Blackjack - results

All policies were tested on 10000 random episodes.

1 Deterministic optimal policy

```
####################
Deterministic policy: GameStats(wins=0.4298, draws=0.0962, losses=0.474)
```

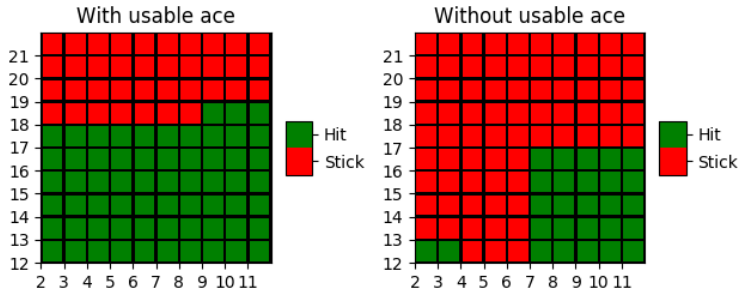

Figure 1: Deterministic policy

2 Monte Carlo Exploring States

```
####################
MonteCarloExploringStates (50000 epochs, 4.62s):
    GameStats(wins=0.4291, draws=0.096, losses=0.4749)
MonteCarloExploringStates (500000 epochs, 47.48s):
    GameStats(wins=0.4327, draws=0.0889, losses=0.4784)
```

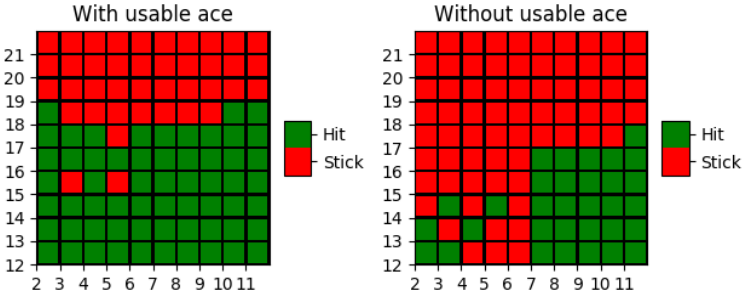

Figure 2: Policy after 50k episodes.

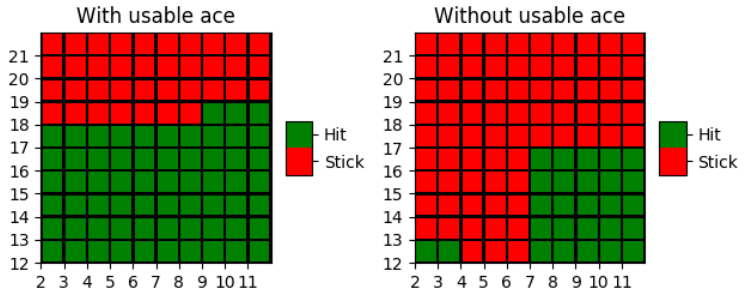

Figure 3: Policy after 500k episodes.

3 On-policy first visit Monte Carl control

```
####################
MonteCarloOnPolicyFirstVisit (50000 epochs, 5.85s):
    GameStats(wins=0.3981, draws=0.0747, losses=0.5272) eps: 0.2
MonteCarloOnPolicyFirstVisit (50000 epochs, 5.86s):
    GameStats(wins=0.3993, draws=0.0873, losses=0.5134) eps: 0.1
MonteCarloOnPolicyFirstVisit (50000 epochs, 5.71s):
    GameStats(wins=0.4127, draws=0.0838, losses=0.5035) eps: 0.05
MonteCarloOnPolicyFirstVisit (500000 epochs, 57.42s):
    GameStats(wins=0.3995, draws=0.0752, losses=0.5253) eps: 0.2
MonteCarloOnPolicyFirstVisit (500000 epochs, 58.44s):
    GameStats(wins=0.4169, draws=0.0862, losses=0.4969) eps: 0.1
MonteCarloOnPolicyFirstVisit (500000 epochs, 57.97s):
    GameStats(wins=0.4161, draws=0.0848, losses=0.4991) eps: 0.05
MonteCarloOnPolicyFirstVisit (2000000 epochs, 239.49s):
    GameStats(wins=0.4067, draws=0.0802, losses=0.5131) eps: 0.2
MonteCarloOnPolicyFirstVisit (2000000 epochs, 238.11s):
    GameStats(wins=0.4108, draws=0.0892, losses=0.5) eps: 0.1
MonteCarloOnPolicyFirstVisit (2000000 epochs, 225.79s):
    GameStats(wins=0.4197, draws=0.0928, losses=0.4875) eps: 0.05
```

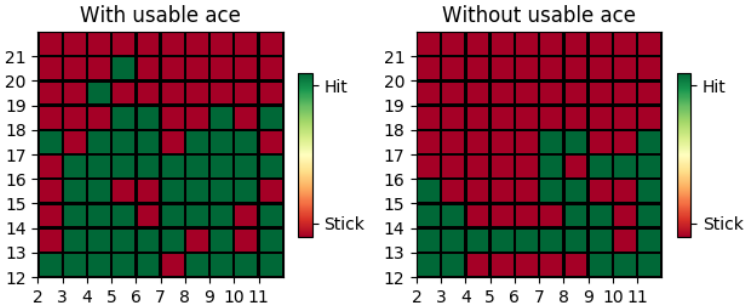

Figure 4: Policy after 50k episodes for $\epsilon = 0.05$ (5% chance of taking action with worse value)

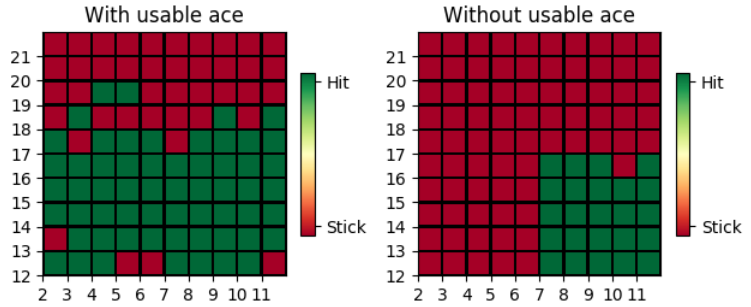

Figure 5: Policy after 500k episodes for $\epsilon = 0.05$ (5% chance of taking action with worse value)

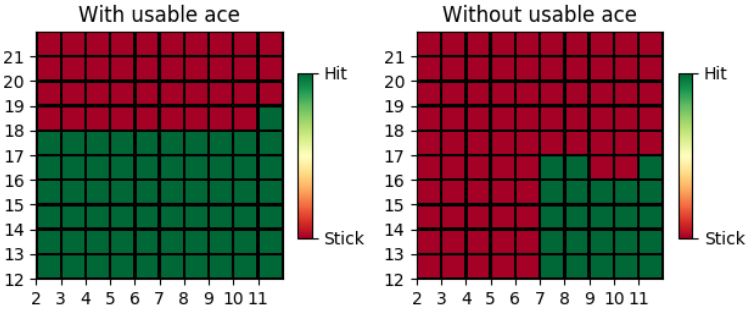

Figure 6: Policy after 2M episodes for $\epsilon = 0.05$ (5% chance of taking action with worse value)

4 Sarsa

To obtain exploring policy while keeping it updated to the computed values, episodes were played with $\epsilon - greedy$ policy with $\epsilon = 0.1$.

```
####################
Sarsa (50000 epochs, 5.29s):
    GameStats(wins=0.4004, draws=0.0792, losses=0.5204) alpha: 0.01 gamma: 0.5
Sarsa (50000 epochs, 5.16s):
    GameStats(wins=0.3957, draws=0.08, losses=0.5243) alpha: 0.01 gamma: 0.9
Sarsa (50000 epochs, 5.19s):
    GameStats(wins=0.3952, draws=0.0758, losses=0.529) alpha: 0.01 gamma: 0.99
Sarsa (50000 epochs, 5.39s):
    GameStats(wins=0.3924, draws=0.0788, losses=0.5288) alpha: 0.1 gamma: 0.5
Sarsa (50000 epochs, 5.22s):
    GameStats(wins=0.3876, draws=0.0781, losses=0.5343) alpha: 0.1 gamma: 0.9
Sarsa (50000 epochs, 5.27s):
    GameStats(wins=0.4007, draws=0.081, losses=0.5183) alpha: 0.1 gamma: 0.99
Sarsa (50000 epochs, 5.49s):
    GameStats(wins=0.3754, draws=0.0763, losses=0.5483) alpha: 0.5 gamma: 0.5
Sarsa (50000 epochs, 5.37s):
    GameStats(wins=0.3787, draws=0.0752, losses=0.5461) alpha: 0.5 gamma: 0.9
Sarsa (50000 epochs, 5.27s):
    GameStats(wins=0.3824, draws=0.0724, losses=0.5452) alpha: 0.5 gamma: 0.99
Sarsa (500000 epochs, 54.75s):
    GameStats(wins=0.4025, draws=0.0834, losses=0.5141) alpha: 0.01 gamma: 0.5
Sarsa (500000 epochs, 53.13s):
    GameStats(wins=0.4003, draws=0.081, losses=0.5187) alpha: 0.01 gamma: 0.9
Sarsa (500000 epochs, 52.52s):
    GameStats(wins=0.3999, draws=0.0803, losses=0.5198) alpha: 0.01 gamma: 0.99
Sarsa (500000 epochs, 54.53s):
```

```
    GameStats(wins=0.3857, draws=0.0803, losses=0.534) alpha: 0.1 gamma: 0.5
Sarsa (500000 epochs, 53.81s):
    GameStats(wins=0.3903, draws=0.0812, losses=0.5285) alpha: 0.1 gamma: 0.9
Sarsa (500000 epochs, 53.97s):
    GameStats(wins=0.3922, draws=0.0848, losses=0.523) alpha: 0.1 gamma: 0.99
Sarsa (500000 epochs, 55.92s):
    GameStats(wins=0.3778, draws=0.0712, losses=0.551) alpha: 0.5 gamma: 0.5
Sarsa (500000 epochs, 52.38s):
    GameStats(wins=0.3947, draws=0.0799, losses=0.5254) alpha: 0.5 gamma: 0.9
Sarsa (500000 epochs, 53.03s):
    GameStats(wins=0.3796, draws=0.0718, losses=0.5486) alpha: 0.5 gamma: 0.99
Sarsa (4000000 epochs, 411.17s):
    GameStats(wins=0.4177, draws=0.0778, losses=0.5245) alpha: 0.01 gamma: 0.5
Sarsa (4000000 epochs, 414.64s):
    GameStats(wins=0.3979, draws=0.0849, losses=0.5172) alpha: 0.01 gamma: 0.9
Sarsa (4000000 epochs, 420.89s):
    GameStats(wins=0.3913, draws=0.0876, losses=0.5211) alpha: 0.01 gamma: 0.99
Sarsa (4000000 epochs, 421.86s):
    GameStats(wins=0.4055, draws=0.0765, losses=0.528) alpha: 0.1 gamma: 0.5
Sarsa (4000000 epochs, 412.72s):
    GameStats(wins=0.396, draws=0.0851, losses=0.5189) alpha: 0.1 gamma: 0.9
Sarsa (4000000 epochs, 412.23s):
    GameStats(wins=0.386, draws=0.0737, losses=0.5403) alpha: 0.1 gamma: 0.99
Sarsa (4000000 epochs, 444.25s):
    GameStats(wins=0.4065, draws=0.0769, losses=0.5366) alpha: 0.5 gamma: 0.5
Sarsa (4000000 epochs, 434.89s):
    GameStats(wins=0.3813, draws=0.0833, losses=0.5354) alpha: 0.5 gamma: 0.9
Sarsa (4000000 epochs, 415.12s):
    GameStats(wins=0.3923, draws=0.0735, losses=0.5342) alpha: 0.5 gamma: 0.99
```

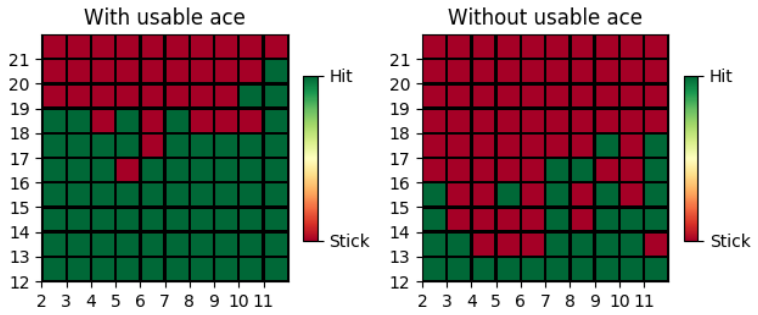

Figure 7: Policy after 50k episodes for $\alpha = 0.01,\ \gamma = 0.99$ (10% chances of taking action with worse value)

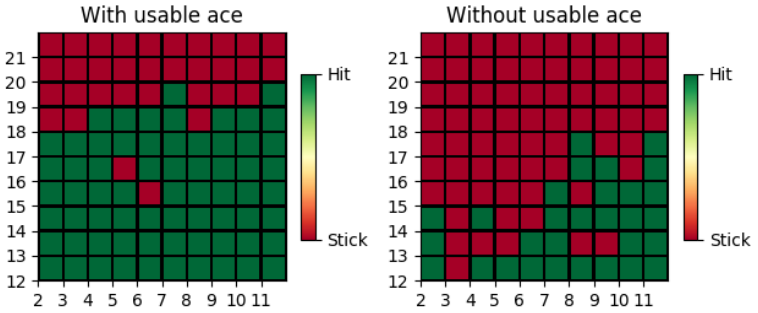

Figure 8: Policy after 500K episodes for $\alpha = 0.01,\ \gamma = 0.99$ (10% chances of taking action with worse value)

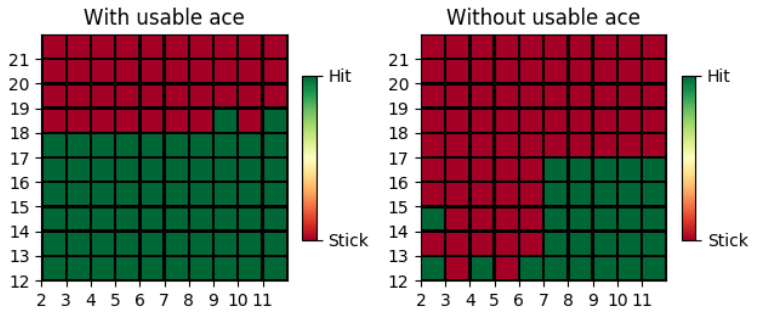

Figure 9: Policy after 2M episodes for $\alpha = 0.01,\ \gamma = 0.99$ (10% chances of taking action with worse value)

5 QLearning

Random choice was used as the behavioral policy.

```
####################
QLearning (50000 epochs, 4.63s):
    GameStats(wins=0.4298, draws=0.0874, losses=0.4828) alpha: 0.01 gamma: 0.5
QLearning (50000 epochs, 4.43s):
    GameStats(wins=0.4239, draws=0.0886, losses=0.4875) alpha: 0.01 gamma: 0.9
QLearning (50000 epochs, 4.55s):
    GameStats(wins=0.4436, draws=0.0892, losses=0.4672) alpha: 0.01 gamma: 0.99
QLearning (50000 epochs, 4.48s):
    GameStats(wins=0.4162, draws=0.085, losses=0.4988) alpha: 0.1 gamma: 0.5
QLearning (50000 epochs, 4.51s):
    GameStats(wins=0.4216, draws=0.0913, losses=0.4871) alpha: 0.1 gamma: 0.9
QLearning (50000 epochs, 4.72s):
    GameStats(wins=0.4204, draws=0.0878, losses=0.4918) alpha: 0.1 gamma: 0.99
QLearning (50000 epochs, 4.81s):
    GameStats(wins=0.4168, draws=0.0822, losses=0.501) alpha: 0.5 gamma: 0.5
QLearning (50000 epochs, 4.81s):
    GameStats(wins=0.401, draws=0.0683, losses=0.5307) alpha: 0.5 gamma: 0.9
QLearning (50000 epochs, 4.79s):
    GameStats(wins=0.3975, draws=0.0808, losses=0.5217) alpha: 0.5 gamma: 0.99
QLearning (500000 epochs, 47.90s):
    GameStats(wins=0.4198, draws=0.0925, losses=0.4877) alpha: 0.01 gamma: 0.5
QLearning (500000 epochs, 48.65s):
    GameStats(wins=0.4169, draws=0.0927, losses=0.4904) alpha: 0.01 gamma: 0.9
QLearning (500000 epochs, 48.54s):
    GameStats(wins=0.4276, draws=0.096, losses=0.4764) alpha: 0.01 gamma: 0.99
QLearning (500000 epochs, 48.92s):
    GameStats(wins=0.4291, draws=0.0894, losses=0.4815) alpha: 0.1 gamma: 0.5
QLearning (500000 epochs, 48.18s):
    GameStats(wins=0.4099, draws=0.0875, losses=0.5026) alpha: 0.1 gamma: 0.9
QLearning (500000 epochs, 46.76s):
    GameStats(wins=0.4257, draws=0.0887, losses=0.4856) alpha: 0.1 gamma: 0.99
QLearning (500000 epochs, 47.45s):
    GameStats(wins=0.4046, draws=0.0776, losses=0.5178) alpha: 0.5 gamma: 0.5
QLearning (500000 epochs, 48.33s):
    GameStats(wins=0.4019, draws=0.0804, losses=0.5177) alpha: 0.5 gamma: 0.9
QLearning (500000 epochs, 48.20s):
    GameStats(wins=0.4039, draws=0.0835, losses=0.5126) alpha: 0.5 gamma: 0.99
QLearning (2000000 epochs, 173.71s):
    GameStats(wins=0.4246, draws=0.0893, losses=0.4861) alpha: 0.01 gamma: 0.5
QLearning (2000000 epochs, 180.44s):
```

```
    GameStats(wins=0.4327, draws=0.0917, losses=0.4756) alpha: 0.01 gamma: 0.9
QLearning (2000000 epochs, 182.71s):
    GameStats(wins=0.4325, draws=0.0937, losses=0.4738) alpha: 0.01 gamma: 0.99
QLearning (2000000 epochs, 178.15s):
    GameStats(wins=0.423, draws=0.0843, losses=0.4927) alpha: 0.1 gamma: 0.5
QLearning (2000000 epochs, 173.01s):
    GameStats(wins=0.4185, draws=0.0885, losses=0.493) alpha: 0.1 gamma: 0.9
QLearning (2000000 epochs, 185.81s):
    GameStats(wins=0.427, draws=0.0898, losses=0.4832) alpha: 0.1 gamma: 0.99
QLearning (2000000 epochs, 183.98s):
    GameStats(wins=0.3995, draws=0.0814, losses=0.5191) alpha: 0.5 gamma: 0.5
QLearning (2000000 epochs, 175.78s):
    GameStats(wins=0.4076, draws=0.0847, losses=0.5077) alpha: 0.5 gamma: 0.9
QLearning (2000000 epochs, 168.81s):
    GameStats(wins=0.403, draws=0.0881, losses=0.5089) alpha: 0.5 gamma: 0.99
```

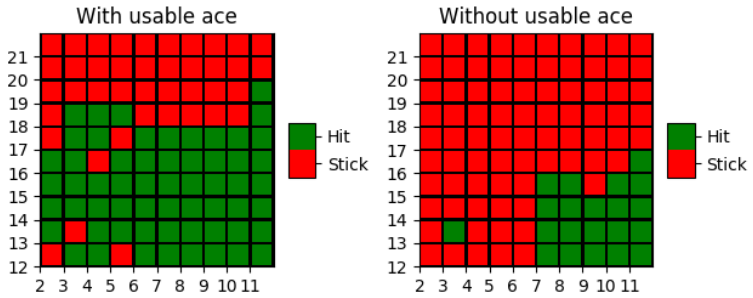

Figure 10: Policy after 50k episodes for $\alpha = 0.01,\ \gamma = 0.99$

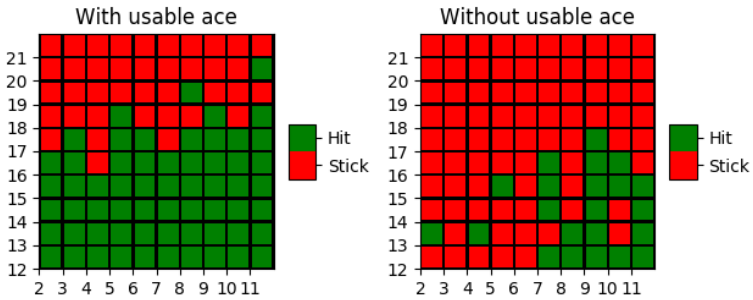

Figure 11: Policy after 500k episodes for $\alpha = 0.01,\ \gamma = 0.99$

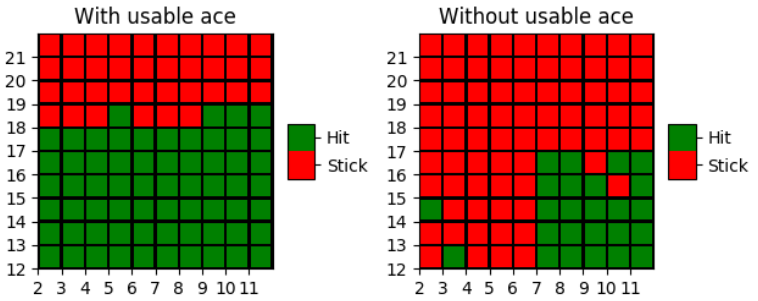

Figure 12: Policy after 2M episodes for $\alpha = 0.01,\ \gamma = 0.99$