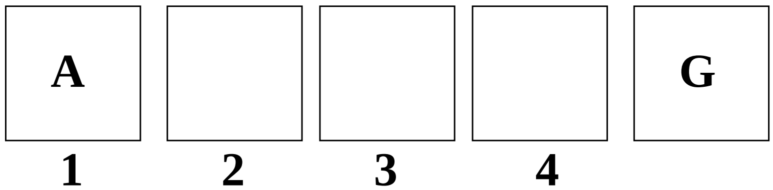
CS 6300 # MDP Practice **February 16, 2021**

A soccer robot A is on a fast break toward the goal, starting in position 1. From positions 1 through 3, it can either shoot (S) or dribble the ball forward (D). From 4 it can only shoot. If it shoots, it either scores a goal (state G) or misses (state M). If it dribbles, it either advances a square or loses the ball, ending up in M. When shooting, the robot is more likely to score a goal from states closer to the goal; when dribbling, the likelihood of missing is independent of the current state.



In this MDP, the states k are 1, 2, 3, 4, G and M, where G and M are terminal states. The transition model depends on the parameter y, which is the probability of dribbling success. Assume a discount of $\gamma = 1$.

$$
\begin{aligned}
T(k, S, G) &= \frac{k}{6} \\
T(k, S, M) &= 1 - \frac{k}{6} \\
T(k, D, k+1) &= y \text{ for } k \in \{1, 2, 3\} \\
T(k, D, M) &= 1 - y \text{ for } k \in \{1, 2, 3\} \\
R(k, S, G) &= 1
\end{aligned}
$$

Rewards are 0 for all other transitions.

1. Using $y = 3/4$, compute the first two iterations of value iteration.

i	$Q_i(1,S)$	$Q_i(2,S)$	$Q_i(3,S)$	$Q_i(4,S)$
0	0	0	0	0
1	1/6	1/3	1/2	2/3
2	1/6	1/3	1/2	2/3

i	$Q_i(1,D)$	$Q_i(2,D)$	$Q_i(3,D)$
0	0	0	0
1	0	0	0
2	1/4	3/8	1/2

i	$V_i(1)$	$V_i(2)$	$V_i(3)$	$V_i(4)$
0	0	0	0	0
1	1/6	1/3	1/2	2/3
2	1/4	3/8	1/2	2/3

The equations for value iteration with $\gamma = 1$ are:

$$
\begin{aligned}
Q^*_{i+1}(s,a) &= \sum_{s'} T(s,a,s')[R(s,a,s') + V^*_i(s')] \\
V^*_{i+1}(s) &= \max_{a_i} Q^*_{i+1}(s,a)
\end{aligned}
$$

Initially, all values are zero. For iteration 1, please note that the end states G, M have no values. For action S, the Q-states are:

$$
\begin{aligned}
Q_1(1,S) &= T(1,S,G)[(R(1,S,G) + V^*_0(G)] + T(1,S,M)[(R(1,S,M) + V^*_0(M)] \\
&= \frac{1}{6}[1+0] + \frac{5}{6}[0+0] = \frac{1}{6} \\
Q_1(2,S) &= T(2,S,G)[(R(2,S,G) + V^*_0(G)] + T(2,S,M)[(R(2,S,M) + V^*_0(M)] \\
&= \frac{1}{3}[1+0] + \frac{2}{3}[0+0] = \frac{1}{3} \\
Q_1(3,S) &= T(3,S,G)[(R(3,S,G) + V^*_0(G)] + T(3,S,M)[(R(3,S,M) + V^*_0(M)] \\
&= \frac{1}{2}[1+0] + \frac{1}{2}[0+0] = \frac{1}{2} \\
Q_1(4,S) &= T(4,S,G)[(R(4,S,G) + V^*_0(G)] + T(4,S,M)[(R(4,S,M) + V^*_0(M)] \\
&= \frac{2}{3}[1+0] + \frac{1}{3}[0+0] = \frac{2}{3}
\end{aligned}
$$

For action D, the Q-states are:

$$
\begin{aligned}
Q_1(1,D) &= T(1,D,2)[(R(1,D,2) + V^*_0(2)] + T(1,D,M)[(R(1,D,M) + V^*_0(M)] \\
&= \frac{3}{4}[0+0] + \frac{1}{4}[0+0] = 0 \\
Q_1(2,D) &= T(2,D,3)[(R(2,D,3) + V^*_0(3)] + T(2,D,M)[(R(2,D,M) + V^*_0(M)] \\
&= \frac{3}{4}[0+0] + \frac{1}{4}[0+0] = 0 \\
Q_1(3,D) &= T(3,D,4)[(R(3,D,4) + V^*_0(4)] + T(3,D,M)[(R(3,D,M) + V^*_0(M)] \\
&= \frac{3}{4}[0+0] + \frac{1}{4}[0+0] = 0
\end{aligned}
$$

The values are now updated.

$$
\begin{aligned}
V_1(1) &= \max_{a\in\{S,D\}} Q_1^*(1,a) = \frac{1}{6} \\
V_1(2) &= \max_{a\in\{S,D\}} Q_1^*(2,a) = \frac{1}{3} \\
V_1(3) &= \max_{a\in\{S,D\}} Q_1^*(3,a) = \frac{1}{2} \\
V_1(4) &= \max_{a\in\{S,D\}} Q_1^*(4,a) = \frac{2}{3}
\end{aligned}
$$

For iteration 2, the Q-states for action S don't change.

$$
\begin{aligned}
Q_2(1,S) &= T(1,S,G)[(R(1,S,G)+V_1^*(G)] + T(1,S,M)[(R(1,S,M)+V_1^*(M)] \\
&= \frac{1}{6}[1+0] + \frac{5}{6}[0+0] = \frac{1}{6} \\
Q_2(2,S) &= T(2,S,G)[(R(2,S,G)+V_1^*(G)] + T(2,S,M)[(R(2,S,M)+V_1^*(M)] \\
&= \frac{1}{3}[1+0] + \frac{2}{3}[0+0] = \frac{1}{3} \\
Q_2(3,S) &= T(3,S,G)[(R(3,S,G)+V_1^*(G)] + T(3,S,M)[(R(3,S,M)+V_1^*(M)] \\
&= \frac{1}{2}[1+0] + \frac{1}{2}[0+0] = \frac{1}{2} \\
Q_2(4,S) &= T(4,S,G)[(R(4,S,G)+V_1^*(G)] + T(4,S,M)[(R(4,S,M)+V_1^*(M)] \\
&= \frac{2}{3}[1+0] + \frac{1}{3}[0+0] = \frac{2}{3}
\end{aligned}
$$

At iteration 2, the Q-states for action D are updated:

$$
\begin{aligned}
Q_2(1,D) &= T(1,D,2)[(R(1,D,2)+V_1^*(2)] + T(1,D,M)[(R(1,D,M)+V_1^*(M)] \\
&= \frac{3}{4}[0+\frac{1}{3}] + \frac{1}{4}[0+0] = \frac{1}{4} \\
Q_2(2,D) &= T(2,D,3)[(R(2,D,3)+V_1^*(3)] + T(2,D,M)[(R(2,D,M)+V_1^*(M)] \\
&= \frac{3}{4}[0+\frac{1}{2}] + \frac{1}{4}[0+0] = \frac{3}{8} \\
Q_2(3,D) &= T(3,D,4)[(R(3,D,4)+V_1^*(4)] + T(3,D,M)[(R(3,D,M)+V_1^*(M)] \\
&= \frac{3}{4}[0+\frac{2}{3}] + \frac{1}{4}[0+0] = \frac{1}{2}
\end{aligned}
$$

The values are now updated.

$$
\begin{aligned}
V_2(1) &= \max_{a\in\{S,D\}} Q_2^*(1,a) = \frac{1}{4} \\
V_2(2) &= \max_{a\in\{S,D\}} Q_2^*(2,a) = \frac{3}{8} \\
V_2(3) &= \max_{a\in\{S,D\}} Q_2^*(3,a) = \frac{1}{2} \\
V_2(4) &= \max_{a\in\{S,D\}} Q_2^*(4,a) = \frac{2}{3}
\end{aligned}
$$

2. After two iterations, perform policy extraction.

   The equation for policy extraction for $\gamma = 1$ is:

$$\pi_i^*(s) \quad = \quad \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + V_i^*(s')]$$

   Solving,

$$\begin{aligned}
\pi_2^*(1) &= \arg \max_{a\in\{S,D\}} \begin{cases} T(1,S,G)[R(1,S,G) + V_2^*(G)] + T(1,S,M)[R(1,S,M) + V_2^*(M)] \\ T(1,D,2)[R(1,D,2) + V_2^*(2)] + T(1,D,M)[R(1,D,M) + V_2^*(M)] \end{cases} \\
&= \arg \max_{a\in\{S,D\}} \begin{cases} \frac{1}{6}[1+0] + \frac{5}{6}[0+0] = \frac{1}{6} \\ \frac{3}{4}[0+\frac{3}{8}] + \frac{1}{4}[0+0] = \frac{9}{32} \end{cases} \\
&= D
\end{aligned}$$

$$\begin{aligned}
\pi_2^*(2) &= \arg \max_{a\in\{S,D\}} \begin{cases} T(2,S,G)[R(2,S,G) + V_2^*(G)] + T(2,S,M)[R(2,S,M) + V_2^*(M)] \\ T(2,D,3)[R(2,D,3) + V_2^*(3)] + T(2,D,M)[R(2,D,M) + V_2^*(M)] \end{cases} \\
&= \arg \max_{a\in\{S,D\}} \begin{cases} \frac{1}{3}[1+0] + \frac{2}{3}[0+0] = \frac{1}{3} \\ \frac{3}{4}[0+\frac{1}{2}] + \frac{1}{4}[0+0] = \frac{3}{8} \end{cases} \\
&= D
\end{aligned}$$

$$\begin{aligned}
\pi_2^*(3) &= \arg \max_{a\in\{S,D\}} \begin{cases} T(3,S,G)[R(3,S,G) + V_2^*(G)] + T(3,S,M)[R(3,S,M) + V_2^*(M)] \\ T(3,D,4)[R(3,D,4) + V_2^*(4)] + T(3,D,M)[R(3,D,M) + V_2^*(M)] \end{cases} \\
&= \arg \max_{a\in\{S,D\}} \begin{cases} \frac{1}{2}[1+0] + \frac{1}{2}[0+0] = \frac{1}{2} \\ \frac{3}{4}[0+\frac{2}{3}] + \frac{1}{4}[0+0] = \frac{1}{2} \end{cases} \\
&= D \text{ or } S
\end{aligned}$$

   Policy $\pi_i^*(4) = S$ by definition.

3. Do two iterations of policy iteration for the initial policy $\pi_0^*(s) = S$.

The equations for policy evaluation and policy extraction for iteration $i+1$ are:

$$\begin{aligned} V_{i+1}^{\pi_k}(s) &= \sum_{s'} T(s,\pi_k(s),s')[R(s,\pi_k(s),s') + V_i^{\pi_k}(s')] \\ \pi_{k+1}(s) &= \arg\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + V^{\pi_k}(s')] \end{aligned}$$

The policy evaluation results are exactly the same as for $Q_i^*(s,S)$ in value iteration.

$$\begin{aligned} V_2^{\pi_0}(1) &= \frac{1}{6} \\ V_2^{\pi_0}(2) &= \frac{1}{3} \\ V_2^{\pi_0}(3) &= \frac{1}{2} \\ V_2^{\pi_0}(4) &= \frac{2}{3} \end{aligned}$$

Policy extraction yields:

$$\begin{aligned} \pi_2(1) &= \arg\max_{a\in\{S,D\}} \begin{cases} T(1,S,G)[R(1,S,G)+V_2^*(G)] + T(1,S,M)[R(1,S,M)+V_2^*(M)] \\ T(1,D,2)[R(1,D,2)+V_2^{\pi_0}(2)] + T(1,D,M)[R(1,D,M)+V_2^*(M)] \end{cases} \\ &= \arg\max_{a\in\{S,D\}} \begin{cases} \frac{1}{6}[1+0] + \frac{5}{6}[0+0] = \frac{1}{6} \\ \frac{3}{4}[0+\frac{1}{3}] + \frac{1}{4}[0+0] = \frac{1}{4} \end{cases} \\ &= D \end{aligned}$$

$$\begin{aligned} \pi_2(2) &= \arg\max_{a\in\{S,D\}} \begin{cases} T(2,S,G)[R(2,S,G)+V_2^*(G)] + T(2,S,M)[R(2,S,M)+V_2^*(M)] \\ T(2,D,3)[R(2,D,3)+V_2^{\pi_0}(3)] + T(2,D,M)[R(2,D,M)+V_2^*(M)] \end{cases} \\ &= \arg\max_{a\in\{S,D\}} \begin{cases} \frac{1}{3}[1+0] + \frac{2}{3}[0+0] = \frac{1}{3} \\ \frac{3}{4}[0+\frac{1}{2}] + \frac{1}{4}[0+0] = \frac{3}{8} \end{cases} \\ &= D \end{aligned}$$

$$
\begin{aligned}
\pi_2(3) &= \arg\max_{a\in\{S,D\}} \begin{cases} T(3,S,G)[R(3,S,G)+V_2^*(G)] + T(3,S,M)[R(3,S,M)+V_2^*(M)] \\ T(3,D,4)[R(3,D,4)+V_2^{\pi_0}(4)] + T(3,D,M)[R(3,D,M)+V_2^*(M)] \end{cases} \\
&= \arg\max_{a\in\{S,D\}} \begin{cases} \frac{1}{2}[1+0] + \frac{1}{2}[0+0] = \frac{1}{2} \\ \frac{3}{4}[0+\frac{2}{3}] + \frac{1}{4}[0+0] = \frac{1}{2} \end{cases} \\
&= D \text{ or } S
\end{aligned}
$$

The policy is the same as for part 2.