Assignment 2

Kevin McCoy

1. Policy Evaluation 20pts

1.1 5pts

In the class, we derived the following relation between the Q_π and V_π values of a policy π

$$Q_\pi(s,a) = R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V_\pi(s') \tag{1}$$

$$= \sum_{s' \in S} T(s'|s,a) R(s,a,s') + \gamma T(s'|s,a) V_\pi(s') \tag{2}$$

$$= \sum_{s' \in S} T(s'|s,a) \{R_d(s,a,s') + \gamma V_\pi(s')\} \tag{3}$$

as $\sum_{s' \in S} T(s'|s,a) R(s,a,s') = R(s,a)$.

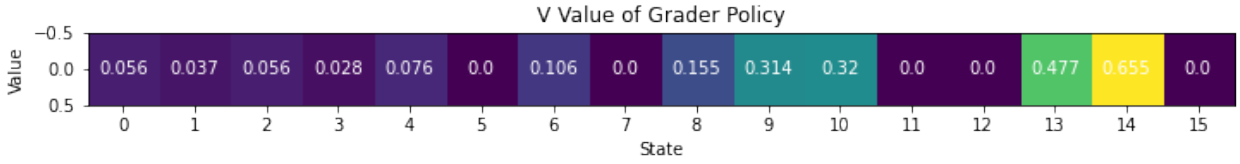
If the reward signal is a deterministic function (denoted by $R_d(s')$) of only the next state, we can re-write the above as:

$$Q_\pi(s,a) = \sum_{s' \in S} T(s'|s,a) \{R_d(s') + \gamma V_\pi(s')\} \tag{4}$$

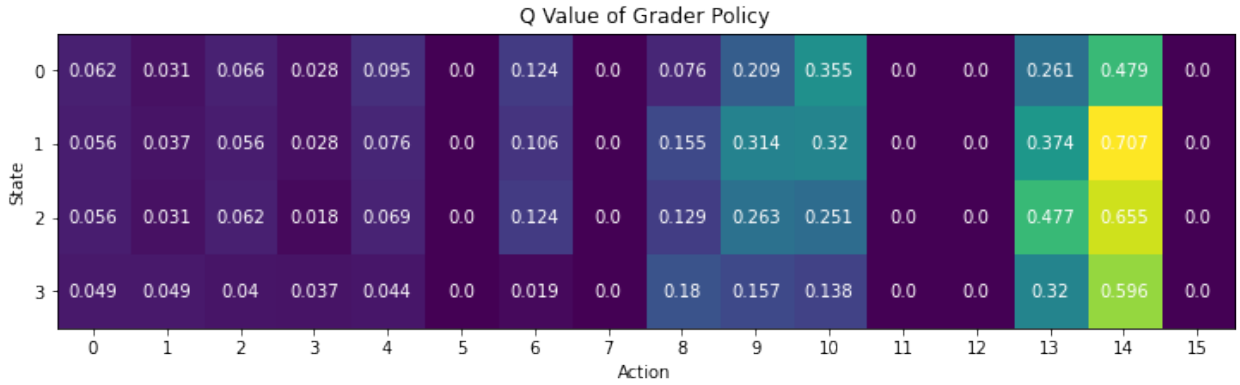
$$= \sum_{s' \in S} T(s'|s,a) V_\pi(s) \tag{5}$$

$$\tag{6}$$

1.2 10pts

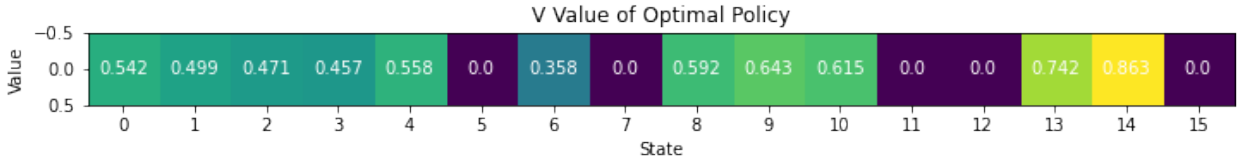


1.3 5pts

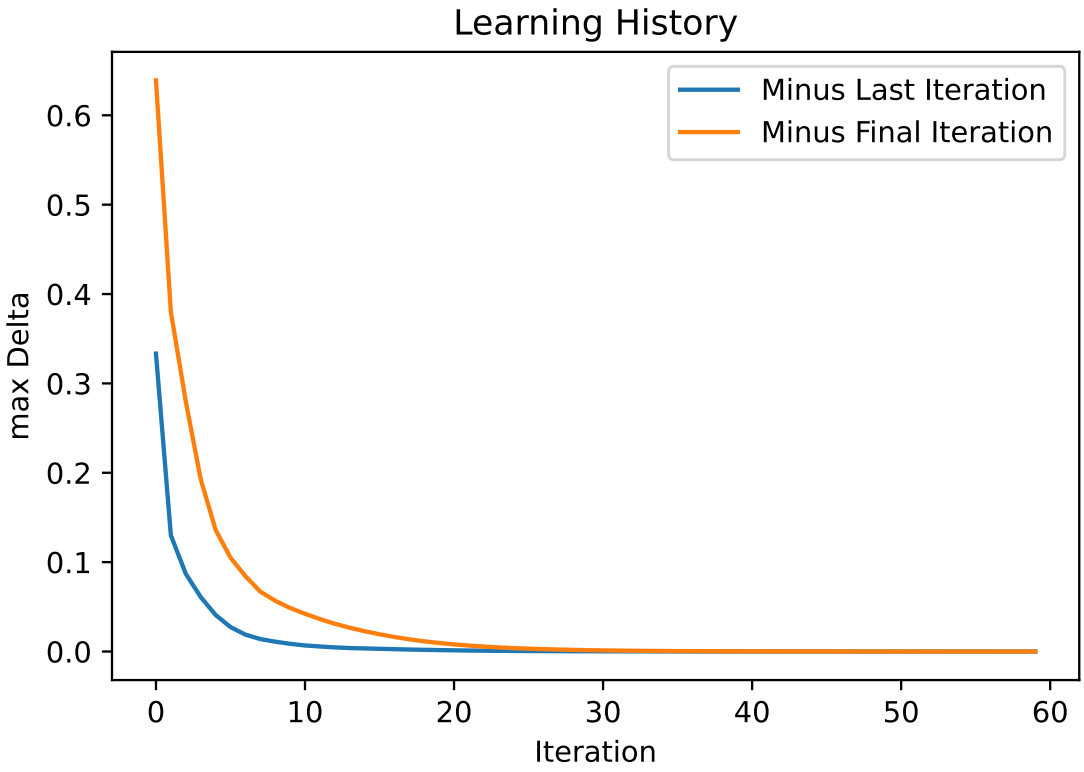


2. Value Iteration 20pts

2.1 10pts



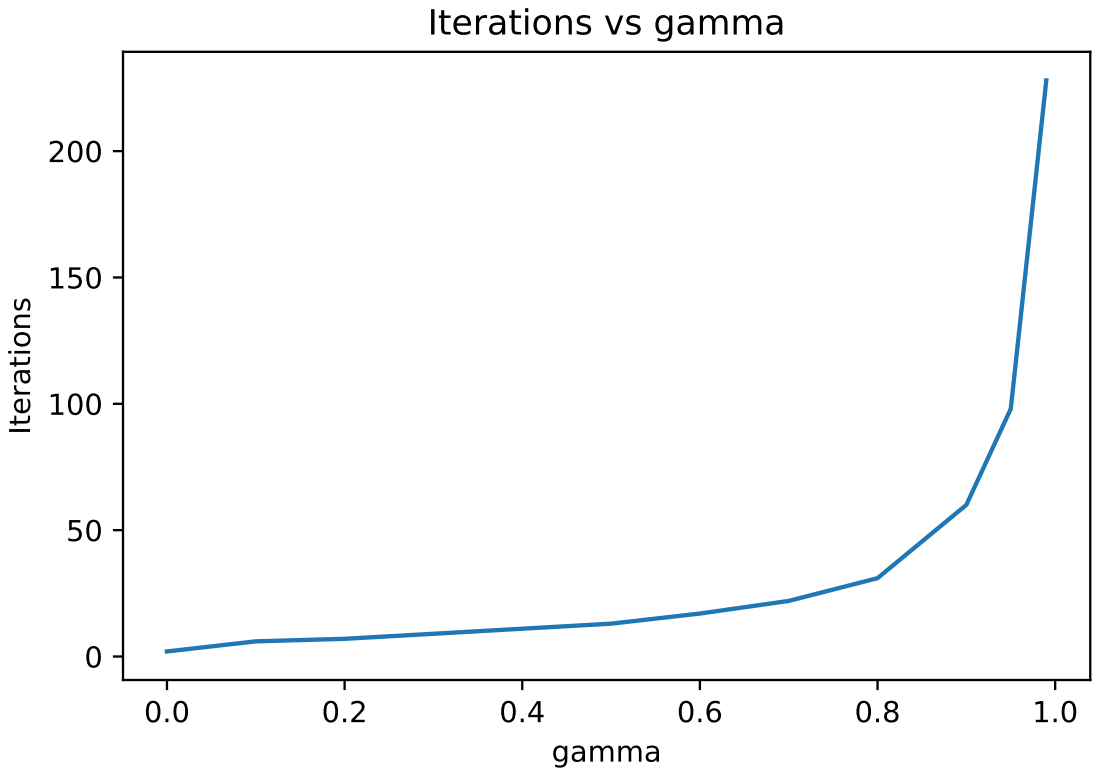
2.2 3pts



2.3 2pts

The algorithm converges after 60 iterations. I am fairly certain that the value iteration algorithm has found the optimal policy, as it is guaranteed convergence to such a policy even though policy evaluation is only conducted once before policy improvement.

2.4 **3pts**



The number of iterations is positively correlated with γ because lower values of gamma correspond with a more greedy value evaluation. In other words, the value of a state will depend much more on its immediate reward and depend less on the rewards of possible future states. This independence of values makes it faster for the algorithm to converge, but will ultimately converge to a greedy value estimation.

2.5 **2pts**

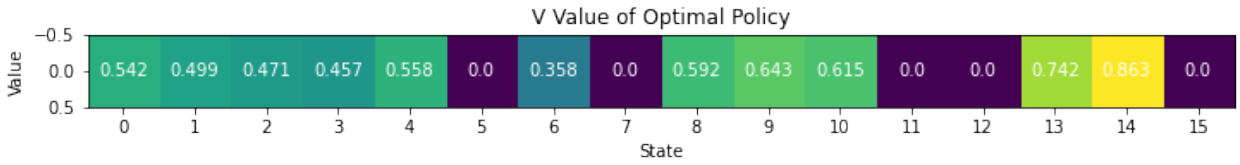
```
Without guess policy:
Number of iterations = 228
With guess policy:
Number of iterations = 155
```

3. Policy Iteration 20pts

3.1 5pts

$$|\pi| = |A|^{|S|} = 4^{16} = 2^{32} = 4,294,967,296$$

3.2 12pts



3.3 3pts

The policy iteration only looped 7 times before obtaining the optimal policy. This is obviously much lower than the brute force approach. This is because not every deterministic policy really needs to be tested. Instead, by alternating between policy evaluation and improvement, we make continual strides toward the correct solution.

4. Monte Carlo Control **20pts**

(a) Average reward of -1.0

(b) Average reward -0.098

Figure 1: Grader Code Output

5. Temporal Difference Learning 20pts

(a) Average reward of -1.0

(b) Average reward of -0.112

Figure 2: Grader Code Output