CS221 Fall 2018 Homework 4

SUNet ID: 05794739
Name: Luis Perez
Collaborators:

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

(a) We give the value for each iteration. We note that $V^0_{\text{opt}}(s) = 0$ to start out. We also note that since for $s_t \in \{-2, 2\}$ we are at a terminal state, we'll have $V_{\text{opt}}(s_t) = 0$ for all iterations.

(a) After iteration 0, we'll have:

$$\begin{aligned} V^0_{\text{opt}}(-1) &= 0 \\ V^0_{\text{opt}}(0) &= 0 \\ V^0_{\text{opt}}(1) &= 0 \end{aligned}$$

(b) After the first iteration, we'll have the following values:

$$\begin{aligned} V^1_{\text{opt}}(-1) &= \max_{a \in \{-1,1\}} \{0.8[20 + V^0_{\text{opt}}(-2)] + 0.2[-5 + V^0_{\text{opt}}(0)], 0.7[20 + V^0_{\text{opt}}(-2)] + 0.3[-5V^0_{\text{opt}}(0)]\} \\ &= 15 \\ V^1_{\text{opt}}(0) &= \max_{a \in \{-1,1\}} \{0.8[-5 + V^0_{\text{opt}}(-1)] + 0.2[-5 + V^0_{\text{opt}}(1)], 0.7[-5 + V^0_{\text{opt}}(-1)] + 0.3[-5 + V^0_{\text{opt}}(\\ &= -5 \\ V^1_{\text{opt}}(1) &= \max_{a \in \{-1,1\}} \{0.8[-5 + V^0_{\text{opt}}(0)] + 0.2[100 + V^0_{\text{opt}}(2)], 0.7[-5 + V^0_{\text{opt}}(0)] + 0.3[100 + V^0_{\text{opt}}(2)\} \\ &= 26.5 \end{aligned}$$

(c) Finally, after the second iteration, we'll have:

$$\begin{aligned} V^2_{\text{opt}}(-1) &= \max_{a \in \{-1,1\}} \{0.8[20 + V^1_{\text{opt}}(-2)] + 0.2[-5 + V^1_{\text{opt}}(0)], 0.7[20 + V^1_{\text{opt}}(-2)] + 0.3[-5 + V^1_{\text{opt}}(0) \\ &= 14 \\ V^2_{\text{opt}}(0) &= \max_{a \in \{-1,1\}} \{0.8[-5 + V^1_{\text{opt}}(-1)] + 0.2[-5 + V^1_{\text{opt}}(1)], 0.7[-5 + V^1_{\text{opt}}(-1)] + 0.3[-5 + V^1_{\text{opt}}(\\ &= 13.45 \\ V^2_{\text{opt}}(1) &= \max_{a \in \{-1,1\}} \{0.8[-5 + V^1_{\text{opt}}(0)] + 0.2[100 + V^1_{\text{opt}}(2)], 0.7[-5 + V^1_{\text{opt}}(0)] + 0.3[100 + V^1_{\text{opt}}(2)\} \\ &= 23 \end{aligned}$$

(b) We interpret this question as asking for the resulting optimal policy for non-terminal states after two iterations. In that case, we have:

$$
\begin{aligned}
\pi^2_{\text{opt}}(-1) &= \arg\max_{a \in \{-1,1\}} \{0.8[20 + V^1_{\text{opt}}(-2)] + 0.2[-5 + V^1_{\text{opt}}(0)], 0.7[20 + V^1_{\text{opt}}(-2)] + 0.3[-5 + V^1_{\text{opt}}(0)]\} \\
&= -1 \\
\pi^2_{\text{opt}}(0) &= \arg\max_{a \in \{-1,1\}} \{0.8[-5 + V^1_{\text{opt}}(-1)] + 0.2[-5 + V^1_{\text{opt}}(1)], 0.7[-5 + V^1_{\text{opt}}(-1)] + 0.3[-5 + V^1_{\text{opt}}(1) \\
&= 1 \\
\pi^2_{\text{opt}}(1) &= \arg\max_{a \in \{-1,1\}} \{0.8[-5 + V^1_{\text{opt}}(0)] + 0.2[100 + V^1_{\text{opt}}(2)], 0.7[-5 + V^1_{\text{opt}}(0)] + 0.3[100 + V^1_{\text{opt}}(2)\} \\
&= 1
\end{aligned}
$$