Solution 2.6

Initially, all the actions have the same estimates as 5, so the algorithm randomly picks action. The estimated values are much higher than the expected rewards for few steps in the beginning and when the best bandit is selected everytime, estimated values are still optimistic due to which the agents try all the actions once. Hence, the curve oscillates in the early steps and is so until all the actions' estimates start to converge to the expected rewards.

Also, after few initial steps, the agents realize that the algorithm has tried all the actions, which are low in rewards but must have got some action with higher reward. Hence, almost all the agents tend to pick up the same action as the optimal action at the same time step. Due to this, we see the spike in the curve at some time step after initial few steps. But this action turns out to be more disappointing and hence the agents start seeing the other actions again.

With this plot, it also indicates that the optimistic greedy is not a better approach for small number of steps or runs to learn the optimal solution for any reinforcement learning problem, here the bandit arm problem.