```
366 Programming Assignment 1 Results

Eric Smith
Peter Sterling

Question 2 Results:

Deterministic policy (As printed with printPolicy()):

Average return:  -0.04248

 Usable Ace:
S H H H H S H H S S 20
S H H S S H S H S S 19
H S H H H S S H H H 18
S H H H H S H H H H 17
H H H H H H H H H H 16
H H H H H H H H H H 15
H H H H H H H H H H 14
H H H H H H H H H H 13
H H H H H H H H H H 12
1 2 3 4 5 6 7 8 9 10

 No Usable Ace:
S S S S S S S S S S 20
S S S S S S S S S S 19
S S S S S S S S S S 18
S S S S S S S S S S 17
H S S S S S S S S H 16
H H S S S H H H H H 15
H H S S H H H H H H 14
H H H S S H H H H H 13
H H H H S H H H H H 12
1 2 3 4 5 6 7 8 9 10

Question 3

We had fairly inconsistant results with many of the settings for alpha and epsilon but we found
some values that were on average better. The inconsistantcy might suggest that more learning
episodes are required or a smaller alpha if an increased number of episodes doesn't seem to be
helping (because it might be chasing noise rather than improving).

The final values we settled on were:
    - α = 0.0005
    - epsilonPi = 0
    - epsilonMu = 0.15 (learning policy explores more)
    - Number of episodes = 3 000 000 (3 000 000 runs to learn then 3 000 000 more runs to find
    the average return of the deterministic greedy policy)

epsilonMu = 0.2
epsilonPi = 0.0
alpha = 0.0005

Average return:  -0.0336803333333
Average returns for other runs with the same parameters (not corresponding to the policy below):
{-0.034194, -0.033047}

 Usable Ace:
S S S S H S S S S S 20
S S S S H S S S S S 19
S H S S H S S S H S 18
S H H H S S S H H H 17
H H H H H H H H H H 16
H H H H H H H H H H 15
H H H H H H H H H H 14
H H H H H H H H H H 13
```

```
H H H H H H H H H H 12
1 2 3 4 5 6 7 8 9 10

 No Usable Ace:
S S S S S S S S S S 20
S S S S S S S S S S 19
S S S S S S S S S S 18
S S S S S S S S S S 17
H S S S S S S S H H 16
H H S S S H H H H H 15
H H S S S H H H H H 14
H H H H H H H H H H 13
H H H H H H H H H H 12
1 2 3 4 5 6 7 8 9 10
```

An indication that we might not have enough learning episodes is that the Usable Ace side of the policy makes some irrational (though not fatal) decisions like hitting with a 20 when the dealer is showing a 8. A likely reason for these irrational policy decisions is that there is a lack of good data at each of those states (usable ace is fairly rare), therefore more episodes would help.

Another way I think the usable ace data could be improved is having the policy assume initially that the best action with an ace is to stay. The return after taking a stay action is a lot more deterministic and gives more reliable information about how good that action is. Given the small number of data points, and the action of hitting could go either way in any state with a usable ace, the data could be fairly inaccurate.