ELL729 Assignment 3

January 17, 2021

1 Problem 1

Q learning algorithm was used with epsilon greedy method. The iteration we use is as follows:-

$$Q_{n+1}(x,u) = Q_n(x,u) + a(\nu(x,u)) * I_{\{Xn=x,Un=u\}} * [g(x,u) + \alpha * min_{u'} Q_n(Y(x,u),u') - Q_n(x,u)]$$

where $a(n) = 1/n^{0.6}$ and it satisfies the properties $\sum a(n) = \infty$, $\sum a(n)^2 < \infty$, $\lim_{n->\infty} a(n) = 0$. Y(x,u) is the nest state when action u is taken on state x. $\alpha = 1$ in our case.g(x,u) is -1 always except the case where the transition is made to the destination that is F.

1.1 Optimal Path

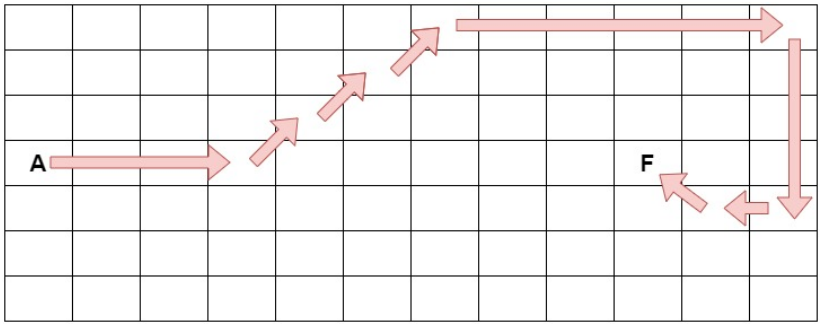

Figure 1: optimal path for problem 1

1.2 Optimal Policy

E	E	E	E	E	E	E	E	E	E	E	S
E	E	E	E	E	E	E	E	E	E	E	S
E	E	E	E	E	E	E	E	E	E	E	S
E	E	E	E	E	E	E	E	E	S	E	S
E	E	E	E	E	E	E	E	E	E	W	W
E	E	E	E	E	E	E	E	E	N	E	N
E	E	E	E	E	E	E	E	S	W	N	W

Figure 2: optimal policy for problem 1

1.3 Comments

- The optimal policy is as expected. The winds are strong so even if it tries to counter it by taking steps to the right and downwards, in order to go on a straight path, the strong winds will anyway take it to the top. So its best that the ship itsel goes right always and wind pushes it to the topright, from where it can go down and left to reach F.
- We see from the optimal policy that all states lead to F only.
- For all values of epsilon, the same optimal pllicy was obtained after the iteration was over.
- Sum of values of the Q matrix were plotted, and a typical convergence plot is as folows

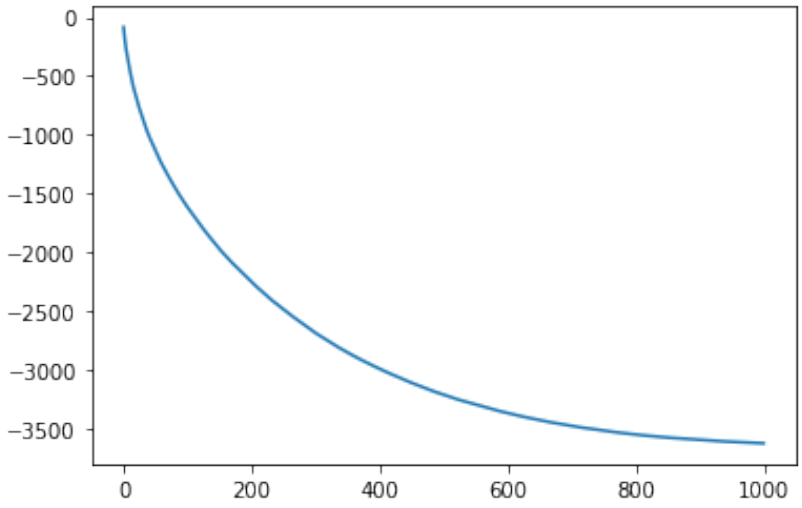

Figure 3: Sum of all elements of Q matrix vs iterations

2 Problem 2

Algorithm is the same, but g(x,u) has changed according to the question and also that if the ship goes into pirate infested water then the next state is A and the cost incurred is -100.All other costs are -1 except cost of transitioning to F which is 0.

2.1 Optimal Path

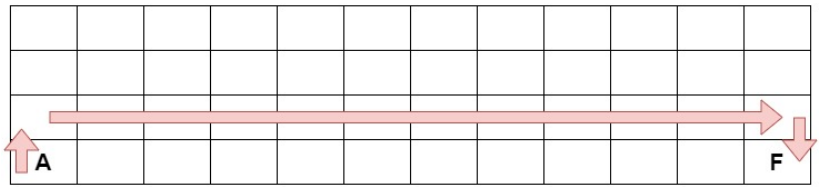

Figure 4: optimal path for problem 2

2.2 Optimal Policy

E	E	E	E	E	E	E	E	E	E	E	S
E	E	E	E	E	E	E	E	E	E	E	S
E	E	E	E	E	E	E	E	E	E	E	S
N	N	N	E	E	E	E	E	E	E	E	E

Figure 5: optimal policy for problem 2

2.3 Comments

- The optimal policy is as expected. There are no winds and the ship should not go into the pirates area. So it goes one step up, then right, till the end and then downwards to reach F. This is clearly the optimal policy for the ship.
- For all values of epsilon, the same optimal pllicy was obtained after the iteration was over. For larger values of epsilon, the number of iterations required for convergence was more(or equivalently the convergence was late). This is expected becasue for larger epsilon it explores more than it exploits and hence the convergence to the optimal takes more time
- Sum of values of the Q matrix were plotted, and a typical convergence plot is as folows

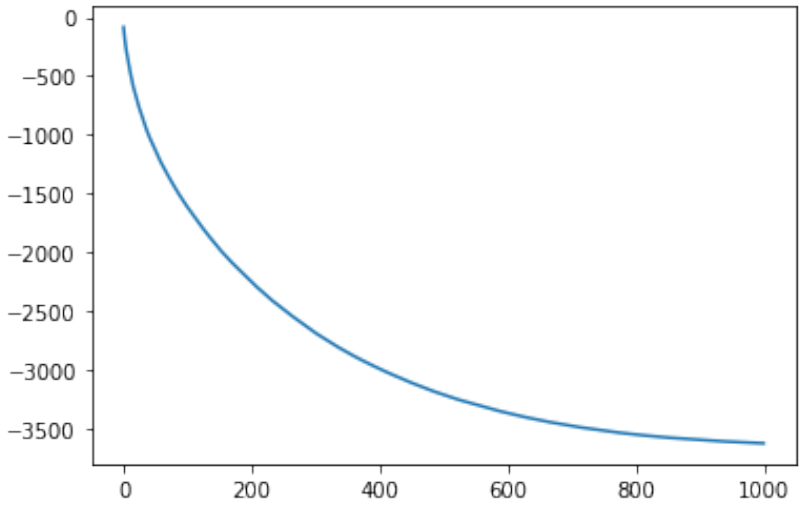

Figure 6: Sum of all elements of Q matrix vs iterations

2.4 Last part of Problem 2

- Yes this task is episodic as it has a defined end state which is absorbing. ie the whole task runs for a finite amount of time till it reaches that absorbing state which is the end state
- The problem with epsilon greedy strategy is that, when running our iterations after convergence of Q learning, the epsilon greedy strategy may take the ship into the pirates area and that would decrease the average reward.
- There is discrepancy. The average reward should be -1*12 = -12 according to our optimal policy but actually it comes lower (eg-600, -700) when we run 1000 episodes an take the average and this is happening due to the epsilon greedy strategy