Lund University
Faculty of Engineering (LTH)

EDAN95

Applied Machine Learning

Project 7

Author:
Marcel Attar, 941127-2173

The report was handed in on: December 19, 2019

Blackjack, Exercise in Reinforcement Learning

You will write a report of about two pages on your experiments:

Part 1

You will describe the program you wrote and plot the graph showing your optimal policy. With how many episodes and how many policy improvement cycles was your result achieved?

I first implemented the function *make_epsilon_greedy_policy* which returns a function *policy_fn* that spits out an array of length nA that is the number of actions given a certain state. The array values represents the probability that we will chose a certain action (we won't always pick the "best one"), the different indices are the different actions. The best action (the greedy action) gets the value $1 - \epsilon + \frac{\epsilon}{nA}$ and the remaining actions (the non-greedy ones) gets the value $\frac{\epsilon}{nA}$.

Then the main function was implemented, *mc_control_epsilon_greedy*. The pseudocode from [1, p.99] was used as reference for the implementation. This yielded the following results:

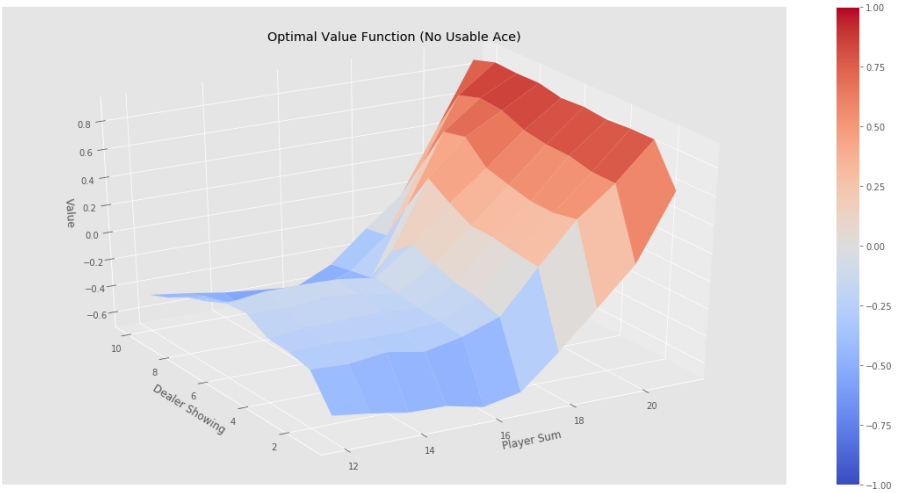

Figure 1: The optimal value function with no usable ace after 500 000 episodes.

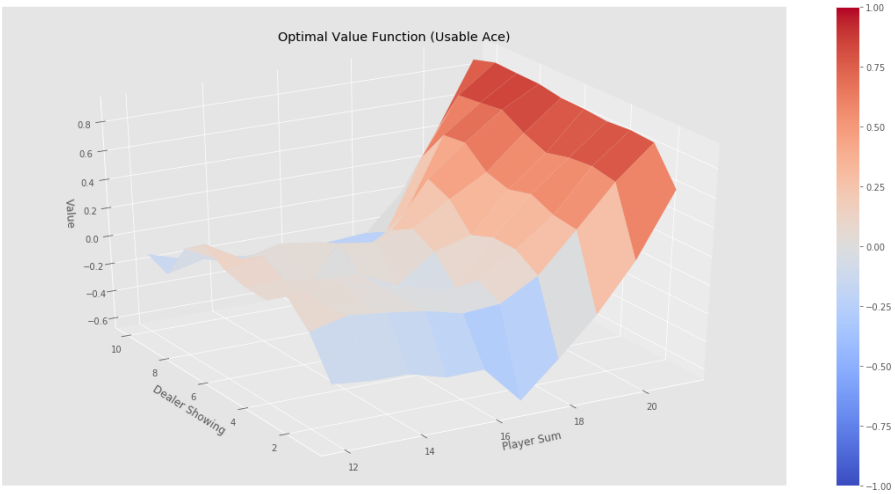

Figure 2: The optimal value function with an usable ace after 500 000 episodes.

This result was achieved after 500 000 episodes. As can be seen in fig. 1 the break even point (value 0) for no usable ace is reached if the player sum i 16-17 for most of the dealers cards. If the player has an usable ace, as in fig. 2, the break even point is reached for a lot more states, i.e. it's a stronger hand in general.

The number of policy improvement cycles is the same as the amount of times that Q is improved since the policy is then improved implicitly. This happened 630 846 times.

Part 2

Please read chapters 5-5.4 in the book "Reinforcement Learning, 2nd ed." by Sutton and Barto and discuss the following questions.

1. *Why do you have to use the Q function instead of the Value function? for MC control?*

2. *Sect 5.3 talked about MC with exploring starts. We have not talked about exploring starts in the lecture. Why are exploring starts important, what is the problem with them and why can we omit them when using greedy MC?*

Answer 1: The value function requires you to know the immediate reward R_s^a and the transition probability $P_{ss'}^a$ since the policy is improved by

$$\pi'(s) = \arg\max_{a \in A} \left(R_s^a + P_{ss'}^a v_k \left(s' \right) \right) \tag{1}$$

However, if those two variables are unknown (there's no model) we have to use the Q function. The policy is then improved by

$$\pi'(s) = \arg\max_{a \in A} Q(s, a) \tag{2}$$

Monte Carlo control this is the case, it is model free and we therefore can't use the value function.

Answer 2: Exploring starts are important because it guaranties that the policy reaches the optimum policy. In order for our policy to update closer to the optimum for each iteration we need to ensure that all actions are selected infinitely often, this can be done by using exploring starts. However, since we use the ε-greedy we chose the non-optimum action with probability $\frac{\varepsilon}{|\mathcal{A}(s)|}$ and therefore our algorithm has exploration built into it, making the exploring start unnecessary.

The problem with exploring starts is that it requires hugh amount of episodes, more than when you are using ε-greedy. This is because you have to start in *all* states an infinite amount of time to reach the optimum policy.

References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL: `http://incompleteideas.net/book/the-book-2nd.html`.