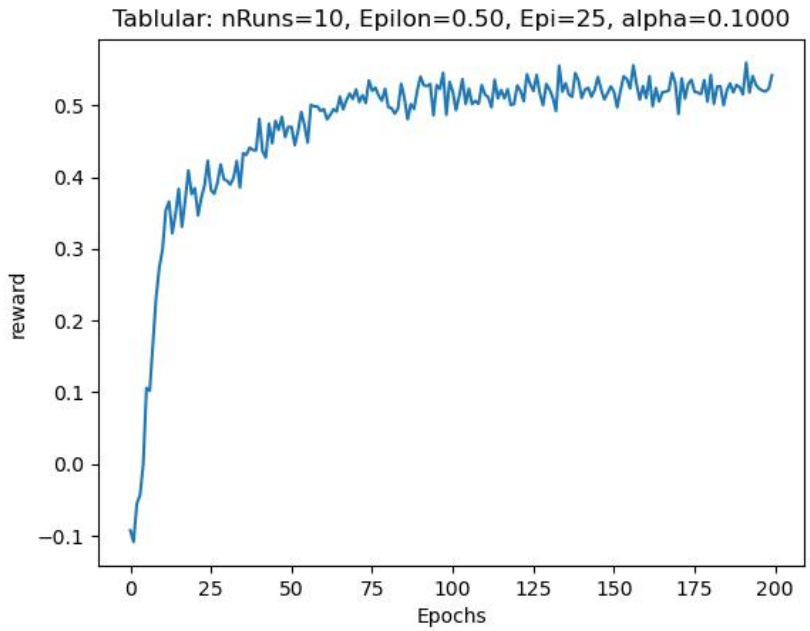
Question2



Showing in the figure, in the first estimated 12 epochs, it went up quickly, then it slowly rose and moved to converge. We can estimate that when it comes to converge, the number of epoch is almost 100.

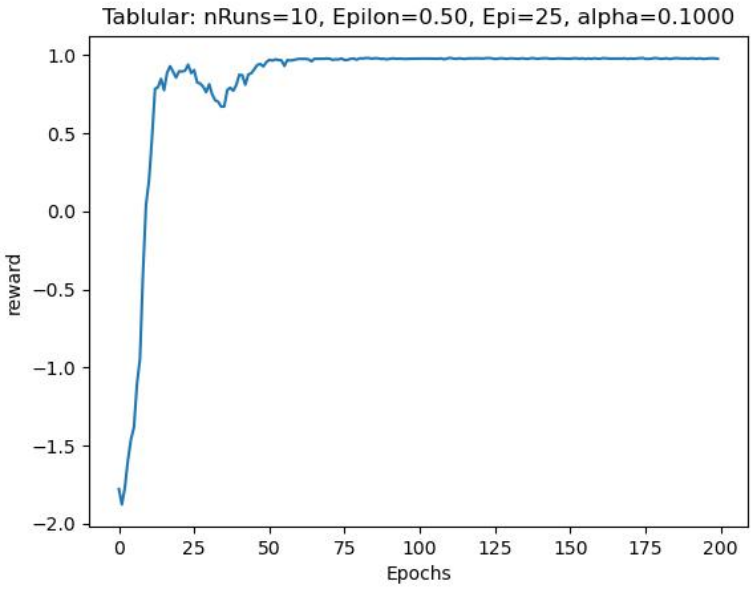
Question 3

```
Avg reward: 0.469538 | Ewma reward: 0.529629: 100%|| 200/200 [00:01<00:00, 135.
Avg reward: 0.471308 | Ewma reward: 0.528880: 100%|| 200/200 [00:01<00:00, 132.
Avg reward: 0.471772 | Ewma reward: 0.519437: 100%|| 200/200 [00:01<00:00, 136.
Avg reward: 0.474021 | Ewma reward: 0.537871: 100%|| 200/200 [00:01<00:00, 137.
Avg reward: 0.471540 | Ewma reward: 0.524827: 100%|| 200/200 [00:01<00:00, 137.
Avg reward: 0.464660 | Ewma reward: 0.519332: 100%|| 200/200 [00:01<00:00, 135.
Avg reward: 0.471304 | Ewma reward: 0.517616: 100%|| 200/200 [00:01<00:00, 135.
Avg reward: 0.471185 | Ewma reward: 0.524430: 100%|| 200/200 [00:01<00:00, 132.
Avg reward: 0.469813 | Ewma reward: 0.531181: 100%|| 200/200 [00:01<00:00, 136.
Avg reward: 0.468187 | Ewma reward: 0.529592: 100%|| 200/200 [00:01<00:00, 135.
```

The figure shows that the average reward is around 0.468187.

Question 4

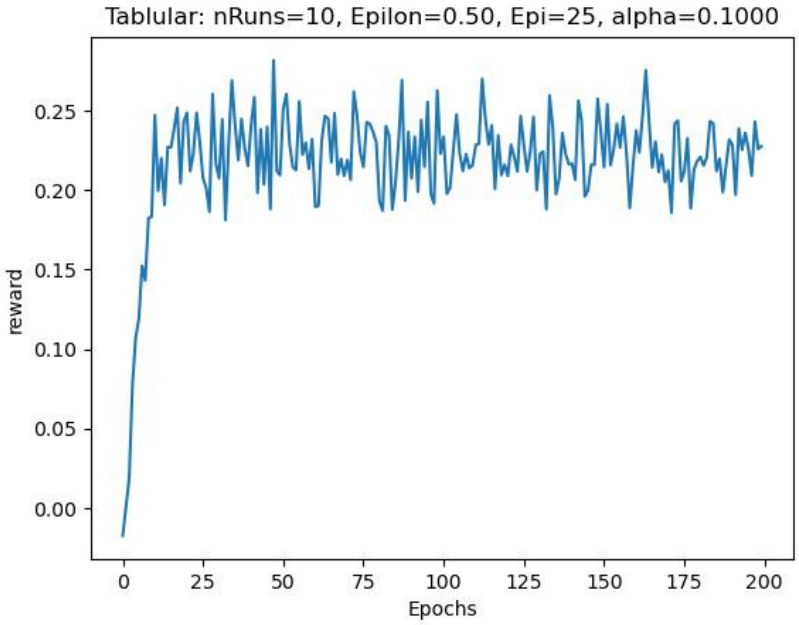
(a) For GAMMA = 1




```
Avg reward: 0.836605 | Ewma reward: 0.979417: 100%|| 200/200 [00:01<00:00, 136.
Avg reward: 0.835730 | Ewma reward: 0.980586: 100%|| 200/200 [00:01<00:00, 135.
Avg reward: 0.837650 | Ewma reward: 0.980568: 100%|| 200/200 [00:01<00:00, 139.
Avg reward: 0.836081 | Ewma reward: 0.979172: 100%|| 200/200 [00:01<00:00, 140.
Avg reward: 0.835355 | Ewma reward: 0.978326: 100%|| 200/200 [00:01<00:00, 140.
Avg reward: 0.830439 | Ewma reward: 0.977905: 100%|| 200/200 [00:01<00:00, 135.
Avg reward: 0.832641 | Ewma reward: 0.978250: 100%|| 200/200 [00:01<00:00, 138.
Avg reward: 0.830180 | Ewma reward: 0.977666: 100%|| 200/200 [00:01<00:00, 139.
Avg reward: 0.834440 | Ewma reward: 0.978378: 100%|| 200/200 [00:01<00:00, 142.
Avg reward: 0.836547 | Ewma reward: 0.979952: 100%|| 200/200 [00:01<00:00, 139.
```

The smooth line in the figure shows that it rises very fast at first, also comes to converge at almost the 50th epoch. Thus it converges faster compared to GAMMA = 0.5.

(b) For GAMMA = 0.000001

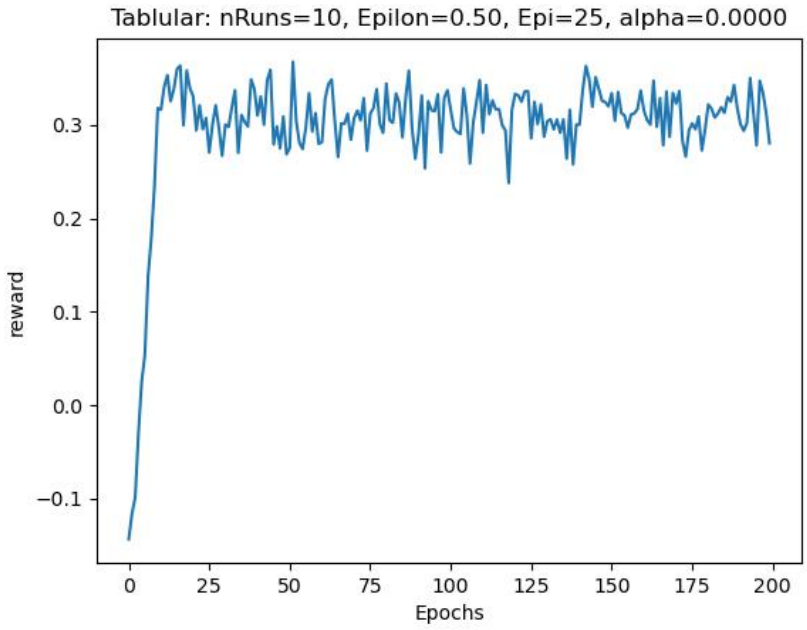


```
Avg reward: 0.213272 | Ewma reward: 0.230130: 100%|| 200/200 [00:02<00:00, 68.2
Avg reward: 0.215671 | Ewma reward: 0.215365: 100%|| 200/200 [00:02<00:00, 68.3
Avg reward: 0.220682 | Ewma reward: 0.254737: 100%|| 200/200 [00:02<00:00, 68.5
Avg reward: 0.216781 | Ewma reward: 0.203247: 100%|| 200/200 [00:02<00:00, 70.3
Avg reward: 0.214303 | Ewma reward: 0.236803: 100%|| 200/200 [00:02<00:00, 69.6
Avg reward: 0.213838 | Ewma reward: 0.215965: 100%|| 200/200 [00:02<00:00, 69.5
Avg reward: 0.220922 | Ewma reward: 0.222569: 100%|| 200/200 [00:02<00:00, 68.7
Avg reward: 0.229028 | Ewma reward: 0.236364: 100%|| 200/200 [00:02<00:00, 68.2
Avg reward: 0.217787 | Ewma reward: 0.220694: 100%|| 200/200 [00:02<00:00, 68.6
Avg reward: 0.219364 | Ewma reward: 0.208346: 100%|| 200/200 [00:02<00:00, 67.5
```

The figure shows that this time it has a huge fluctuation, and it is really hard to say that it converged. Thus it converges slower compared to GAMMA = 0.5.

Question 5

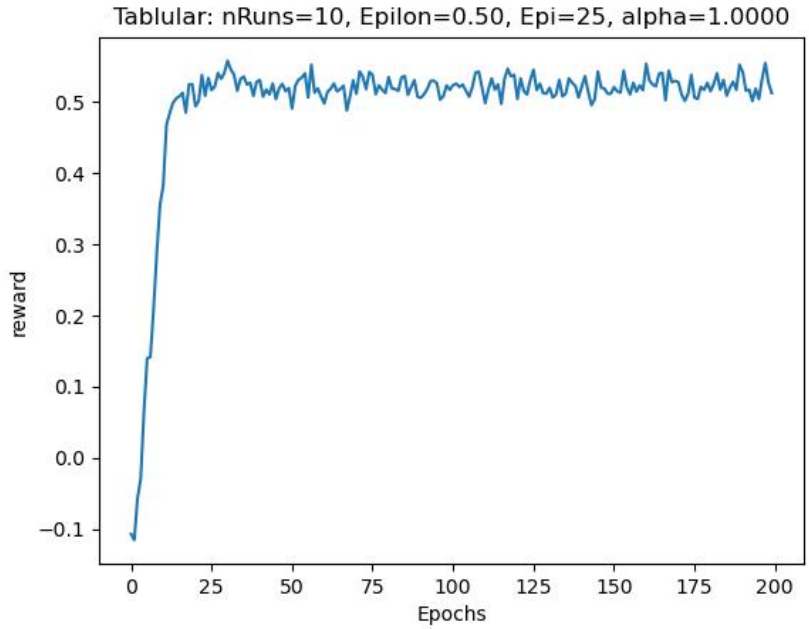
For ALPHA = 1e-6




```
Avg reward: 0.295714 | Ewma reward: 0.287749: 100%|| 200/200 [00:02<00:00, 72.1
Avg reward: 0.288622 | Ewma reward: 0.325320: 100%|| 200/200 [00:02<00:00, 70.3
Avg reward: 0.295156 | Ewma reward: 0.321001: 100%|| 200/200 [00:02<00:00, 71.7
Avg reward: 0.298885 | Ewma reward: 0.328025: 100%|| 200/200 [00:02<00:00, 70.8
Avg reward: 0.300296 | Ewma reward: 0.285677: 100%|| 200/200 [00:02<00:00, 72.3
Avg reward: 0.303919 | Ewma reward: 0.328025: 100%|| 200/200 [00:02<00:00, 71.4
Avg reward: 0.306558 | Ewma reward: 0.322751: 100%|| 200/200 [00:02<00:00, 73.8
Avg reward: 0.303000 | Ewma reward: 0.308914: 100%|| 200/200 [00:02<00:00, 72.7
Avg reward: 0.301526 | Ewma reward: 0.312553: 100%|| 200/200 [00:02<00:00, 72.2
Avg reward: 0.294090 | Ewma reward: 0.304873: 100%|| 200/200 [00:02<00:00, 72.4
```

When we set the ALPHA = 1e-6, it fluctuates a lot and does not achieve good convergence.

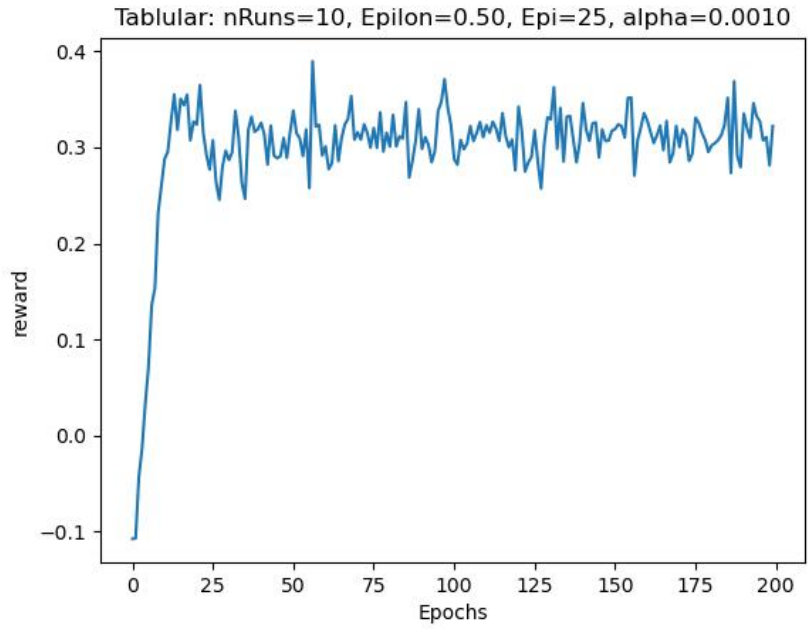
For ALPHA = 1




```
Avg reward: 0.498282 | Ewma reward: 0.527738: 100%|| 200/200 [00:01<00:00, 154.
Avg reward: 0.498095 | Ewma reward: 0.523651: 100%|| 200/200 [00:01<00:00, 153.
Avg reward: 0.497042 | Ewma reward: 0.526109: 100%|| 200/200 [00:01<00:00, 154.
Avg reward: 0.496030 | Ewma reward: 0.519801: 100%|| 200/200 [00:01<00:00, 154.
Avg reward: 0.499031 | Ewma reward: 0.522884: 100%|| 200/200 [00:01<00:00, 155.
Avg reward: 0.496270 | Ewma reward: 0.530373: 100%|| 200/200 [00:01<00:00, 156.
Avg reward: 0.499872 | Ewma reward: 0.520100: 100%|| 200/200 [00:01<00:00, 157.
Avg reward: 0.499975 | Ewma reward: 0.507520: 100%|| 200/200 [00:01<00:00, 157.
Avg reward: 0.505406 | Ewma reward: 0.536753: 100%|| 200/200 [00:01<00:00, 156.
Avg reward: 0.498136 | Ewma reward: 0.523074: 100%|| 200/200 [00:01<00:00, 153.
```

This time is better, the line is more smooth and it converges quickly.

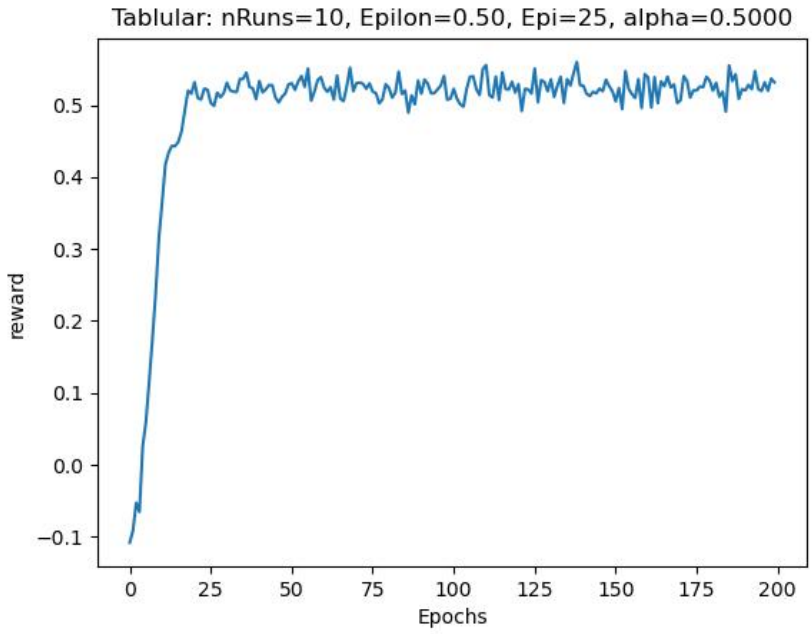
For ALPHA = 0.001



```
Avg reward: 0.302570 | Ewma reward: 0.307657: 100%|| 200/200 [00:02<00:00, 71.1
Avg reward: 0.298516 | Ewma reward: 0.313506: 100%|| 200/200 [00:02<00:00, 71.0
Avg reward: 0.297752 | Ewma reward: 0.286085: 100%|| 200/200 [00:02<00:00, 70.8
Avg reward: 0.295932 | Ewma reward: 0.322225: 100%|| 200/200 [00:02<00:00, 70.9
Avg reward: 0.302144 | Ewma reward: 0.334589: 100%|| 200/200 [00:02<00:00, 71.2
Avg reward: 0.304939 | Ewma reward: 0.312994: 100%|| 200/200 [00:02<00:00, 71.7
Avg reward: 0.306546 | Ewma reward: 0.328005: 100%|| 200/200 [00:02<00:00, 72.7
Avg reward: 0.296189 | Ewma reward: 0.299035: 100%|| 200/200 [00:02<00:00, 70.8
Avg reward: 0.296199 | Ewma reward: 0.328020: 100%|| 200/200 [00:02<00:00, 71.9
Avg reward: 0.289979 | Ewma reward: 0.309610: 100%|| 200/200 [00:02<00:00, 71.4
```

Still, the fluctuation is large.

For ALPHA = 0.5




```
Avg reward: 0.502296 | Ewma reward: 0.536700: 100%|| 200/200 [00:01<00:00, 149.
Avg reward: 0.497169 | Ewma reward: 0.529343: 100%|| 200/200 [00:01<00:00, 149.
Avg reward: 0.496694 | Ewma reward: 0.530253: 100%|| 200/200 [00:01<00:00, 152.
Avg reward: 0.495953 | Ewma reward: 0.523842: 100%|| 200/200 [00:01<00:00, 150.
Avg reward: 0.499701 | Ewma reward: 0.537597: 100%|| 200/200 [00:01<00:00, 152.
Avg reward: 0.492283 | Ewma reward: 0.522644: 100%|| 200/200 [00:01<00:00, 152.
Avg reward: 0.495175 | Ewma reward: 0.499814: 100%|| 200/200 [00:01<00:00, 150.
Avg reward: 0.497592 | Ewma reward: 0.523761: 100%|| 200/200 [00:01<00:00, 152.
Avg reward: 0.491164 | Ewma reward: 0.532357: 100%|| 200/200 [00:01<00:00, 151.
Avg reward: 0.497267 | Ewma reward: 0.533734: 100%|| 200/200 [00:01<00:00, 151.
```

This result is good.

In conclusion, the algorithm does not converge for all values of α in less than 200 epochs, also the smaller the α the slower the convergence.