SAiDL ASSIGNMENT

Aadetya Jaiswal

31st August 2019

1 Assignment 2 problem statement

Dear Friend, Some time ago, I bought this old house, but found it to be haunted by ghostly sardonic laughter. As a result it is hardly habitable. There is hope, however, for by actual testing I have found that this haunting is subject to certain laws, obscure but infallible, and that the laughter can be affected by my playing the organ or burning incense. In each minute, the laughter occurs or not, it shows no degree. What it will do during the ensuing minute depends, in the following exact way, on what has been happening during the preceding minute: Whenever there is laughter, it will continue in the succeeding minute unless I play the organ, in which case it will stop. But continuing to play the organ does not keep the house quiet. I notice, however, that whenever I burn incense when the house is quiet and do not play the organ it remains quiet for the next minute. At this minute of writing, the laughter is going on. Please tell me what manipulations of incense and organ I should make to get that house quiet, and to keep it so. Sincerely, At Wits End.

a) Formulate this problem as an MDP (for the sake of uniformity , formulate it as a continuing discounted problem with $\gamma = 0.9$. Let the reward be +1 on any transition into the silent state, and -1 on any transition into the laughing state). Explicitly give the state set, action sets, state transition and reward function.

1.1 Answer for question 1.

State set: (L, Q), where L indicates that there is laughter in the room, and Q indicates that the room is quiet.
Action set: $(O \wedge I, O \wedge \neg I, \neg O \wedge I, \neg O \wedge \neg I)$, where O corresponds to playing the organ, and I corresponds to burning incense.
We consider this as a continuing discounted problem with $\gamma = 0.9$ and we let the

reward be +1 on any transition into the silent state, and -1 on any transition into the laughing state.

b) Starting with the policy $\pi(laughing) = \pi(silent) = \pi(incense, noorgan)$, perform a couple of policy iterations(by hand) until you find an optimal policy.(Clearly show and label each step. If you are taking a lot of iterations, stop and reconsider your formulation).
Do a couple of value iterations as well.

1.2 Answer for question 2.

1.2.1 Notation.

$P^{\pi}_{s1 \rightarrow s2}$ denotes the probability of transition from state1 to state2 following policy π

δ denotes the difference between the value function in consecutive iterations.

θ denotes the threshold such that if $\delta < \theta$ policy improvement will take place.

1.2.2 Policy Iteration

Initialisation: V (L) = V (Q) = 0; π(L) = π(Q) = $\neg$O $\wedge$ I Assuming θ = 0.7 and given, γ = 0.9
Evaluation:
δ=0
V(L)=$P^{\pi(L)}_{L\rightarrow L}$*[$R^{\pi(L)}_{L\rightarrow L}$ + γ*V(L)] + $P^{\pi(L)}_{L\rightarrow Q}$*[$R^{\pi(L)}_{L\rightarrow Q}$ + γ*V(L)] = -1
V(Q)=$P^{\pi(Q)}_{Q\rightarrow L}$*[$R^{\pi(Q)}_{Q\rightarrow L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\rightarrow Q}$*[$R^{\pi(Q)}_{Q\rightarrow Q}$ + γ*V(Q)] = +1

δ=1
V(L)=$P^{\pi(L)}_{L\rightarrow L}$*[$R^{\pi(L)}_{L\rightarrow L}$ + γ*V(L)] + $P^{\pi(L)}_{L\rightarrow Q}$*[$R^{\pi(L)}_{L\rightarrow Q}$ + γ*V(L)] = -1.9
V(Q)=$P^{\pi(Q)}_{Q\rightarrow L}$*[$R^{\pi(Q)}_{Q\rightarrow L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\rightarrow Q}$*[$R^{\pi(Q)}_{Q\rightarrow Q}$ + γ*V(Q)] = +1.9

δ=0.9
V(L)=$P^{\pi(L)}_{L\rightarrow L}$*[$R^{\pi(L)}_{L\rightarrow L}$ + γ*V(L)] + $P^{\pi(L)}_{L\rightarrow Q}$*[$R^{\pi(L)}_{L\rightarrow Q}$ + γ*V(L)] = -2.71
V(Q)=$P^{\pi(Q)}_{Q\rightarrow L}$*[$R^{\pi(Q)}_{Q\rightarrow L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\rightarrow Q}$*[$R^{\pi(Q)}_{Q\rightarrow Q}$ + γ*V(Q)] = +2.71

δ=0.81
V(L)=$P^{\pi(L)}_{L\rightarrow L}$*[$R^{\pi(L)}_{L\rightarrow L}$ + γ*V(L)] + $P^{\pi(L)}_{L\rightarrow Q}$*[$R^{\pi(L)}_{L\rightarrow Q}$ + γ*V(L)] = -3.439
V(Q)=$P^{\pi(Q)}_{Q\rightarrow L}$*[$R^{\pi(Q)}_{Q\rightarrow L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\rightarrow Q}$*[$R^{\pi(Q)}_{Q\rightarrow Q}$ + γ*V(Q)] = +3.439

δ=0.729
V(L)=$P^{\pi(L)}_{L\rightarrow L}$*[$R^{\pi(L)}_{L\rightarrow L}$ + γ*V(L)] + $P^{\pi(L)}_{L\rightarrow Q}$*[$R^{\pi(L)}_{L\rightarrow Q}$ + γ*V(L)] = -4.0951

V(Q)=$P^{\pi(Q)}_{Q\to L}$*[$R^{\pi(Q)}_{Q\to L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\to Q}$*[$R^{\pi(Q)}_{Q\to Q}$ + γ*V(Q)] = +4.0951

δ = 0.6561
Improvement:
π(L) = O $\wedge$ I (or O $\wedge$ $\neg$ I)
π(Q) = $\neg$O $\wedge$ I
Evaluation:
δ=0
V(L)=$P^{\pi(L)}_{L\to L}$*[$R^{\pi(L)}_{L\to L}$ + γ*V(L)] + $P^{\pi(L)}_{L\to Q}$*[$R^{\pi(L)}_{L\to Q}$ + γ*V(L)] = -4.6856
V(Q)=$P^{\pi(Q)}_{Q\to L}$*[$R^{\pi(Q)}_{Q\to L}$ + γ*V(Q)] + $P^{\pi(Q)}_{Q\to Q}$*[$R^{\pi(Q)}_{Q\to Q}$ + γ*V(Q)] = +4.6856
δ=0.5905
Improvement:
π(L) = O $\wedge$ I (or O $\wedge$ $\neg$I)
π(Q) = $\neg$O $\wedge$ I
No change in policy. This is an optimal policy

1.2.3 Value Iterations

Initialisation: V (L) = V (Q) = 0; Assuming θ = 0.9
δ = 0
V(L)=max_a($P^{\pi(a)}_{L\to L}$*[$R^{\pi(a)}_{L\to L}$ + γ*V(L)] + $P^{\pi(a)}_{L\to Q}$*[$R^{\pi(a)}_{L\to Q}$ + γ*V(L)]) = +1
For a = O $\wedge$ I or O $\wedge$ $\neg$I:
V(Q)=max_a($P^{\pi(a)}_{Q\to L}$*[$R^{\pi(a)}_{Q\to L}$ + γ*V(Q)] + $P^{\pi(a)}_{Q\to Q}$*[$R^{\pi(a)}_{Q\to Q}$ + γ*V(Q)]) = +1
For a = $\neg$ O $\wedge$ I :

δ = 0.9
V(L)=max_a($P^{\pi(a)}_{L\to L}$*[$R^{\pi(a)}_{L\to L}$ + γ*V(L)] + $P^{\pi(a)}_{L\to Q}$*[$R^{\pi(a)}_{L\to Q}$ + γ*V(L)]) = +1.9
For a = O $\wedge$ I or O $\wedge$ $\neg$I:
V(Q)=max_a($P^{\pi(a)}_{Q\to L}$*[$R^{\pi(a)}_{Q\to L}$ + γ*V(Q)] + $P^{\pi(a)}_{Q\to Q}$*[$R^{\pi(a)}_{Q\to Q}$ + γ*V(Q)]) = +1.9
For a = $\neg$ O $\wedge$ I :

δ = 0.81
V(L)=max_a($P^{\pi(a)}_{L\to L}$*[$R^{\pi(a)}_{L\to L}$ + γ*V(L)] + $P^{\pi(a)}_{L\to Q}$*[$R^{\pi(a)}_{L\to Q}$ + γ*V(L)]) = +2.71
For a = O $\wedge$ I or O $\wedge$ $\neg$I:
V(Q)=max_a($P^{\pi(a)}_{Q\to L}$*[$R^{\pi(a)}_{Q\to L}$ + γ*V(Q)] + $P^{\pi(a)}_{Q\to Q}$*[$R^{\pi(a)}_{Q\to Q}$ + γ*V(Q)]) = +2.71
For a = $\neg$ O $\wedge$ I :

Value of δ decreases by a factor of γ in every iteration therefore continuing evaluation will lead to an convergence value of ±10 ±1(1+γ+γ^2...)
Therefore:
Deterministic policy:
π(L) = O $\wedge$ I (or O $\wedge$ $\neg$I)
π(Q) = $\neg$O $\wedge$ I

c) What are the resulting optimal state-action values for all state-action pairs?

1.3 Answer for question 3.

$q_*(s,a)$ denotes the expected value of reward
Optimal state-action values:

Current state	**Action**	**Next State**	$q_*(s,a)$
L	**O $\wedge$ I**	**Q**	**±10**
L	**O $\wedge$ ¬I**	**Q**	**±10**
L	**¬O $\wedge$ I**	**L**	**±8**
L	**¬O $\wedge$ ¬ I**	**L**	**±8**
Q	**O $\wedge$ I**	**L**	**±8**
Q	**O $\wedge$ ¬I**	**L**	**±8**
Q	**¬O $\wedge$ I**	**Q**	**±10**
Q	**¬O $\wedge$ ¬ I**	**L**	**±8**

Taking a sub optimal action initially and then following the optimal policy results in $q_*(s,a)$=8 (-1+γ*10)
On the contrary following the optimal policy results in $q_*(s,a)$=10

d) What is your advice to At Wits End?

1.4 Answer to question 4

If there is laughter, play the organ; if room is quite, do not play the organ and burn incense

END OF THE ASSIGNMENT PROBLEM