# Homework 4


## Tyler Amos


6 June 2019

# 1a

# 1b

# 1c

# 2a

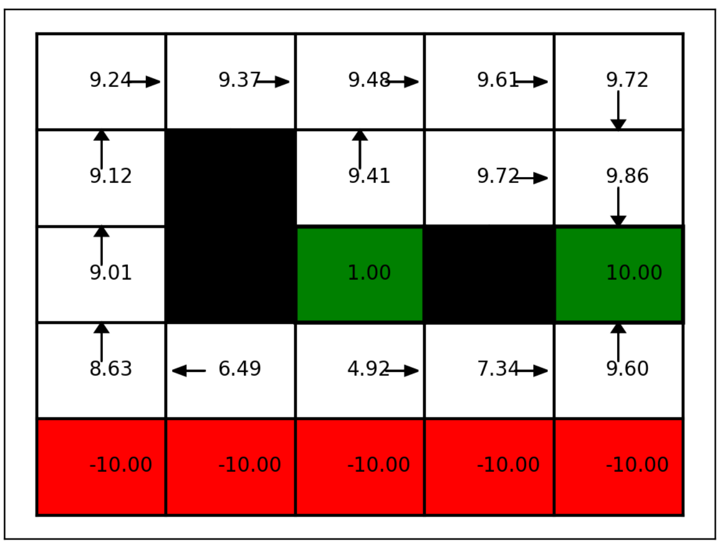


Figure 1: Optimal Value Function and Policy for 2a

## 2b

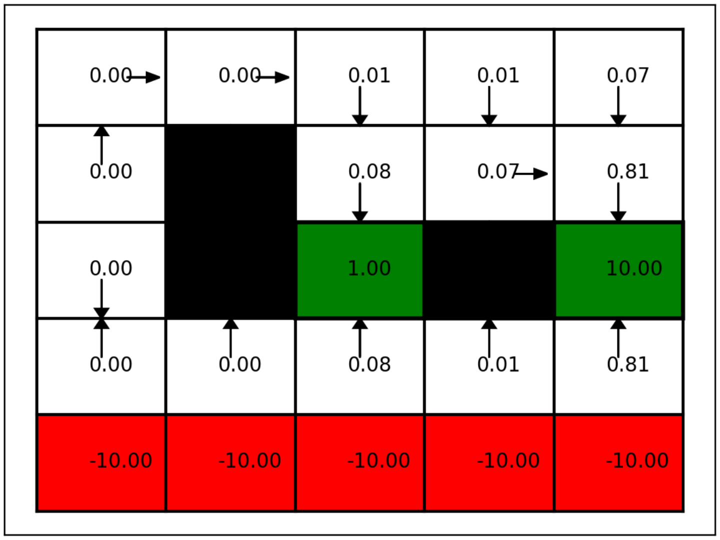


Figure 2: Optimal Value Function and Policy for 2b,i

These two policies differ in their discounting factor. The first (i) is highly myopic, while the second (ii) is very farsighted. This means under the first policy, the agent will prioritize short-term rewards. This leads to, for example, the cautious policy in the second row from the bottom. In all of these cells, the agent pursues the policy of moving upwards. By always going upward only, the agent avoids catastrophic outcomes in the bottom row.

By contrast, in the second policy the agent engages in more 'risk-taking' behaviour by moving left and right (east/west) in the second row. This runs the risk of entering the cliff in the bottom row. However, if the agent does not end up going off the cliff, the potential rewards are greater.

I find these results to be quite interesting because my initial intuition was actually quite the opposite. I thought an agent which values their future more (i.e., is more farsighted) would always try to move upward in the bottom row - this guarantees they will not fall off the cliff. This would be a very risk-averse strategy. It is interesting then that the farsighted agent may pursue a riskier strategy than the myopic agent.

At the same time, the logic of the agents' planning choices are clear. The myopia

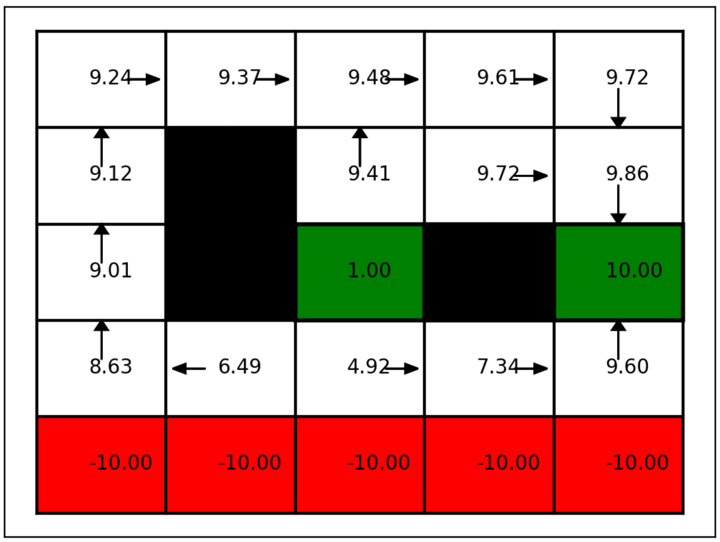


Figure 3: Optimal Value Function and Policy for 2b,ii

of the first agent likely leads it to ‘worry’ excessively about the risk posed by the cliff in the bottom row. This leads it to adopt a very cautious strategy of only moving upwards, which guarantees it will never fall over the cliff. The more farsighted agent still accounts for the risks of the cliff, but tries to benefit from the rewards in the other parts of the environment.

## 2c

It would appear more farsighted agents require more iterations to converge upon a policy (28) than more myopic agents (5).

Agents who have a greater amount of noise also appear to take longer to converge (58) than those with less noise (28). This is logical as estimates of potential rewards which are less certain (i.e., more noisy) will take longer to converge to a global optimum.

## 2d

From 14.5 in Probabilistic Robotics, note the payoff can be written as the sum of a geometric series, so the difference between the current payoff and the optimal payoff can be written as:

$$\Sigma_{i=0}^{inf} ||r_{current_i} - r_{max_i}||$$

Now we know that sequential differences between value functions are less than $\epsilon$, so we can reformulate this as:

$$\Sigma_{i=0}^{inf} \epsilon_\delta$$

Where $\epsilon_\delta$ is a difference of $\epsilon$, or equivalently $2\epsilon$.

Since $\epsilon$ is the reward, or rather the improvement in the reward, it must also be discounted with a $\gamma$ term. We take the above summation and express it with the standard limit of a geometric series formula:

$$\frac{a}{1-r} = \frac{2\epsilon\gamma}{1-\gamma}$$

## 3

## 4a

I did not work with anyone on this problem set.

## 4b

I spent approximately 5 hours on this problem set.