supplementary file for CA2 – RL

Amirmahdi Farzane

I tried different methods for epsilon decay:

1) max(0.1 ,((EPISODES-episode)/EPISODES)*1)

This method depends on number of episodes that finished. More episode, lower epsilon. But I set lower bound 0.1. To save chance of random actions always.

Advantages : first exploring and decreasing per episodes are good but ..

Disadvantages : but after some iterations we find optimal policy and epsilon must be so slow , hence this method is not enough good because it's motivation for exploration is forever(the lowest value is 0.1 and it is bad).

2) 1 / (1+ 10**(episode))

This method fixed previous problem and we can decrease epsilon forever but because of degree 2 for 10 we so soon decrease epsilon so we can't explore as enough as we need.

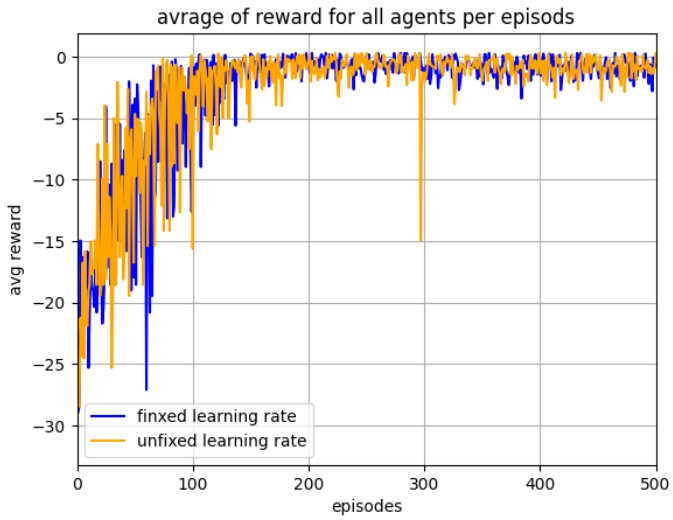
3) 1 / (1+ 10*(episode))

so I decreased used multiplying instead of power and finally I got better result.

1 / (1+ 10*(epsilon))

I did a mistake and set right side of multiplying epsilon instead of episode and this method changed.

So I got this curve :



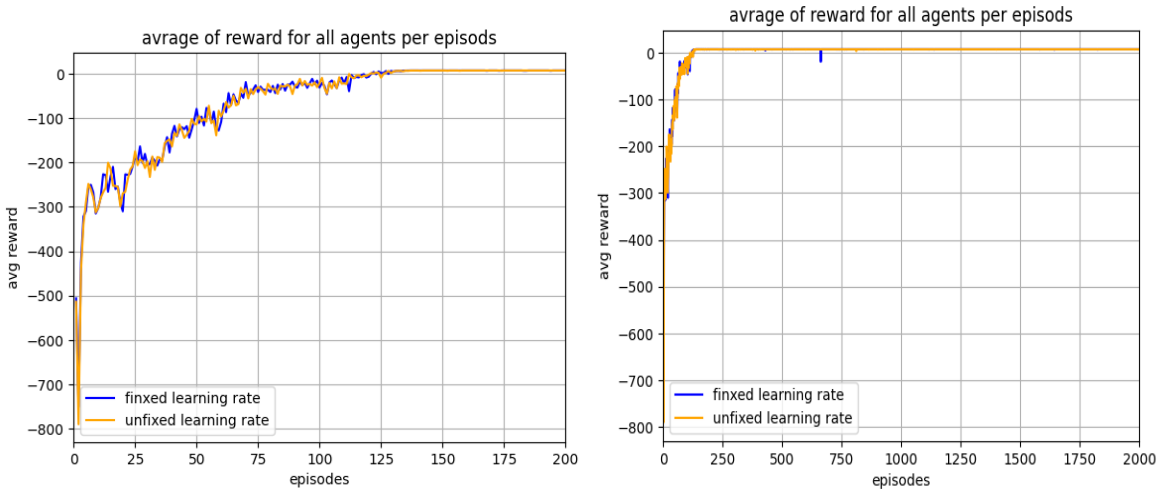
So I understood that I misstated in choosing appropriate epsilon and printed values of epsilon at some iterations and were

like this

1. 0.09090909090909091
2. 0.5238095238095238
3. 0.16030534351145037
4. 0.38416422287390034
5. 0.20654149000605693
6. 0.32622011460185735
7. 0.23462055537527235
8. 0.2988459566922043
9. 0.2507233640509776
10. 0.2851250023521815

And after that I found impact of this wring method . Changing unregular epsilon caused non converging and these unregular change of avg rewards because unregular chance of choosing random or optimal action.

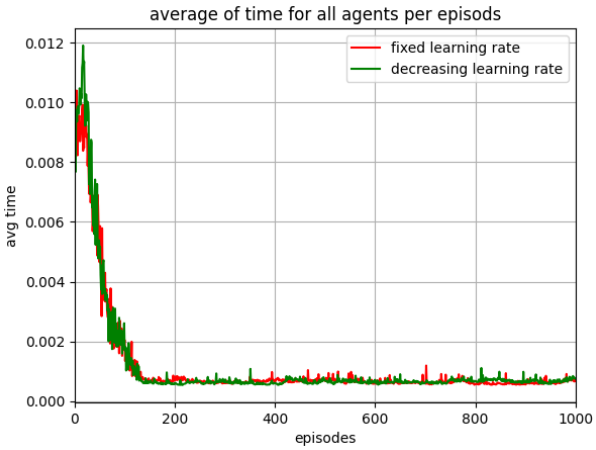
Final result for avg reward:

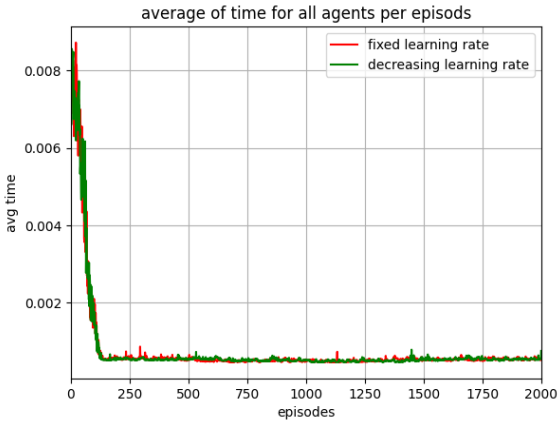


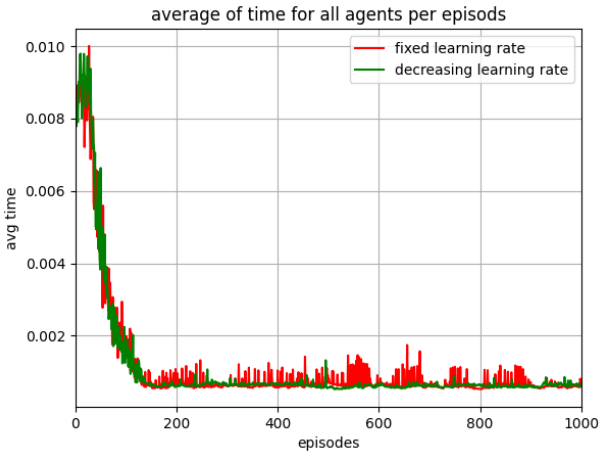
Noises before converging made by choosing random action and after some episodes low epsilon leaded to stable average reward and stable choosing action policy.

Time variance doesn’t depend on change mode of learning rate.

Converging in this two method (fixed or decreasing) is so similar .In decreasing method is a little faster.


average of time for all agents per episods
fixed learning rate
decreasing learning rate
avg time
episodes
0.012
0.010
0.008
0.006
0.004
0.002
0.000
0
200
400
600
800
1000


average of time for all agents per episods
fixed learning rate
decreasing learning rate
avg time
episodes
0.008
0.006
0.004
0.002
0
250
500
750
1000
1250
1500
1750
2000


average of time for all agents per episods
fixed learning rate
decreasing learning rate
avg time
episodes
0.010
0.008
0.006
0.004
0.002
0
200
400
600
800
1000