CSCI-GA.2560-001, Artificial Intelligence April 4, 2022

Solutions to Problem 1 of Homework 8 (5 Points)

Name: Anav Prasad (ap7152) *Due: 5PM on Monday, April 4*

Collaborators:

Gridworld (4x4 - skipped O as too similar to 0)

A	B	C	D
E	F	G	H
I	J	K	L
M	N	P	Q

Similar to the maze from class, in Gridworld you can make 4 possible moves: up, down, left, right. However, in case you are at an edge/corner, rather than lose moves the boundary moves simply transition to self. Multiple ”self” moves are merged.

e.g. the actions from M are I, N, M and from P are {K,N,P,Q}

You may either:

- Use a spreadsheet (Note: search for ”iterative calculation” for info on setting up excel or sheets to allow recursive formulas.)
- Use your Lab 3 solver
- Use a closed form algebraic computation

Please indicate which you did in case of small numerical discrepancies.

Print answers to 3 decimal places, using 150 iterations and tolerance of 0.001.

Answers accepted correct within .01

Remember: α is the ”transition failure probability” for a decision node. So $\alpha = 0.15$ means the success rate is 0.85

You will use the Bellman equation: the utility of a state is equal to the immediate reward for that state, plus the discounted utility of future state(s).

v(s) = R(s) + γ * P * v(s)

Question:

Assuming that:

- A and Q are terminal states with a reward of 2

- all other states give a reward of -3 and are chance nodes with uniform random probabilities (.25 for each of the 4 transitions, .333 for corners).

Solve the value function as a Markov reward process. Print the 14 non-terminal values.

Solution:
Formulas for value iteration:

A $= 2$	B $= -3 + \frac{A+F+C+B}{4}$	C $= -3 + \frac{B+G+D+C}{4}$	D $= -3 + \frac{C+H+D}{3}$
E $= -3 + \frac{E+I+F+A}{4}$	F $= -3 + \frac{E+J+G+B}{4}$	G $= -3 + \frac{F+K+H+C}{4}$	H $= -3 + \frac{G+L+H+D}{4}$
I $= -3 + \frac{I+M+J+E}{4}$	J $= -3 + \frac{I+N+K+F}{4}$	K $= -3 + \frac{J+P+L+G}{4}$	L $= -3 + \frac{K+Q+L+H}{4}$
M $= -3 + \frac{M+N+I}{3}$	N $= -3 + \frac{M+N+P+J}{4}$	P $= -3 + \frac{N+P+Q+K}{4}$	Q $= 2$

So, the values obtained after value iteration are as follows:

A = 2	B = -38.493	C = -55.366	D = -59.865
E = -38.493	F = -50.117	G = -55.741	H = -55.366
I = -55.366	J = -55.741	K = -50.118	L = -38.494
M = -59.865	N = -55.366	P = -38.494	Q = 2

□

CSCI-GA.2560-001, Artificial Intelligence April 4, 2022

Solutions to Problem 2 of Homework 8 (5 Points)

Name: Anav Prasad (ap7152) *Due: 5PM on Monday, April 4*

Collaborators:

Assuming that:

- A and Q are terminal states with a reward of 15 and -15 respectively
- states J and G give a reward of 3, and are chance nodes with uniform random probabilities
- all other states give a reward of -1 and are decision nodes

Using a discount factor γ of 0.9 and a Q-learning α of 0.15 (a.k.a. decision node probability of failure), solve the MDP using value iteration and greedy policy computation.
Print out the learned policy, e.g. {F → E, K → J, ...} and also the 14 non-terminal values under that policy.

Solution:
The learned policy is as follows:

- $B \to A$
- $C \to B$
- $D \to C$
- $E \to A$
- $F \to E$
- $H \to G$
- $I \to E$
- $K \to J$
- $L \to K$
- $M \to I$
- $N \to J$
- $P \to K$

Formulas for value iteration using the learned policies:

$$
\begin{aligned}
A &= 15\\
B &= -1 + \gamma \cdot \Big((1-\alpha)\cdot A + \frac{\alpha}{3}\cdot(F + C + B)\Big)\\
C &= -1 + \gamma \cdot \Big((1-\alpha)\cdot B + \frac{\alpha}{3}\cdot(G + D + C)\Big)\\
D &= -1 + \gamma \cdot \Big((1-\alpha)\cdot C + \frac{\alpha}{2}\cdot(H + D)\Big)\\
E &= -1 + \gamma \cdot \Big((1-\alpha)\cdot A + \frac{\alpha}{3}\cdot(E + I + F)\Big)\\
F &= -1 + \gamma \cdot \Big((1-\alpha)\cdot E + \frac{\alpha}{3}\cdot(J + G + B)\Big)\\
G &= 3 + \frac{\gamma}{4}\cdot(F + K + H + C)\\
H &= -1 + \gamma \cdot \Big((1-\alpha)\cdot G + \frac{\alpha}{3}\cdot(L + H + D)\Big)\\
I &= -1 + \gamma \cdot \Big((1-\alpha)\cdot E + \frac{\alpha}{3}\cdot(I + M + J)\Big)\\
J &= 3 + \frac{\gamma}{4}\cdot(I + N + K + F)\\
K &= -1 + \gamma \cdot \Big((1-\alpha)\cdot J + \frac{\alpha}{3}\cdot(P + L + G)\Big)\\
L &= -1 + \gamma \cdot \Big((1-\alpha)\cdot K + \frac{\alpha}{3}\cdot(Q + L + H)\Big)\\
M &= -1 + \gamma \cdot \Big((1-\alpha)\cdot I + \frac{\alpha}{2}\cdot(M + N)\Big)\\
N &= -1 + \gamma \cdot \Big((1-\alpha)\cdot J + \frac{\alpha}{3}\cdot(M + N + P)\Big)\\
P &= -1 + \gamma \cdot \Big((1-\alpha)\cdot K + \frac{\alpha}{3}\cdot(N + P + Q)\Big)\\
Q &= -15
\end{aligned}
$$

So, the values obtained after value iteration, using the learned policies, are as follows:

A = 15	B = 11.860	C = 9.311	D = 7.173
E = 11.860	F = 9.600	G = 11.037	H = 8.386
I = 9.311	J = 11.037	K = 8.425	L = 5.390
M = 7.173	N = 8.386	P = 5.390	Q = -15

□