Robert Learning

Assignment 1

Ming Qu 2896560

Exercise 1.1

The expected reward is 1.8

Exercise 1.2

The average reward is 1.7447
Very close to the expected reward

Exercise 1.3

The results:
Step100
a1: 8%
a2: 3%
a3: 84%
a4: 4%
a5: 1%
resulting average rewards: 2.8011
Step200
a1: 5%
a2: 2%
a3: 88.5%
a4: 2%
a5: 2.5%
resulting average rewards: 2.8166
Step300
a1: 4%
a2: 1.3333%
a3: 91.3333%
a4: 1.3333%
a5: 2%
resulting average rewards: 2.9334
Step400
a1: 3.5%
a2: 1.25%
a3: 91%
a4: 1.5%
a5: 2.75%
resulting average rewards: 2.8564
Step500
a1: 3.4%
a2: 1.2%
a3: 91.4%
a4: 1.4%
a5: 2.6%
resulting average rewards: 2.8825

```
Step600
a1:  3%
a2:  1.6667%
a3:  91.5%
a4:  1.6667%
a5:  2.1667%
resulting average rewards:  2.8731
Step700
a1:  3.4286%
a2:  1.8571%
a3:  90.8571%
a4:  1.4286%
a5:  2.4286%
resulting average rewards:  2.8809
Step800
a1:  3.125%
a2:  1.875%
a3:  91.5%
a4:  1.25%
a5:  2.25%
resulting average rewards:  2.9146
Step900
a1:  3.1111%
a2:  1.7778%
a3:  91.8889%
a4:  1.2222%
a5:  2%
resulting average rewards:  2.9134
Step1000
a1:  2.9%
a2:  1.8%
a3:  91.9%
a4:  1.5%
a5:  1.9%
resulting average rewards:  2.897
```

Exercise 1.4

Reach the optimal action very quickly. Don’t waste time on non-optimal actions.

The results:

```
Step100
a1:  88%
a2:  3%
a3:  4%
a4:  2%
a5:  3%
resulting average rewards: 2.3494
Step200
a1:  88.5%
a2:  2%
a3:  4%
a4:  2.5%
a5:  3%
resulting average rewards: 2.378
Step300
```

```
a1:  89%
a2:  2%
a3:  3%
a4:  2.6667%
a5:  3.3333%
resulting average rewards: 2.4125
Step400
a1:  87%
a2:  2.75%
a3:  3%
a4:  4%
a5:  3.25%
resulting average rewards: 2.4177
Step500
a1:  88.6%
a2:  2.2%
a3:  2.6%
a4:  3.4%
a5:  3.2%
resulting average rewards: 2.4325
Step600
a1:  89.8333%
a2:  2%
a3:  2.1667%
a4:  3%
a5:  3%
resulting average rewards: 2.4259
Step700
a1:  90.5714%
a2:  1.8571%
a3:  2.1429%
a4:  2.7143%
a5:  2.7143%
resulting average rewards: 2.4289
Step800
a1:  91.125%
a2:  1.75%
a3:  2.125%
a4:  2.375%
a5:  2.625%
resulting average rewards: 2.4583
Step900
a1:  91.3333%
a2:  1.8889%
a3:  2.2222%
a4:  2.2222%
a5:  2.3333%
resulting average rewards: 2.4608
Step1000
a1:  91.3%
a2:  1.9%
a3:  2.4%
a4:  2.1%
a5:  2.3%
resulting average rewards: 2.4699
```

Exercise 1.5:

An initial estimate of +7 is wildly optimistic. Whichever actions are initially selected, the reward is less than the starting estimates (as the results in exercise 1.4); the learner switches to other actions (action 1, 2, 4, 5), being “disappointed” with the rewards it is receiving. The result is that all actions are tried several times before the value estimates converge. The system does a fair amount of exploration even if greedy actions are selected all the time.

The results:

```
Step100
a1:  22%
a2:  14%
a3:  27%
a4:  22%
a5:  15%
resulting average rewards:  2.1358
Step200
a1:  21.5%
a2:  13%
a3:  30%
a4:  22%
a5:  13.5%
resulting average rewards:  2.1459
Step300
a1:  20.6667%
a2:  11.3333%
a3:  33.6667%
a4:  22%
a5:  12.3333%
resulting average rewards:  2.2498
Step400
a1:  23.5%
a2:  10.75%
a3:  31%
a4:  23.25%
a5:  11.5%
resulting average rewards:  2.2129
Step500
a1:  21.8%
a2:  9.4%
a3:  35%
a4:  23.6%
a5:  10.2%
resulting average rewards:  2.255
Step600
a1:  22.1667%
a2:  8.6667%
a3:  37.3333%
a4:  22.8333%
a5:  9%
resulting average rewards:  2.2858
Step700
a1:  20.7143%
```

```
a2:  7.8571%
a3:  37.8571%
a4:  25.2857%
a5:  8.2857%
resulting average rewards:  2.3268
Step800
a1:  18.625%
a2:  7.25%
a3:  44.125%
a4:  22.5%
a5:  7.5%
resulting average rewards:  2.4091
Step900
a1:  17%
a2:  6.6667%
a3:  49%
a4:  20.3333%
a5:  7%
resulting average rewards:  2.4786
Step1000
a1:  15.4%
a2:  6.2%
a3:  53.6%
a4:  18.4%
a5:  6.4%
resulting average rewards:  2.5183
```