AERO 489

Valasek, Selva

4/3/2025

Ian Wilhite

HW4: Q Learning:

a) The reward matrix to represent the problem contains 100 utility for any move that results in being in the goal state (including staying in the goal state). In the first episode (1), we establish the impossible moves as -1, the successful moves into the goal state with a reward state of 100, and other possible moves as having 0 utility.

$$Q_{ep1} = \begin{bmatrix} 0 & -1 & -1 & -1 & 0 & -1 \\ -1 & 0 & -1 & 0 & -1 & 100 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 0 & 0 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 100 \\ -1 & 0 & -1 & -1 & 0 & 100 \end{bmatrix} \tag{1}$$

b) In the second episode (2), we iterate based on the potential utility of the options provided to the agent at any given time. We include staying in the current position as a valid option to allow the agent to save the utility of its current position, and to ease the extraction of the convergent utility of the entire system.

$$Q_{ep2} = \begin{bmatrix} 0 & -1 & -1 & -1 & 0 & -1 \\ -1 & 0 & -1 & 0 & -1 & 100 \\ -1 & -1 & 0 & 0 & -1 & -1 \\ -1 & 80 & 0 & 0 & 0 & -1 \\ 0 & -1 & -1 & 0 & 0 & 100 \\ -1 & 0 & -1 & -1 & 80 & 100 \end{bmatrix} \tag{2}$$

$$Q_{ep3} = \begin{bmatrix} 0 & -1 & -1 & -1 & 80 & -1 \\ -1 & 0 & -1 & 64 & -1 & 100 \\ -1 & -1 & 0 & 64 & -1 & -1 \\ -1 & 80 & 0 & 0 & 80 & -1 \\ 0 & -1 & -1 & 0 & 80 & 100 \\ -1 & 0 & -1 & -1 & 80 & 100 \end{bmatrix} \tag{3}$$

$$Q_{ep4} = \begin{bmatrix} 64 & -1 & -1 & -1 & 80 & -1 \\ -1 & 0 & -1 & 64 & -1 & 100 \\ -1 & -1 & 0 & 64 & -1 & -1 \\ -1 & 80 & 0 & 0 & 80 & -1 \\ 64 & -1 & -1 & 64 & 80 & 100 \\ -1 & 0 & -1 & -1 & 80 & 100 \end{bmatrix} \tag{4}$$

$$Q_{ep5} = \begin{bmatrix} 64 & -1 & -1 & -1 & 80 & -1 \\ -1 & 0 & -1 & 64 & -1 & 100 \\ -1 & -1 & 0 & 64 & -1 & -1 \\ -1 & 80 & 51.2 & 0 & 80 & -1 \\ 64 & -1 & -1 & 64 & 80 & 100 \\ -1 & 0 & -1 & -1 & 80 & 100 \end{bmatrix} \tag{5}$$

c) The final convergent matrix can be found by applying the process many times. After 1000 trials, the convergent matrix was found in (6). Some observations to note include that all columns contain either -1 or a constant. This represents that after converging, moving to that state from any other state will result in the same utility, or conversely, that the value of a given state is independent of the direction of approach. Recalling that the diagonal represents that utility of staying in a given state, combined that all columns contain the same value, the diagonal of the convergent matrix can be utilized to develop a common utility for each state.

$$Q_{conv.} = \begin{bmatrix} 64 & -1 & -1 & -1 & 80 & -1 \\ -1 & 80 & -1 & 64 & -1 & 100 \\ -1 & -1 & 51.2 & 64 & -1 & -1 \\ -1 & 80 & 51.2 & 64 & 80 & -1 \\ 64 & -1 & -1 & 64 & 80 & 100 \\ -1 & 80 & -1 & -1 & 80 & 100 \end{bmatrix} \tag{6}$$

By evaluating the diagonal, we can observe that the convergent Q matrix represents the intuition of the initial problem. The goal state is worth the full goal value of 100. States one and four that allow for one move to the goal state have value of the goal state multiplied by the discount factor, or 80. The subsequent states are each valued at their optimal move value multiplied by the discount factor, or valued at the goal state value multiplied by the discount factor to the power of the minimum number of moves to the goal state. This Q matrix would create an optimal policy to provide an appropriate optimal policy to exit the house.