CS182 Assignment 4

Charles Liu
cliu02@g.harvard.edu

October 29, 2015

1: Model this problem as a Markov Decision Process: Formally specify the states, actions, transition function and reward function.

State	cool, warm, off
Actions	fast, slow
Transition	Pr(cool | cool, slow) = 1 Pr(cool | cool, fast) = 1/4 Pr(warm | cool, fast) = 3/4 Pr(cool | warm, slow) = 1/4 Pr(warm | warm, slow) = 3/4 Pr(warm | warm, fast) = 7/8 Pr(off | warm, fast) = 1/8
Reward	R(cool, slow, cool) = 4 R(cool, fast, cool) = 10 R(cool, fast, warm) = 10 R(warm, slow, cool) = 4 R(warm, slow, warm) = 4 R(warm, fast, warm) = 10 R(warm, fast, off) = 0

2: Write down the U function for this problem in all possible states. Calculate by hand the values of U assuming a policy Π_1 when the rover always drives fast and Π_2 when the rover always drives slow.

$$
\begin{aligned}
U(cool) &= max(\begin{cases} R(cool, slow, cool) + \gamma U(cool) \\ \frac{1}{4}(R(cool, fast, cool) + \gamma U(cool)) + \frac{3}{4}(R(cool, fast, warm) + \gamma U(warm)) \end{cases}) \\
&= max(\begin{cases} 4 + \gamma U(cool) \\ 10 + \frac{1}{4}\gamma U(cool) + \frac{3}{4}\gamma U(warm) \end{cases}) \\
U(warm) &= max(\begin{cases} \frac{1}{4}(R(warm, slow, cool) + \gamma U(cool)) + \frac{3}{4}(R(warm, slow, warm) + \gamma U(warm)) \\ \frac{7}{8}(R(warm, fast, warm) + \gamma U(warm)) + \frac{1}{8}(R(warm, fast, off) + \gamma U(off)) \end{cases}) \\
&= max(\begin{cases} 4 + \frac{1}{4}\gamma U(cool) + \frac{3}{4}\gamma U(warm) \\ \frac{70}{8} + \frac{7}{8}\gamma U(warm) + \frac{1}{8}\gamma U(off) \end{cases}) \\
U(off) &= 0
\end{aligned}
$$

If we have the policy that we only drive fast, then U becomes:

$$\begin{aligned}
U(cool) &= 10 + \frac{1}{4}\gamma U(cool) + \frac{3}{4}\gamma U(warm) \\
U(warm) &= \frac{70}{8} + \frac{7}{8}\gamma U(warm) + \frac{1}{8}\gamma U(off) \\
U(off) &= 0
\end{aligned}$$

Solving this system of equations gives:

$$\begin{aligned}
(1 - \frac{7}{8}\gamma)U(warm) &= \frac{70}{8} \\
U(warm) &= \frac{70}{8 - 7\gamma} \\
(1 - \frac{1}{4}\gamma)U(cool) &= 10 + \frac{3}{4}\gamma(\frac{70}{8 - 7\gamma}) \\
&= 10 + \gamma\frac{105}{16 - 14\gamma} \\
U(cool) &= \frac{10 + \gamma\frac{105}{16-14\gamma}}{1 - \frac{1}{4}\gamma}
\end{aligned}$$

Likewise if we have the policy that we only drive slowly, then U becomes:

$$\begin{aligned}
U(cool) &= 4 + \gamma U(cool) \\
U(warm) &= 4 + \frac{1}{4}\gamma U(cool) + \frac{3}{4}\gamma U(warm) \\
U(off) &= 0
\end{aligned}$$

Solving this system of equations gives:

$$\begin{aligned}
(1 - \gamma)U(cool) &= 4 \\
U(cool) &= \frac{4}{1 - \gamma} \\
(1 - \frac{3}{4}\gamma)U(warm) &= 4 + \frac{\gamma}{1 - \gamma} \\
U(warm) &= \frac{4 + \frac{\gamma}{1-\gamma}}{(1 - \frac{3}{4}\gamma)}
\end{aligned}$$

3: Start with a policy where you drive fast no matter what the condition of the rover is. Simulate the first two iterations of the policy iteration algorithm. Show how the policy evolves as you run the algorithm. What is the policy after the second iteration? For this question assume a discount factor of 0.9.

Iteration 1:

$$
\begin{aligned}
U^0(warm) &= \frac{70}{8-7\gamma} \\
&= 41.1764705882 \\
U^0(cool) &= \frac{10+\gamma\frac{105}{16-14\gamma}}{1-\frac{1}{4}\gamma} \\
&= 48.7666034156 \\
U^0(off) &= 0 \\
\Pi^1(warm) &= max\_action(\begin{cases} slow \rightarrow \frac{1}{4}U^0(cool)+\frac{3}{4}U^0(warm) \\ fast \rightarrow \frac{7}{8}U^0(warm)+\frac{1}{8}U^0(off) \end{cases}) \\
&= max\_action(\begin{cases} slow \rightarrow 43.074003795 \\ fast \rightarrow 36.0294117647 \end{cases}) \\
&= slow \\
\Pi^1(cool) &= max\_action(\begin{cases} slow \rightarrow U^0(cool) \\ fast \rightarrow \frac{1}{4}U^0(cool)+\frac{3}{4}U^0(warm) \end{cases}) \\
&= max\_action(\begin{cases} slow \rightarrow 48.7666034156 \\ fast \rightarrow 43.074003795 \end{cases}) \\
&= slow
\end{aligned}
$$

Iteration 2:

$$
\begin{aligned}
U^1(warm) &= \frac{4+\frac{\gamma}{1-\gamma}}{(1-\frac{3}{4}\gamma)} \\
&= 40 \\
U^1(cool) &= \frac{4}{1-\gamma} \\
&= 40 \\
U^1(off) &= 0 \\
\Pi^2(warm) &= max\_action(\begin{cases} slow \rightarrow \frac{1}{4}U^1(cool)+\frac{3}{4}U^1(warm) \\ fast \rightarrow \frac{7}{8}U^1(warm)+\frac{1}{8}U^1(off) \end{cases}) \\
&= max\_action(\begin{cases} slow \rightarrow 40 \\ fast \rightarrow 35 \end{cases}) \\
&= slow \\
\Pi^2(cool) &= max\_action(\begin{cases} slow \rightarrow U^1(cool) \\ fast \rightarrow \frac{1}{4}U^1(cool)+\frac{3}{4}U^1(warm) \end{cases}) \\
&= max\_action(\begin{cases} slow \rightarrow 40 \\ fast \rightarrow 40 \end{cases}) \\
&= slow
\end{aligned}
$$

The last iteration has a tie, so action is dependent on implementation.