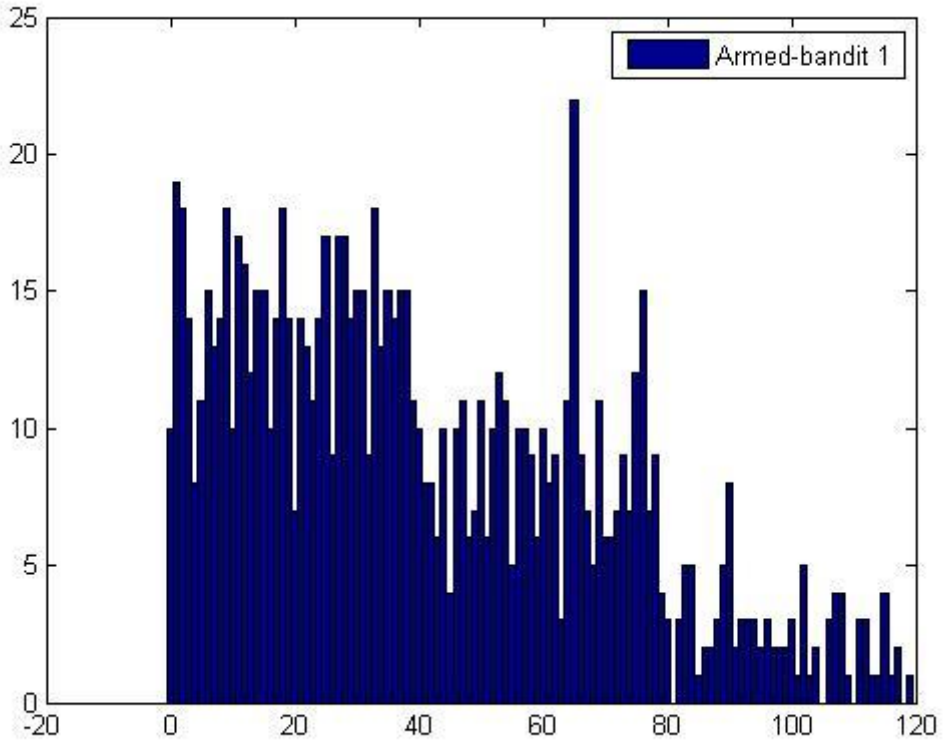
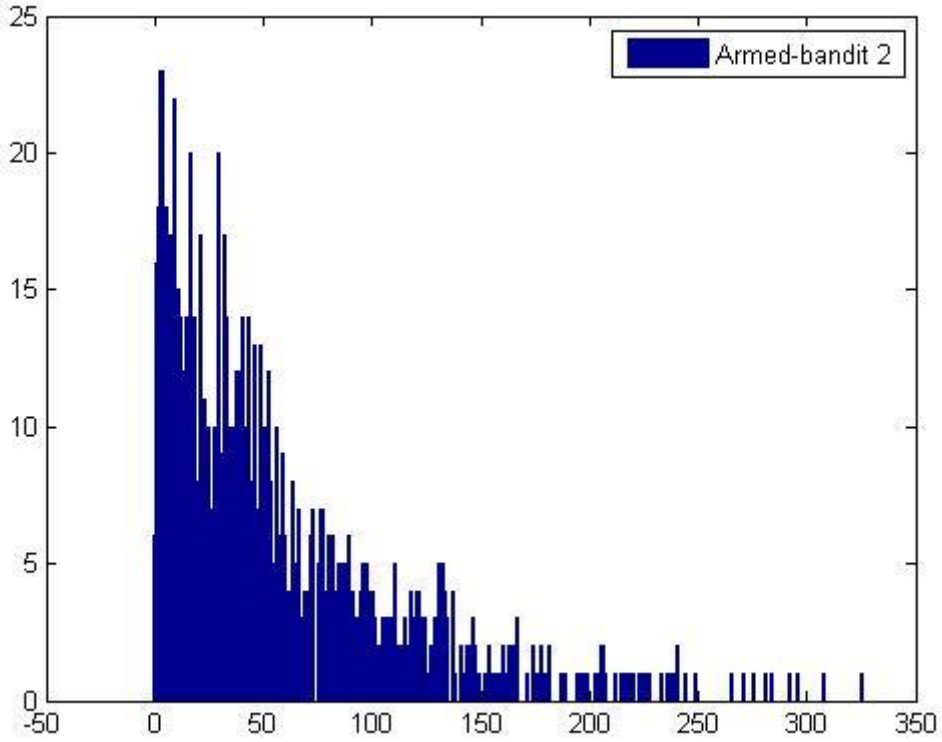
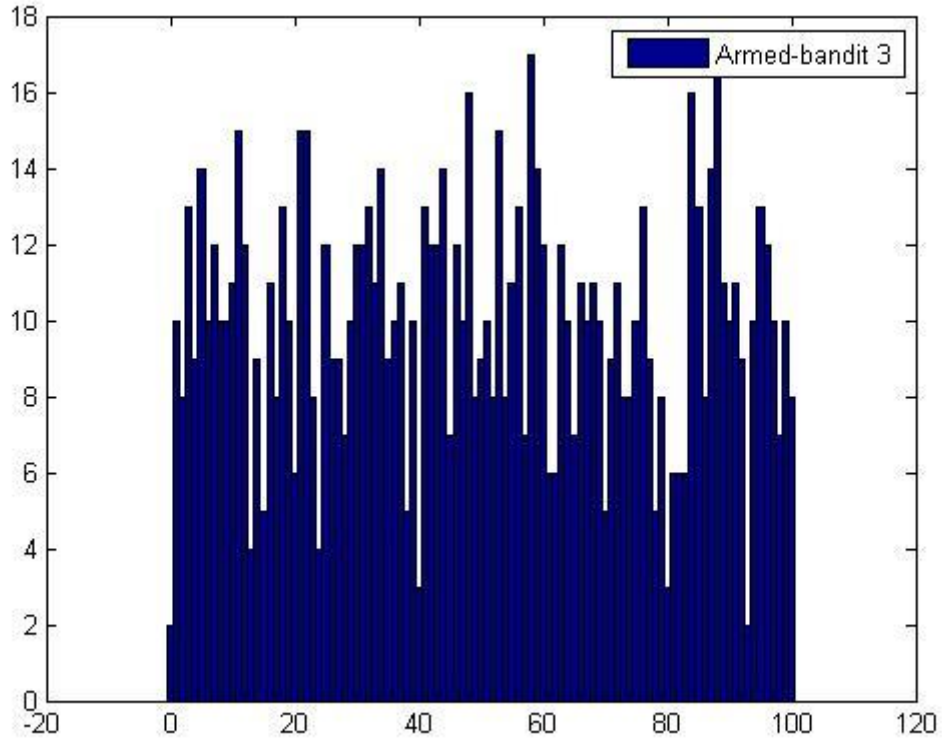
1- Probability density functions of each armed-bandit machine which are generated with Bandit function are illustrated by a histogram in figure 1. The challenging one is the last machine which most of the time produces zero rewards.




Armed-bandit 2
25
20
15
10
5
0
-50
0
50
100
150
200
250
300
350


Armed-bandit 3
18
16
14
12
10
8
6
4
2
0
-20
0
20
40
60
80
100
120

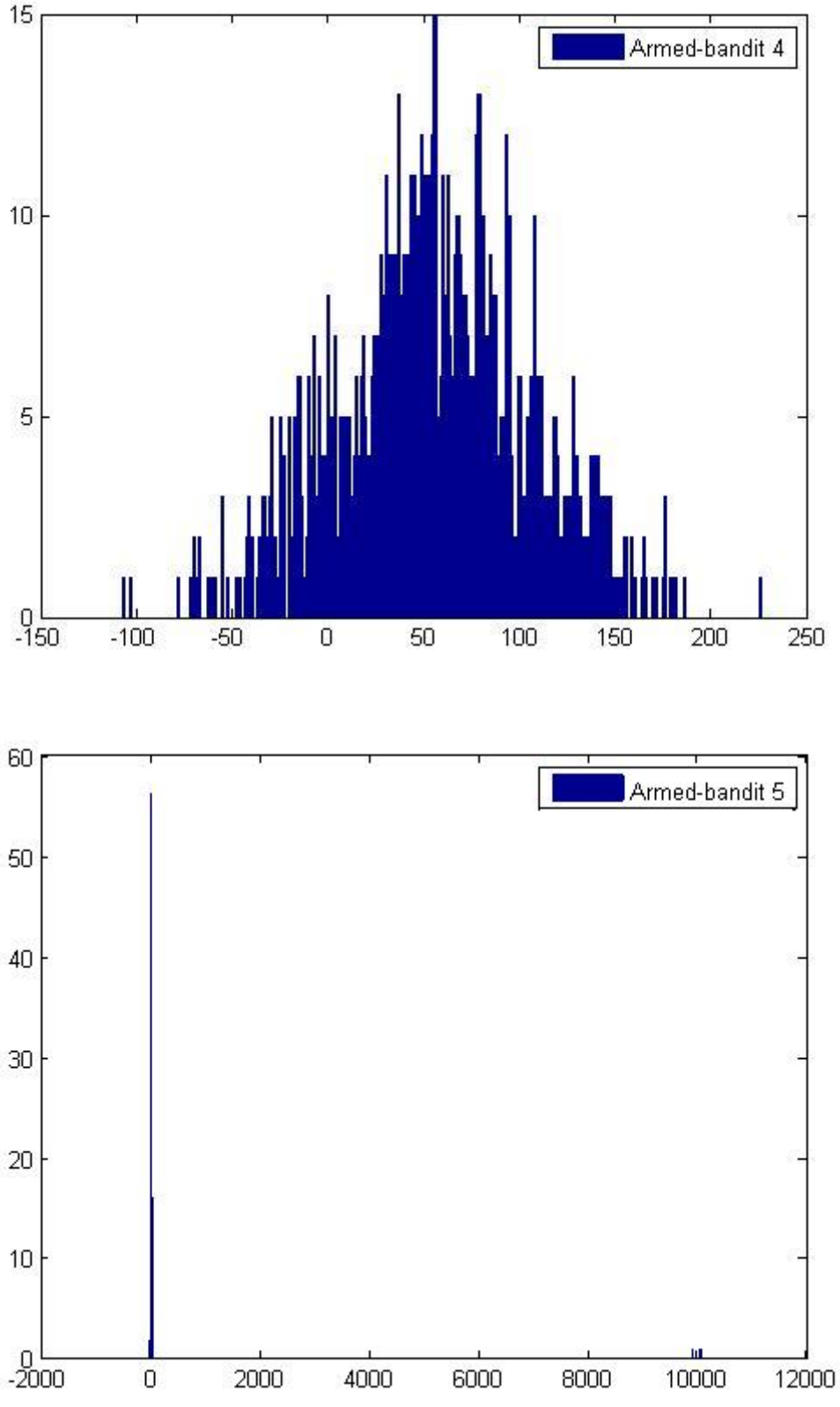
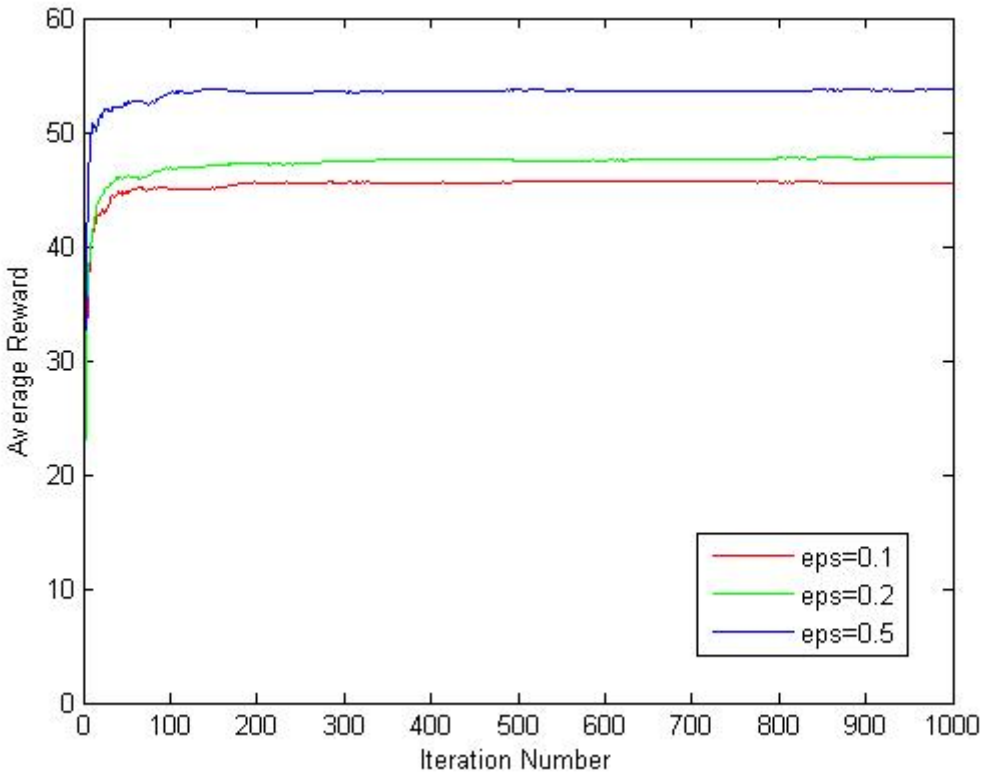
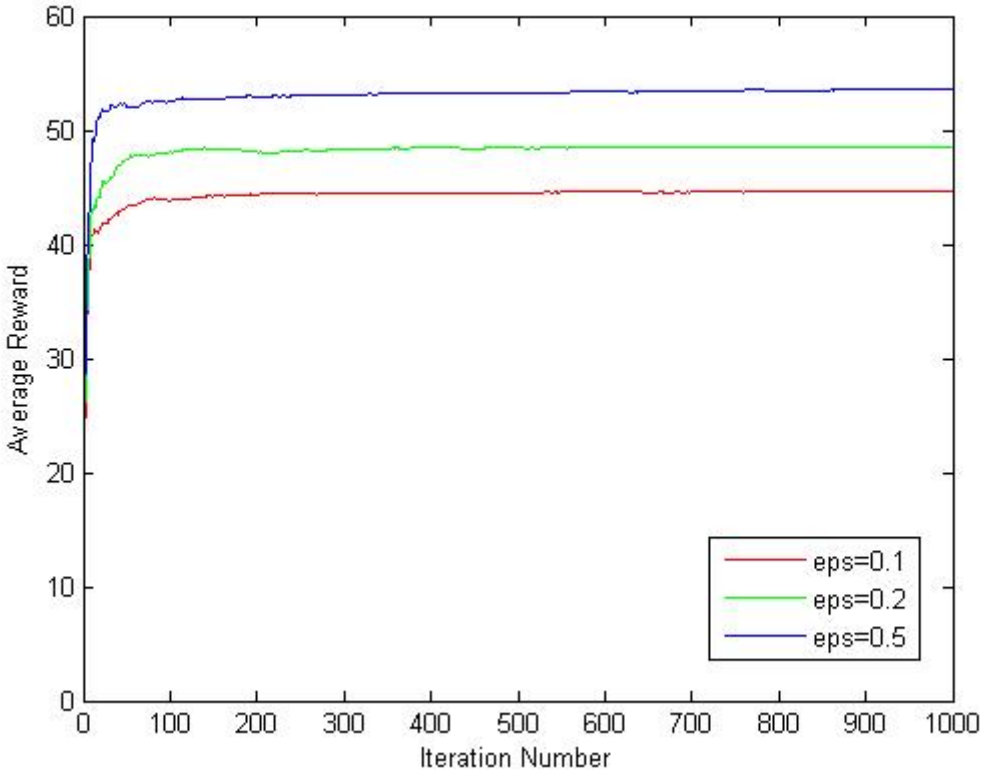


Figure 1 Histograms of armed-bandit machines

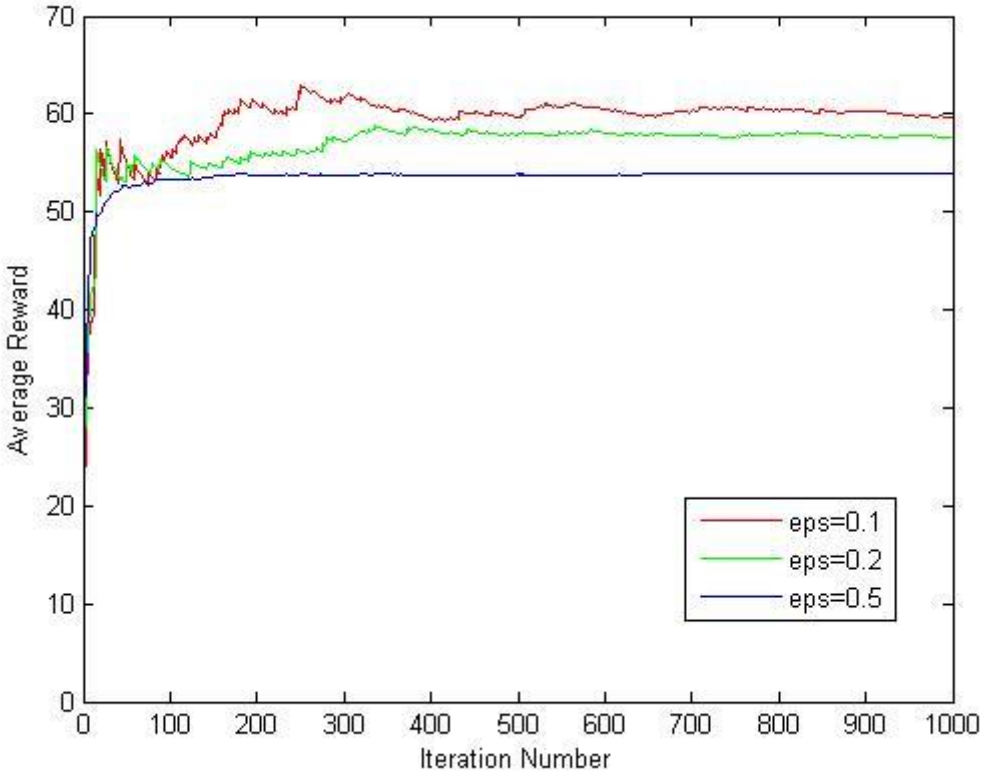
1. The results of Epsilon-greedy action selection for action-value method are illustrated in Figure 2. This problem is too dependent on the initial action value estimates. The results of initialization of Q values with random numbers from [0,100] are demonstrated in Figure 2.a. It shows that higher values of epsilon explore the solution space much more so that can find the optimal answers, whereas lower values of epsilon exploit the first best solution and therefore stuck in suboptimal actions. The results of initializing Q with 100 and zero values are shown in Figure 2.b and Figure 2.c respectively. The explanation of this happening is that the exploration of greedy algorithm is little and so that by initializing with zero values and so following a pessimistic view, the answer may fall into local minima. Conversely, if we use the optimistic policy and initialize all Q values to 100, we would encourage the action value method to explore more and so the chance of finding global minima increases.



a. Random Initialization



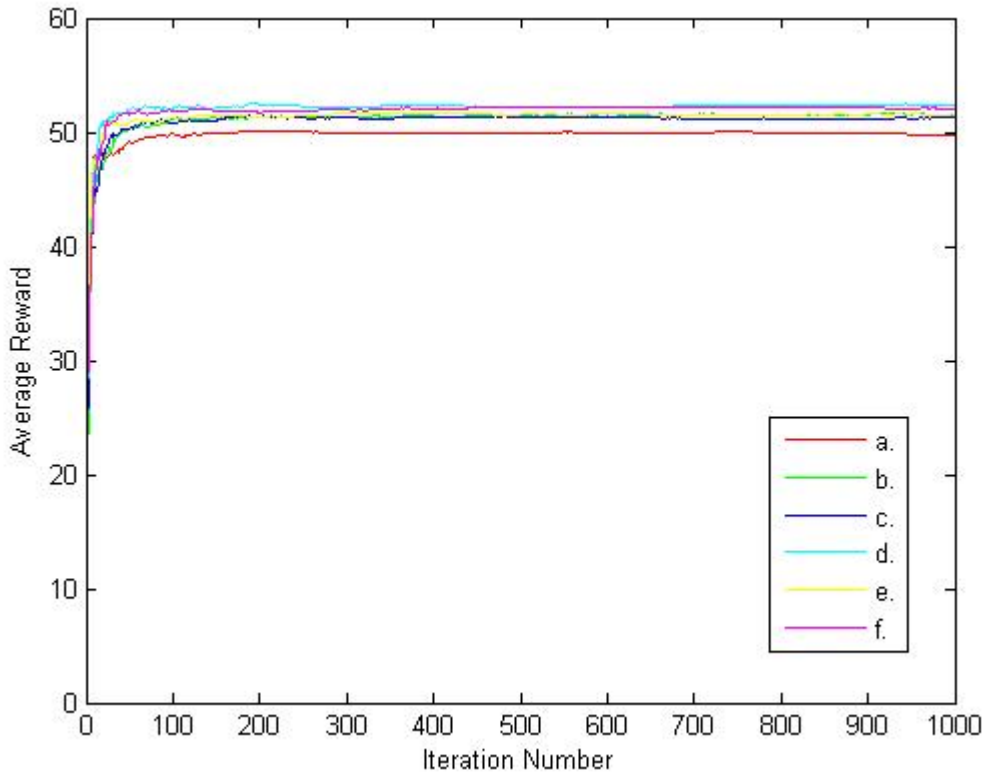
b. Optimistic initialization



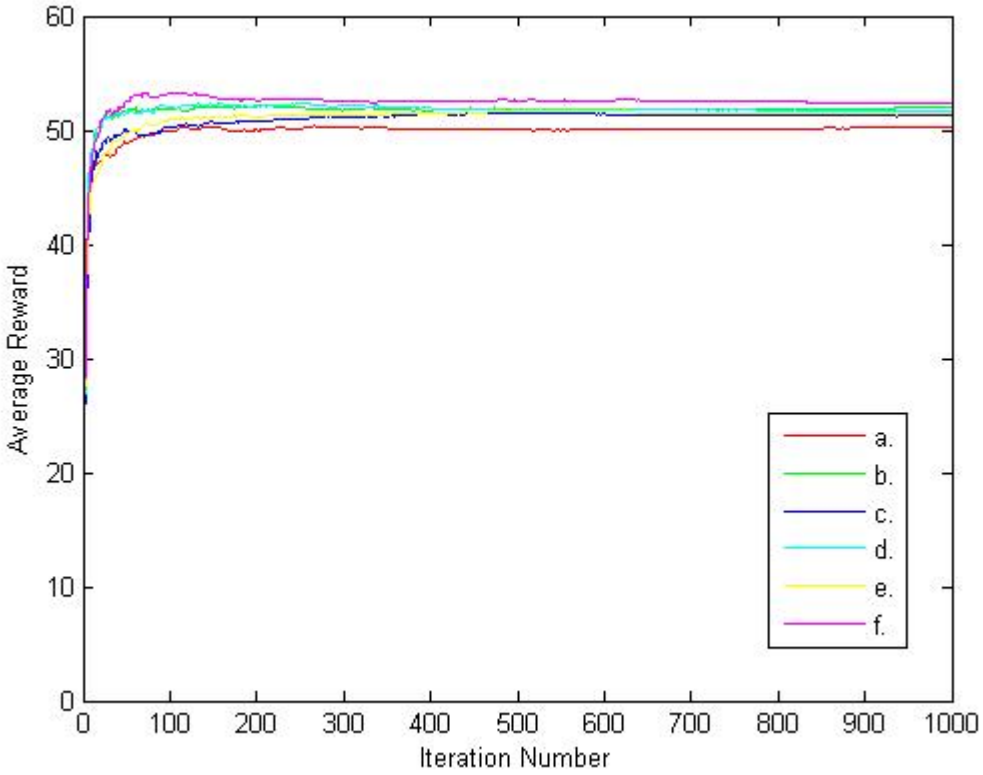
c. Pessimistic initialization

Figure 2 Epsilon-greedy method for different constant epsilons

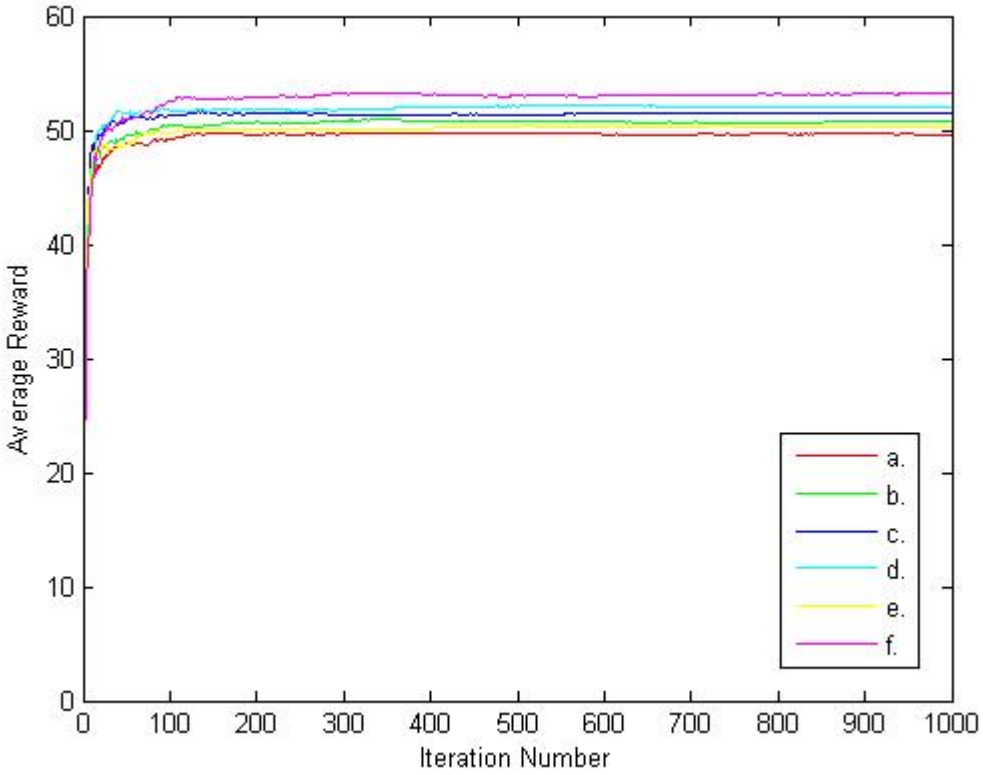
3- The results of applying different adaptive epsilons are shown in figure 3.For the best results, both choice d. and f. which are capable of decreasing epsilon smoothly and slowly over time can be taken into account. However, choice f which follows an exponential distribution dominates choice d frequently. It can be concluded that by reducing the value of epsilon gradually over time, it would be possible to explore the solution space more and therefore converge to a better solution. Again, the same discussion for initialization of Q values can be made here. This time optimistic policy produces much more satisfying answers but random initialization needs a longer time to converge to optimal solutions and so that it produces different answers for every1000 iteration.



a



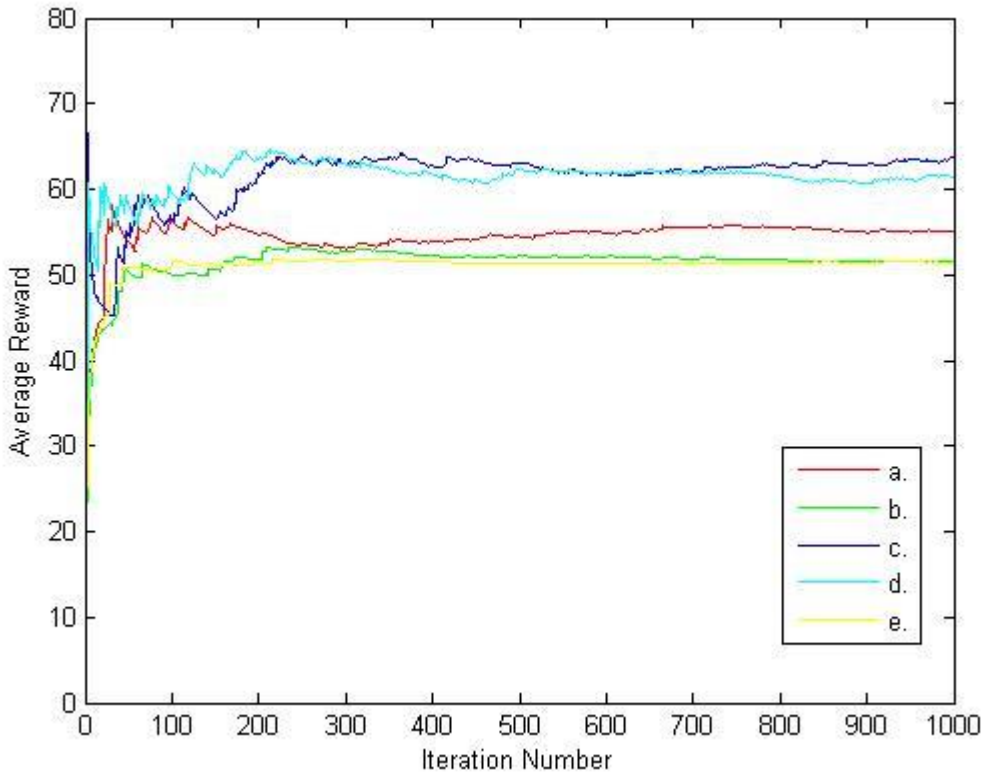
b



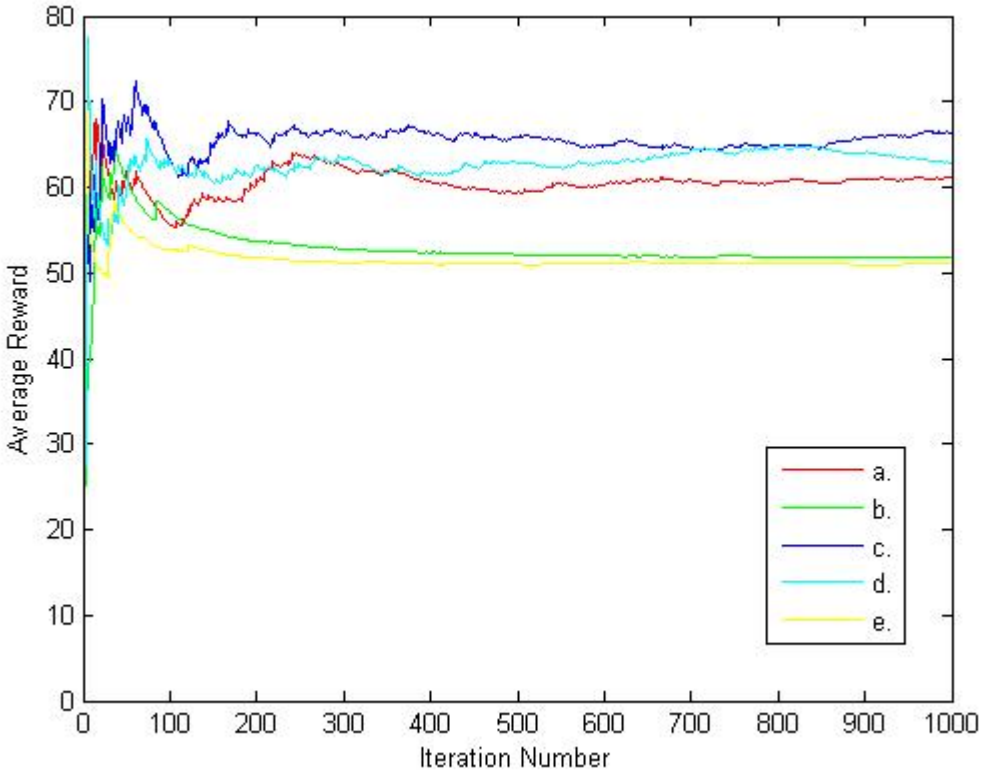
c

Figure 3 Epsilon-greedy method for different Adaptive epsilons

4. Figure 4 demonstrates the results of choosing different kinds of alpha and beta. Here, choice c and d produce nearly the same results as can be seen in Figure 4.a and Figure 4.b. However, choice c dominates choice d in most of the time, because constant alpha works better in non-stationary spaces in which the distribution of rewards changes over time. The justification of the value of beta which here showed to produce better results by reducing over time is that our preference of choosing an action as the number of selection of that action increases must be less influenced by the difference between the instant rewards the and reference reward. Because the algorithm learns the best policy over time and so that we would rely more on the updated preference values than the distance of the instant rewards and reference reward.



a



b

Figure 4 Reinforcement Comparison method for different alpha and beta

5. A comparison between the best results of part 1, part 2, and part 3 is made in figure 5. It depicts that the reinforcement comparison method can produce higher average rewards. It also improves faster than the other two methods. The privilege of this method is due to the attention to the overall reward which make it possible to make much more effective judgments for action selection. In addition, softmax action selection is better than greedy selection. The next method which is capable of producing good solution in term of fast convergence to the optimal solution is the adaptive greedy action selection method.

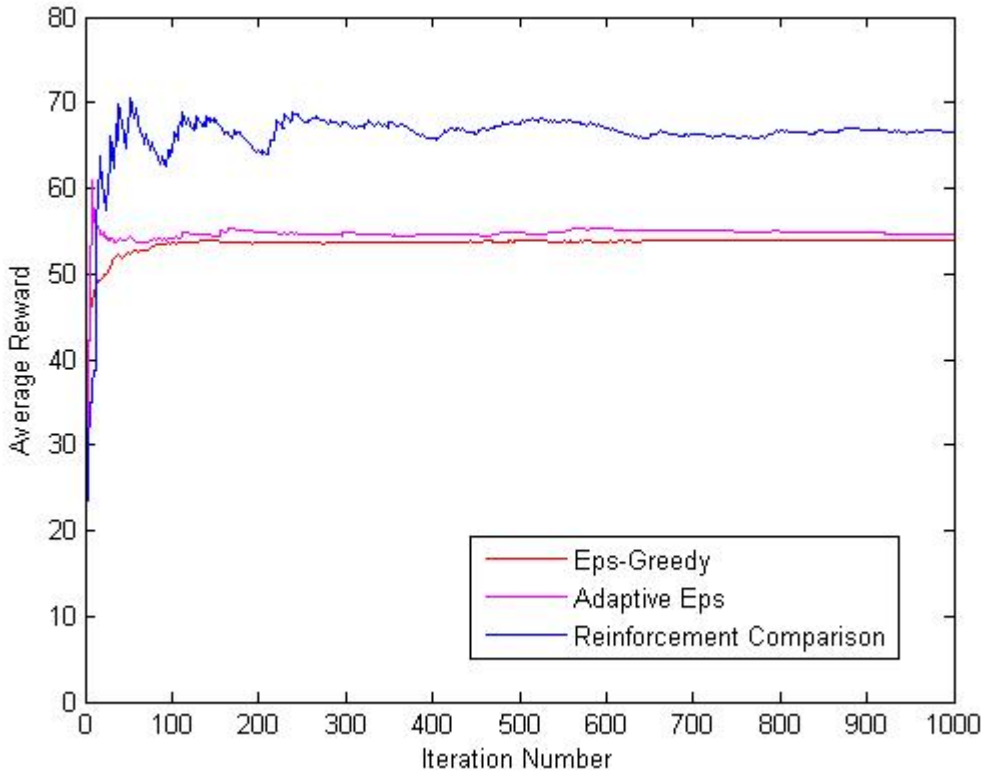


Figure 5 Comparison of the best results