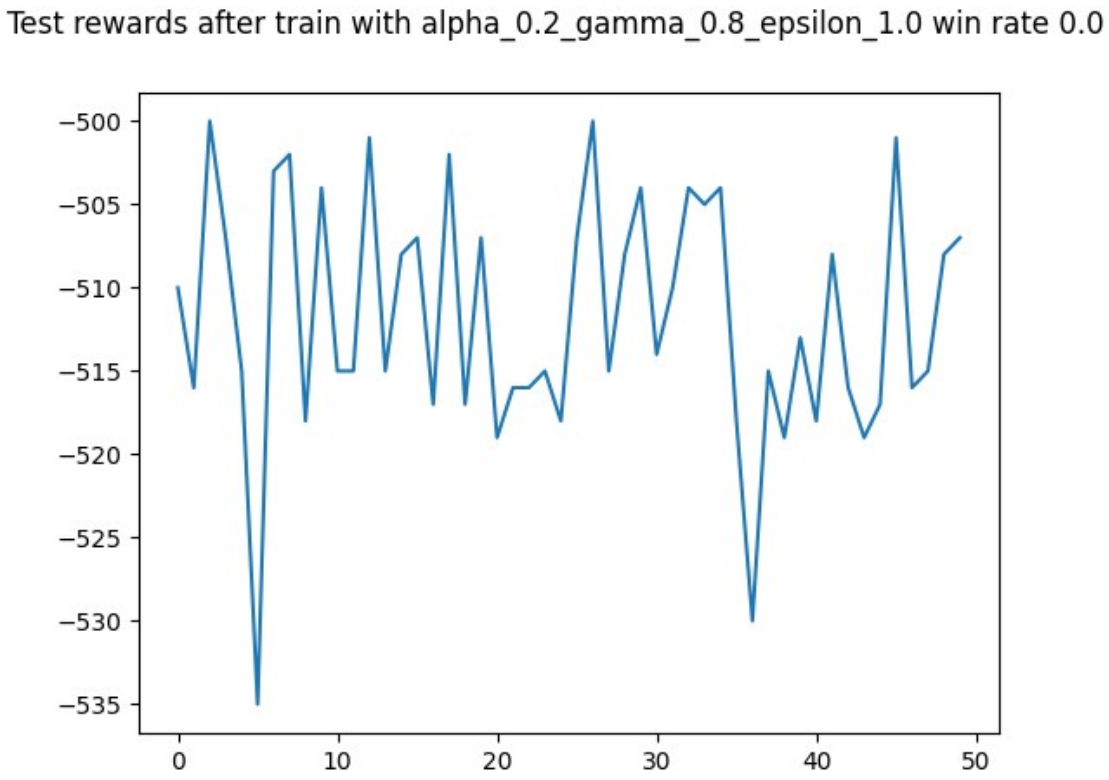
epsilon=0.1


Train rewards 50 ep. avg.alpha_0.2_gamma_0.8_epsilon_0.1
200
100
0
−100
−200
−300
−400
−500
0
500
1000
1500
2000
2500
3000

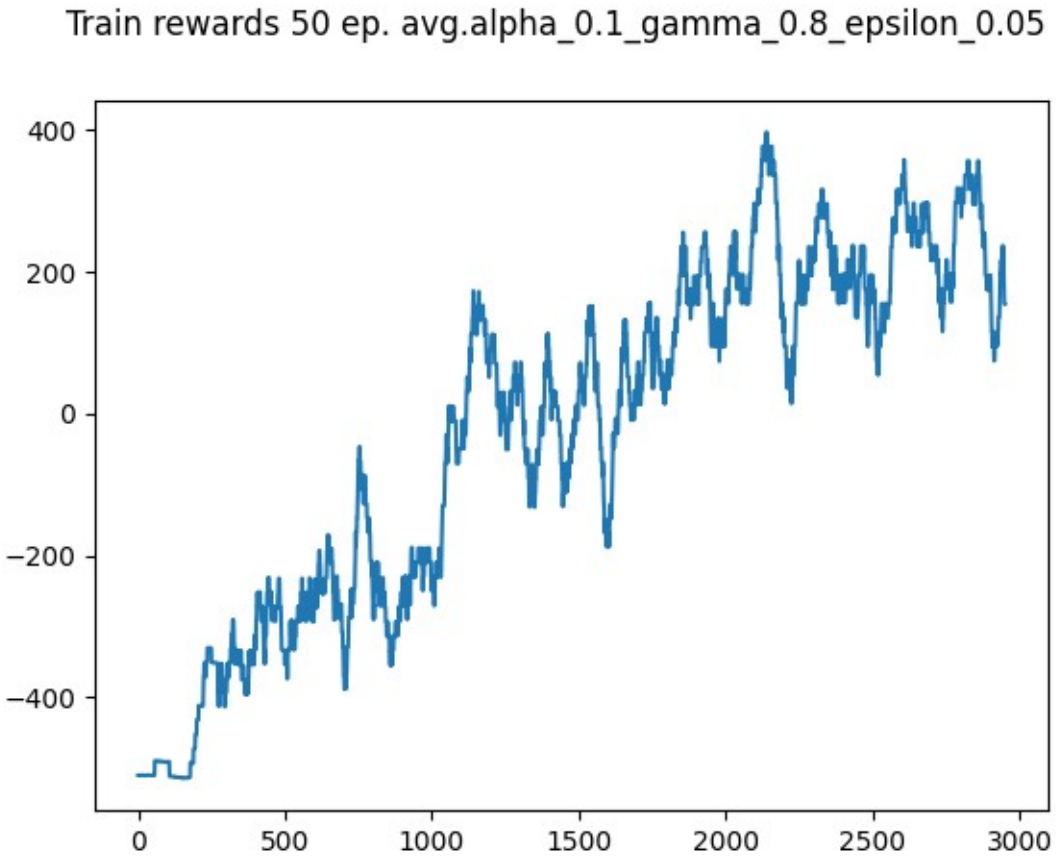

Test rewards after train with alpha_0.2_gamma_0.8_epsilon_0.1 win rate 1.0
503
502
501
500
499
498
497
496
495
0
10
20
30
40
50

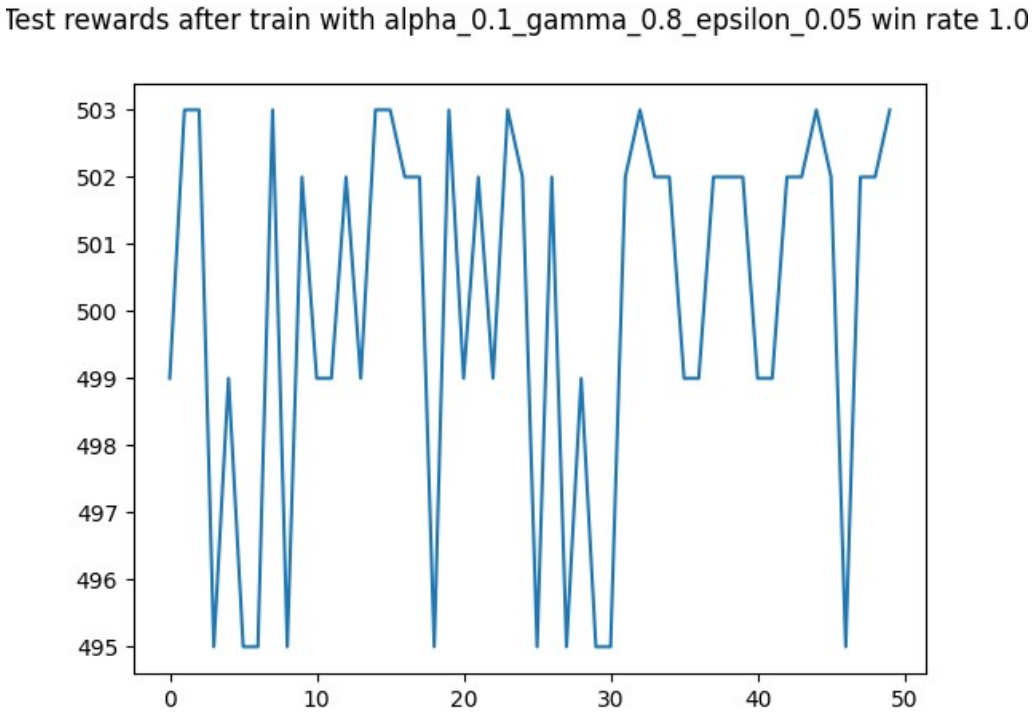
epsilon=1.0


Train rewards 50 ep. avg.alpha_0.2_gamma_0.8_epsilon_1.0
−507.5
−508.0
−508.5
−509.0
−509.5
−510.0
−510.5
−511.0
−511.5
0
500
1000
1500
2000
2500
3000

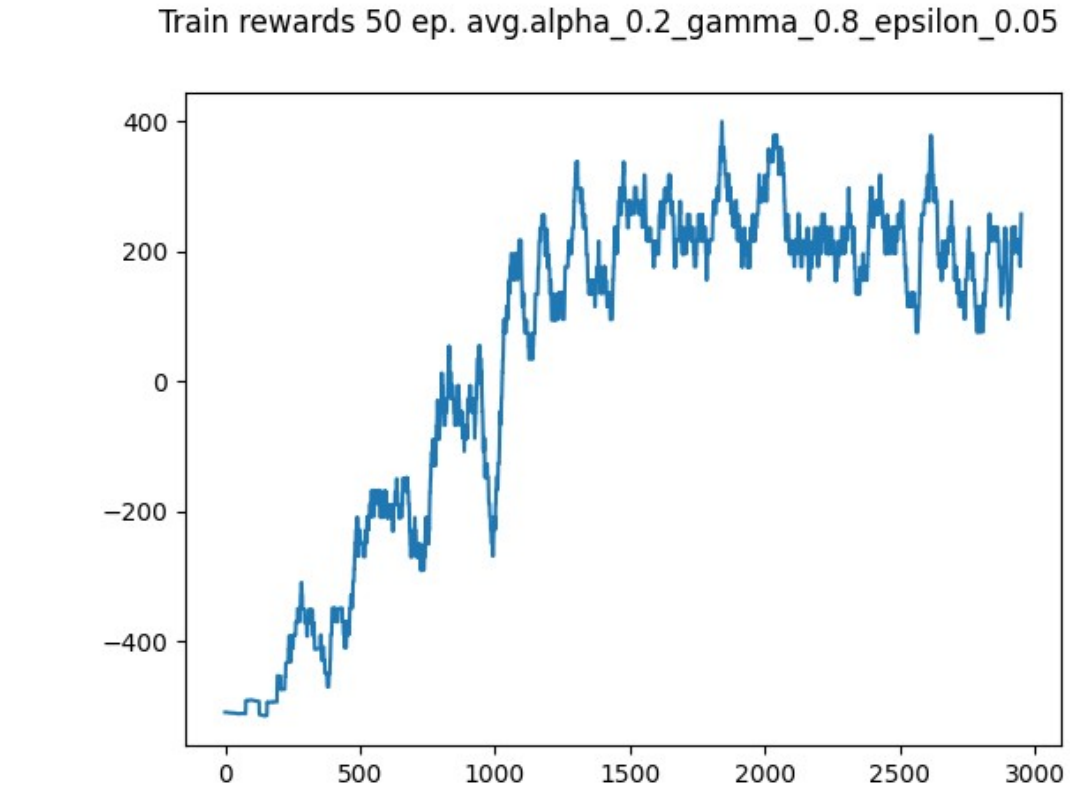

Test rewards after train with alpha_0.2_gamma_0.8_epsilon_1.0 win rate 0.0
−500
−505
−510
−515
−520
−525
−530
−535
0
10
20
30
40
50

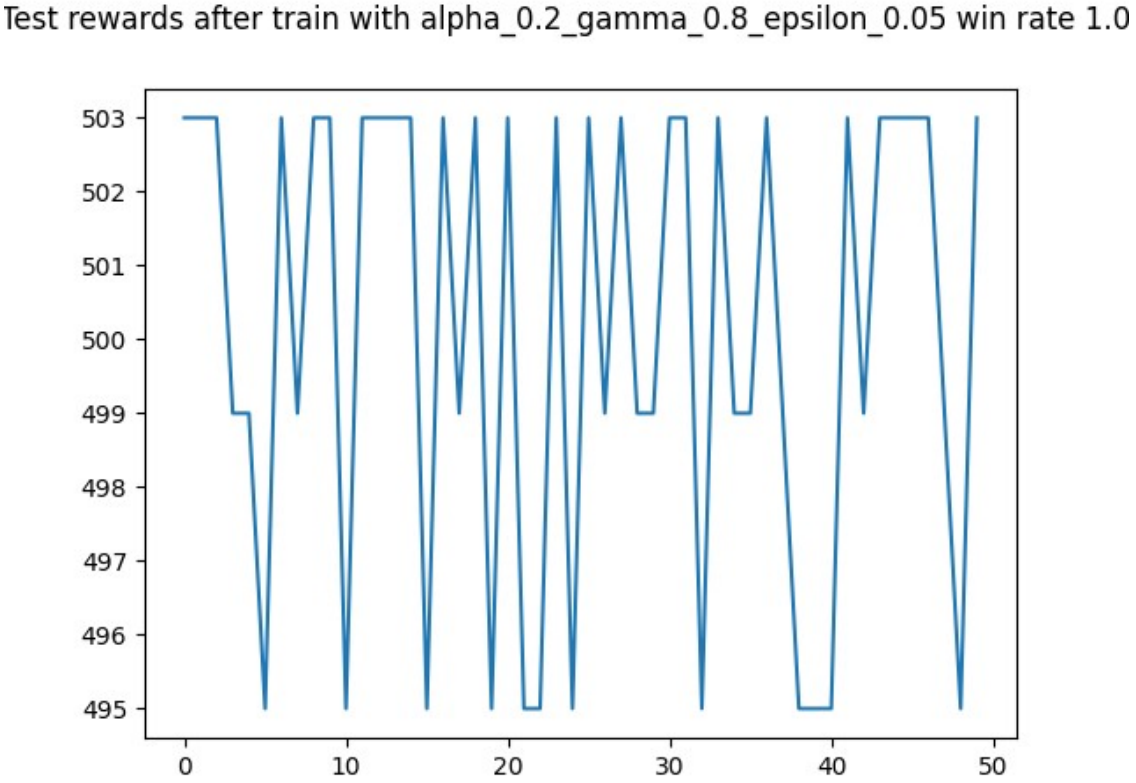
alpha=0.1


Train rewards 50 ep. avg.alpha_0.1_gamma_0.8_epsilon_0.05
400
200
0
−200
−400
0
500
1000
1500
2000
2500
3000

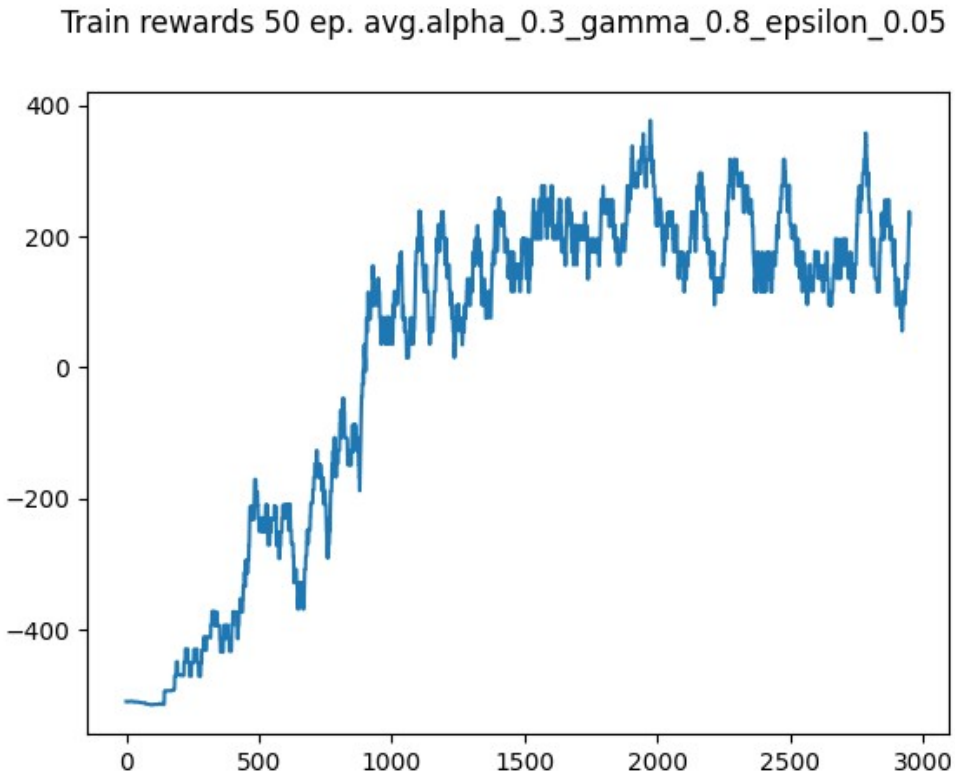

Test rewards after train with alpha_0.1_gamma_0.8_epsilon_0.05 win rate 1.0
503
502
501
500
499
498
497
496
495
0
10
20
30
40
50

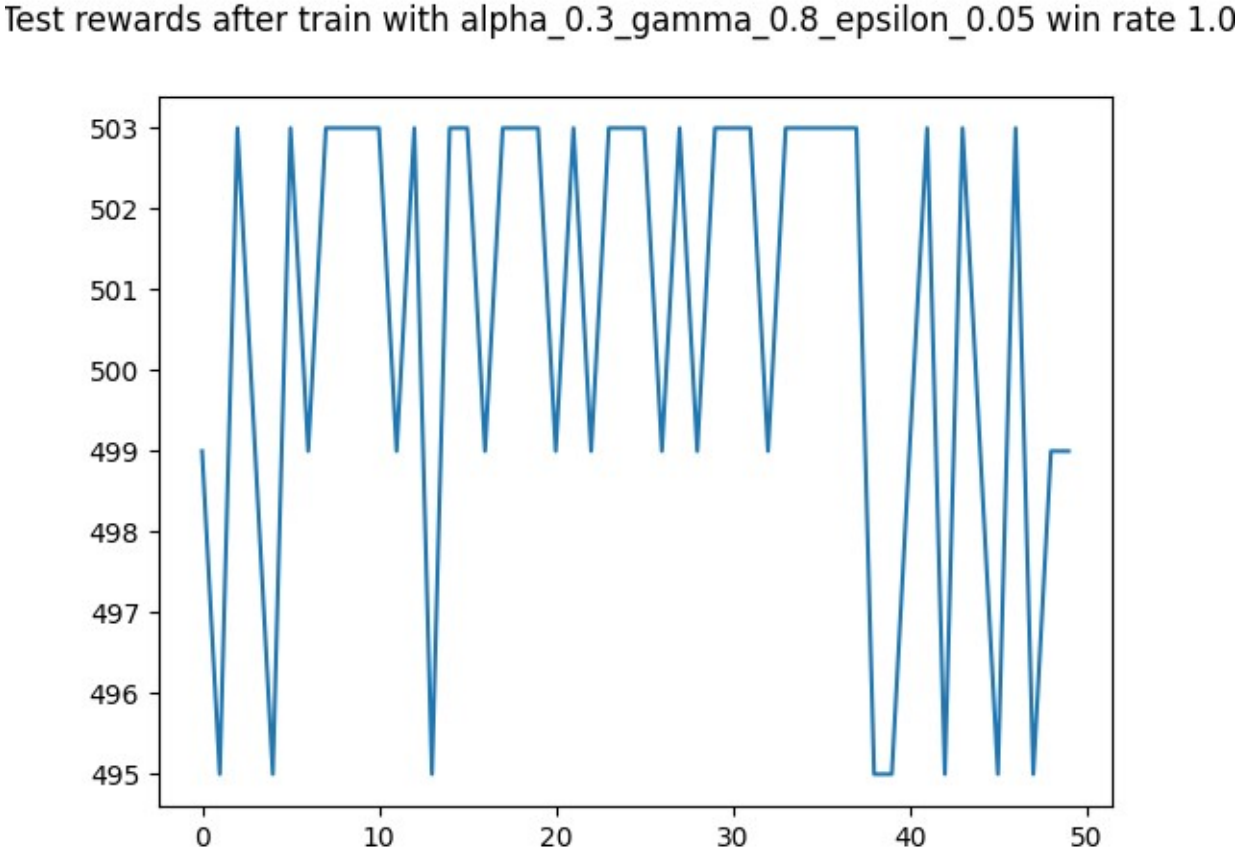
alpha=0.2


Train rewards 50 ep. avg.alpha_0.2_gamma_0.8_epsilon_0.05
400
200
0
−200
−400
0
500
1000
1500
2000
2500
3000


Test rewards after train with alpha_0.2_gamma_0.8_epsilon_0.05 win rate 1.0
503
502
501
500
499
498
497
496
495
0
10
20
30
40
50

alpha=0.3


Train rewards 50 ep. avg.alpha_0.3_gamma_0.8_epsilon_0.05
400
200
0
−200
−400
0
500
1000
1500
2000
2500
3000


Test rewards after train with alpha_0.3_gamma_0.8_epsilon_0.05 win rate 1.0
503
502
501
500
499
498
497
496
495
0
10
20
30
40
50

Question 5. The agent with no exploration works well because of the negative rewards. The Q table is initialized by default to 0 (I even initialized with 500 toincrease exploration) so the max in the q update encourage the agent to try those 0 valued unseen state-actions.

Questions 6+7. The results on the medium layout were about 88%. The number of unseen states (or seen just a few times) is very large:

****

518299 states in Q table
110509 states were visited less than 5 times
75590 states were visited once

****

I made a few changes that rased the test average reward 100%:
When storing the state I zeroed the score and the direction (direction.STOP). The states that only differ in score or direction were unified
This made the q-table a lot smaller and learning more efficient.

****

9148 states in Q table
1357 states were visited less than 5 times
847 states were visited once

****