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; hence the spike.

However, when this estimated value starts decreasing or converging to the expected rewards then all other actions with optimistic values are chosen. Hence, the curve oscillates in the early steps and is so until all the actions' estimates start to converge to the expected rewards. This indicates that the optimistic greedy is not better approach for small number of steps.