Programming Assignment 1

Name: Andriani Yunita <yunita>
Cole Evans <cevans2>

1. Implementing equiprobable-random policy

```
def equiprobable_policy(state):
    if state == 0:
        return 0
    else:
        equiprobable = random.random()
        if equiprobable > 0.5:
            return 0
        else:
            return 1
```

We ran 10 times with 2,000 episodes for each run, and we got:

1 Average return: -0.331
2 Average return: -0.3105
3 Average return: -0.3355
4 Average return: -0.3285
5 Average return: -0.3245
6 Average return: -0.337
7 Average return: -0.363
8 Average return: -0.3565
9 Average return: -0.319
10 Average return: -0.3605

Above results showed that the average return was about -0.3.

2. Expected Sarsa

For the first check, we simply ran for 100,000 episodes with $\alpha = 0.001$ and $\varepsilon_\mu = 1.0$, and we got:

Average return at step 10000 is -0.322632263226
Average return at step 20000 is -0.331566578329
Average return at step 30000 is -0.33077769259
Average return at step 40000 is -0.330658266457
Average return at step 50000 is -0.328906578132
Average return at step 60000 is -0.32718878648
Average return at step 70000 is -0.325576079658
Average return at step 80000 is -0.328104101301
Average return at step 90000 is -0.326870298559
Average return at step 100000 is -0.328963289633
Average return: -0.32896

Now we set $\varepsilon_\mu = \varepsilon_\pi = 0.01$, and ran for 10,000,000 episodes. We observed the return every 10,000 episodes. We found that the learning progresses up to about -0.085 or better. Then, we ran the deterministic learned policy by setting $\varepsilon_\mu = 0$.

(Refer to attachment p1_2.html for the full result)
Learning run result:
Average return at step 10,000 is -0.0934093409341
Average return at step 1,000,000 is -0.0511720511721
Average return at step 10,000,000 is -0.0387959038796

Usable Ace:

S	H	S	S	H	S	H	S	S	S	20
S	H	H	H	H	H	H	S	S	S	19
S	H	H	H	H	H	S	S	H	H	18
H	H	H	H	H	H	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	H	H	H	H	16
H	H	S	S	S	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	S	13
H	H	H	H	H	H	H	H	H	H	12
1	2	3	4	5	6	7	8	9	10	

The result of the deterministic learned policy:
Average return at step 10,000,000 is -0.0339631033963

Usable Ace:

S	H	S	S	H	S	H	S	S	S	20
S	H	H	H	H	H	H	S	S	S	19
S	H	H	H	H	H	S	S	H	H	18
H	H	H	H	H	H	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	H	H	H	H	16
H	H	S	S	S	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	

3. Better policy

We set number of episodes = 10,000,000, $\varepsilon_\mu = 1$, $\varepsilon_\pi = 0.01$, and $\alpha = 0.01$

(Refer to attachment p1_3.html for the full result)
The result of the deterministic learned policy:
Average return at step 10,000,000 is -0.0284725028473

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	H	S	18
H	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	H	H	H	H	16
H	S	H	S	S	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	