AI Planning for Autonomy

Problem Set X: Reward shaping and policy-based approaches

1. Policy Iteration has two main steps, policy evaluation and policy update. In order to evaluate the given policy:

$$\begin{aligned}
V^\pi(Messi) &= Q^\pi(Messi, Pass) \\
&= P_{pass}(Suarez|Messi)[r(Messi, pass, Suarez) + \gamma \cdot V^\pi(Suarez)] \\
&= \gamma \cdot V^\pi(Suarez) - 1 \\
V^\pi(Suarez) &= Q^\pi(Suarez, Pass) \\
&= P_{pass}(Messi|Suarez)[r(Suarez, pass, Messi) + \gamma \cdot V^\pi(Messi)] \\
&= \gamma \cdot V^\pi(Messi) - 1 \\
V^\pi(Scored) &= Q^\pi(Scored, return) \\
&= P_{return}(Messi|Scored)[r(Scored, return, Messi) + \gamma \cdot V^\pi(Messi)] \\
&= \gamma \cdot V^\pi(Messi) + 2
\end{aligned}$$

   Then solve a very basic linear algebra about $V^\pi(Messi)$ and $V^\pi(Suarez)$:

$$\begin{aligned}
V^\pi(Messi) &= 1/(\gamma - 1) \\
V^\pi(Suarez) &= 1/(\gamma - 1) \\
V^\pi(Scored) &= 3 + 1/(\gamma - 1)
\end{aligned}$$

   Then apply $\gamma = 0.8$, the policy evaluation table would be:

Iter	$Q^\pi(Messi, P)$	$Q^\pi(Messi, S)$	$Q^\pi(Suarez, P)$	$Q^\pi(Suarez, S)$	$Q^\pi(Scored)$
0	0	0	0	0	0
1	-5	-5.52	-5	-4.56	-2
2	-4.194	-4.772	-4.355	-3.993	-1.355

   Then implement two iteration of policy update based on value from the policy evaluation table:

Iter	$\pi(Messi)$	$\pi(Suarez)$	$\pi(Scored)$
0	Pass	Pass	Return
1	Pass	Shoot	Return
2	Pass	Shoot	Return

2. The important thing for the reward function is that you need to consider the next goal and whether the key is held. Using normalised Manhattan distance as the estimate, we can define the following potential function:

```
if Key == 0:
   return 1 - NormalizedManhattan(s, K)
elif Key == 1 and M == False:
   return 1 - NormalizedManhattan(s, M)
elif Key == 1 and M == True:
   return 1 - NormalizedManhattan(s, R)
elif Key == 2:
   return 1 - NormalizedManhattan(s, R)
```

   Others are possible, but this one will help to guide the agent early in the exploration.

3. Assuming that the Manhattan function is normalise, we calculate Φ for the following:

Let us s be the current state, s_1 be the state after action Up and s_2 be the state after action $Right$:

$$\begin{array}{rcl} s & = & ((4,0), Key = 1, M = False) \\ s_1 & = & ((4,1), Key = 1, M = False) \\ s_2 & = & ((5,0), Key = 1, M = False) \end{array}$$

$$\begin{array}{rcccl} \Phi(s) & = & 1 - \frac{9}{12} & = & \frac{3}{12} \\ \Phi(s_1) & = & 1 - \frac{8}{12} & = & \frac{4}{12} \\ \Phi(s_2) & = & 1 - \frac{10}{12} & = & \frac{2}{12} \end{array}$$

To update the Up action:

$$\begin{array}{rcl} Q(s, Up) & \leftarrow & Q(s, Up) + \alpha[r(s, Up, s_1) + F(s, s_1) + \gamma \max Q(s_1, a') - Q(s, Up)] \\ & \leftarrow & 0 + 0.2 \times [0 + 0.9 \times \frac{4}{12} - \frac{3}{12} + 0.9 \times 0 - 0] \\ & \leftarrow & 0.01 \end{array}$$

To update the $Right$ action:

$$\begin{array}{rcl} Q(s, Right) & \leftarrow & Q(s, Right) + \alpha[r(s, Right, s_1) + F(s, s_2) + \gamma \max Q(s_2, a') - Q(s, Right)] \\ & \leftarrow & 0 + 0.2 \times [0 + 0.9 \times \frac{2}{12} - \frac{3}{12} + 0.9 \times 0 - 0] \\ & \leftarrow & -0.02 \end{array}$$