Monte Carlo Methods in Reinforcement Learning Technical Report

Riashat Islam
McGill University
Reasoning and Learning Lab
riashat.islam@cs.mcgill.ca

February 3, 2017

1 Introduction

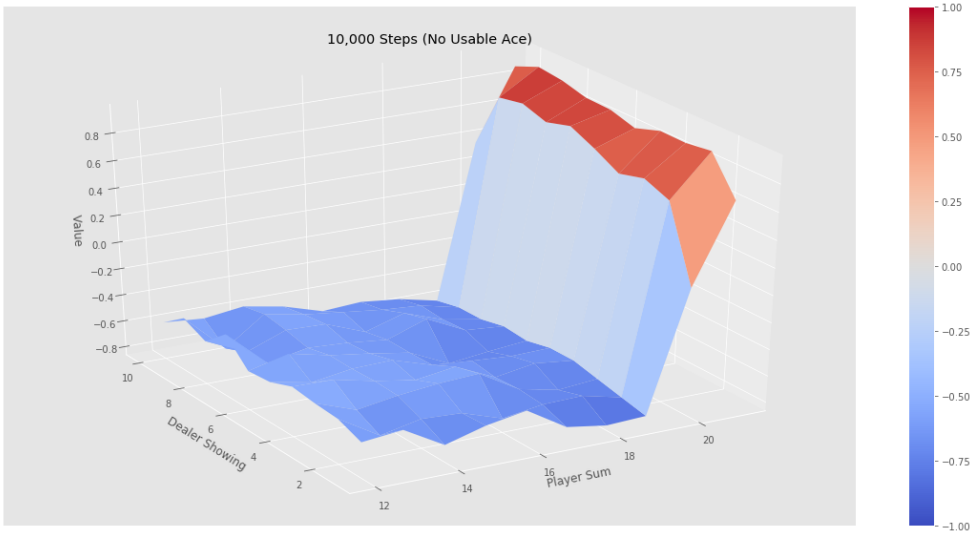
In this work, we analyse the key differences between on-policy and off-policy monte-carlo methods, using the Blackjack game environment as our MDP simulation test bed. We provide experimental demonstrations and key insights to the following methods, and explore variations in MC methods. Our contributions are to analyse the significance of different behaviour policies to be used with importance sampling for variance reduction in off-policy MC methods. We also provide demonstrations of using a control variate (baseline) in the off-policy MC methods. Experimental results are provided for the following:

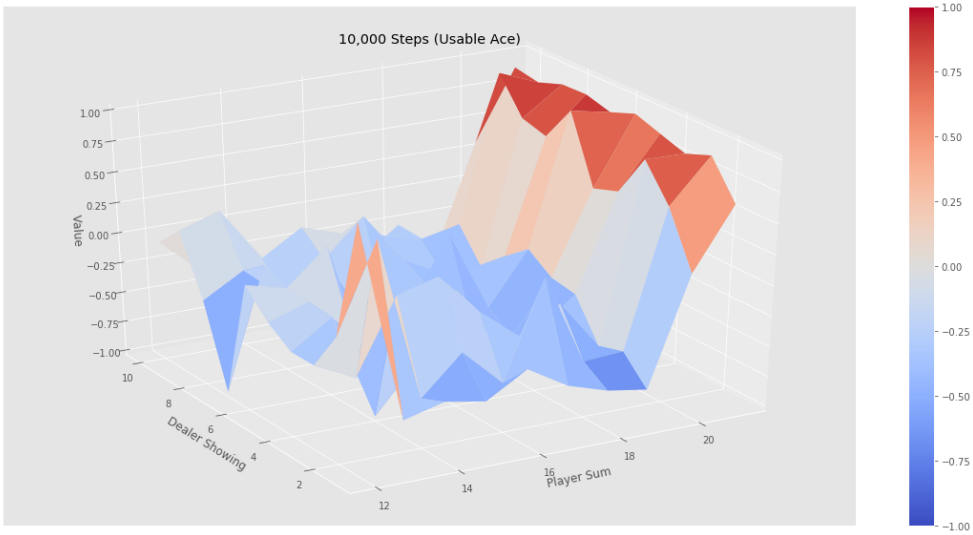
- On-Policy Monte-Carlo for Prediction
- On-Policy Monte-Carlo Control
- Off-Policy Monte-Carlo with Importance Weighted Sampling, using a Random behaviour policy Off-Policy Monte-Carlo with Importance Weighted Sampling, using an ϵ greedy off-policy as well
- Off-Policy Monte-Carlo with Importance Weighted Sampling, using a Boltzmann Policy Off-Policy Monte-Carlo with Importance Weighted Sampling, using a Sigmoid policy (exploratory analysis)
- Variance Reduction Technique : Estimate an Advantage Function with Importance Sampling in Off-Policy MC
- Off-Policy MC with arbitrary and fixed baseline values for variance reduction (control variate)
- Idea : Off-Policy Monte-Carlo Control - updating Q function estimate based on One-Step TD, combined with using importance sampling (using epsilon-greedy behaviour policy). Instead of updating Q towards the mean return in MC, what if we use one step TD and still use importance sampling?

2 Experimental Results

2.1 On-Policy Monte-Carlo Prediction

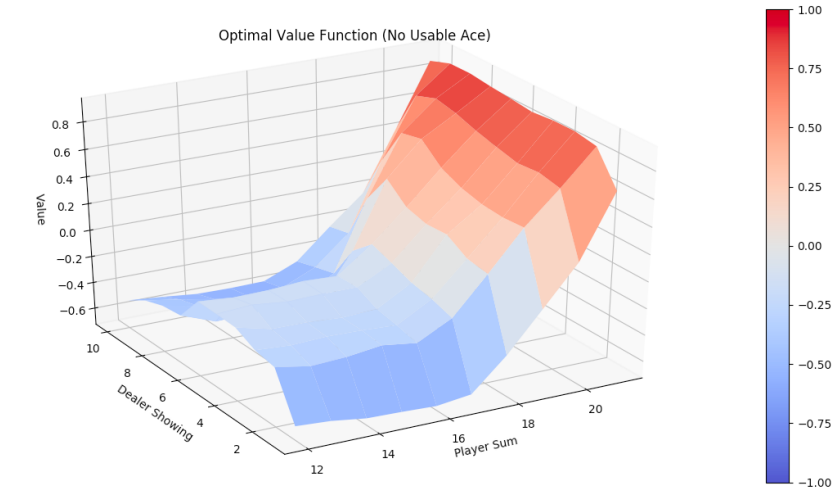
First, we demonstrate the use of Monte-Carlo methods to directly estimate value functions. This inolves using the mean return of G to directly estimate V. **Code:** See $mc\_prediction.py$

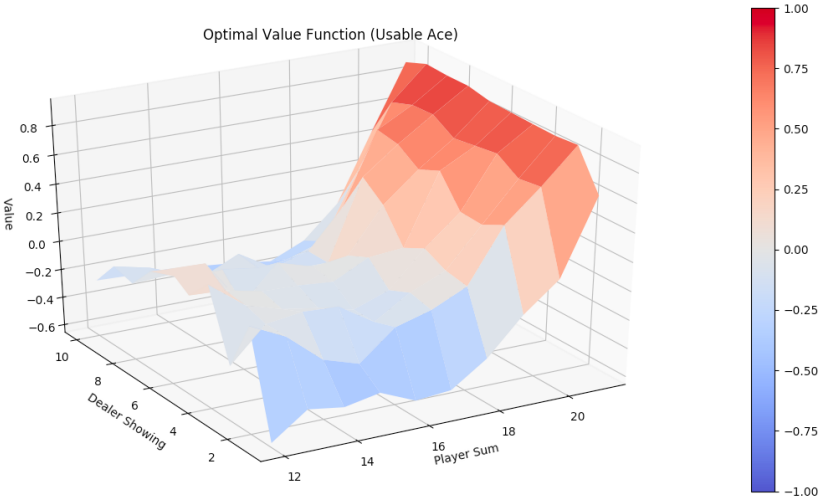




2.2 On-Policy Monte Carlo Control

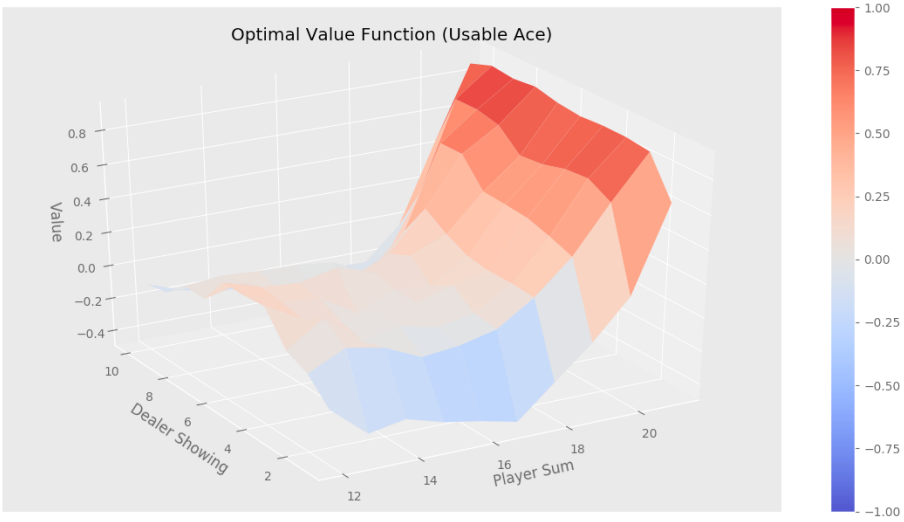
Then we demonstrate the use of on policy for Monte-Carlo methods, to estimate Q(s,a) from which the policy can be derived. Q(s,a) is again estimated from average returns. Policy can then be derived by finding the argmax over Q(s,a) in each state in the episode.





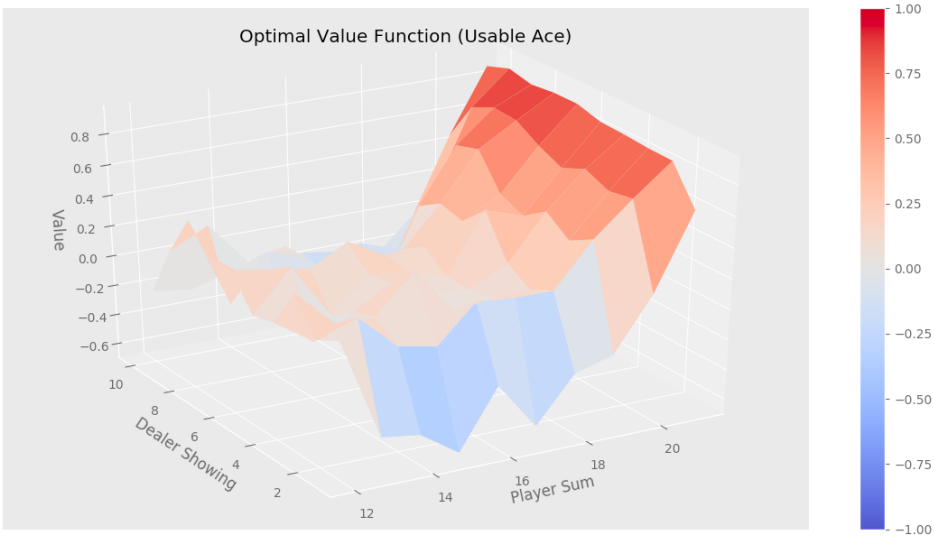
2.3 Off-Policy MC with ϵ Greedy Behaviour Policy

We first demonstrate the use of importance sampling methods for use in off-policy in the Blackjack environment. Here, the off-policy is an ϵ greedy behaviour policy. In other words, in this example, both π and μ are greedy and ϵ greedy policieis repsectively.



2.4 Off-Policy MC with Random Behaviour Policy

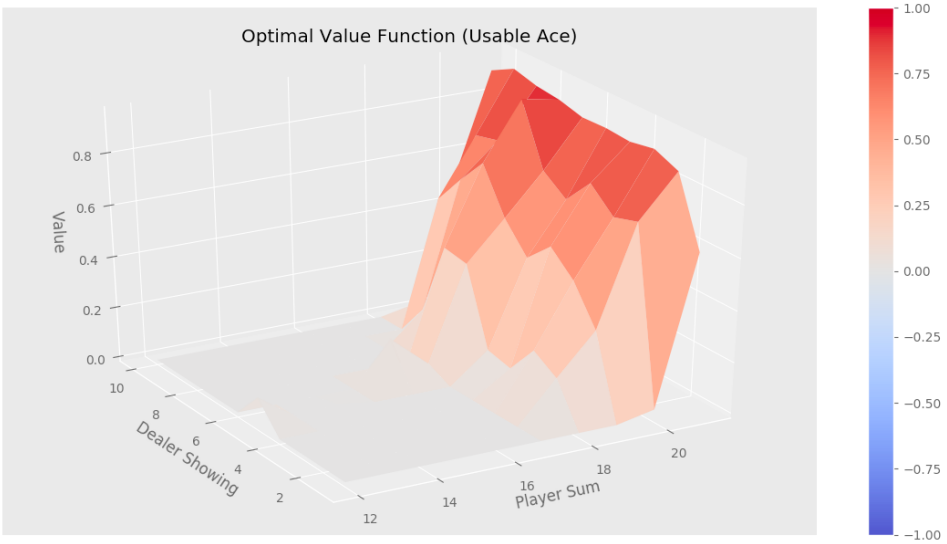
Compared to results above, we find that in the value function estimates, clear differences can be observed between using a greedy behaviour policy and a random behaviour policy. The random behaviour policy in our example here chooses random actions for exploration. Our results show that for random behaviour policies, the variance in value function estimates are higher compared to previous result with ϵ greedy policy.



2.5 Off-Policy MC with Boltzmann Behaviour Policy

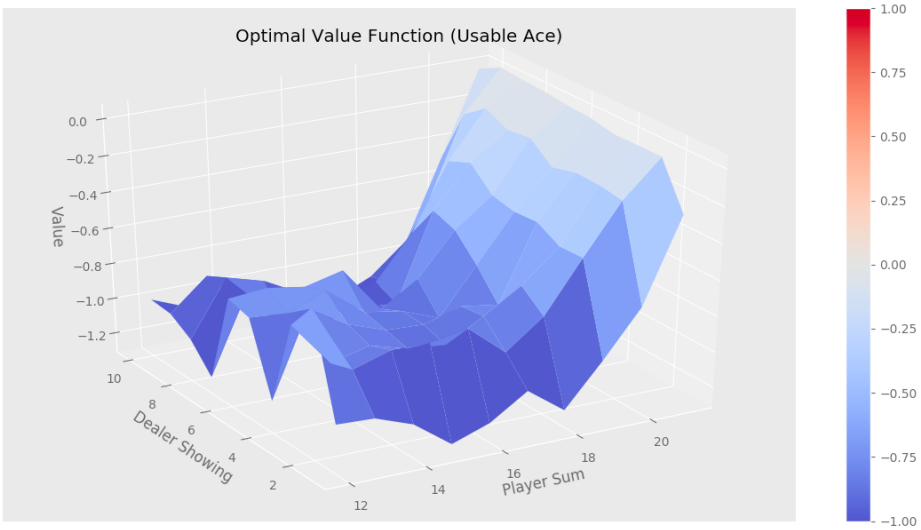
We try with a different variation of behaviour policies for exploration. Here we use a Boltzmann policy, with the temperature parameter set to 1. Ideally, this temperature parameter should be decreasing for better exploration. For ease of comparision with previous results, we set $T = 1$.

Experimental results here are significantly different compared to previous methods. Our results here surprisingly show that the value functions are only high towards the end of the episode, whereas at the beginning of the game trial, V estimates are all 0. This might suggest that due to the exploration approach, only good states with high Value estimates are found towards the end of the episode.

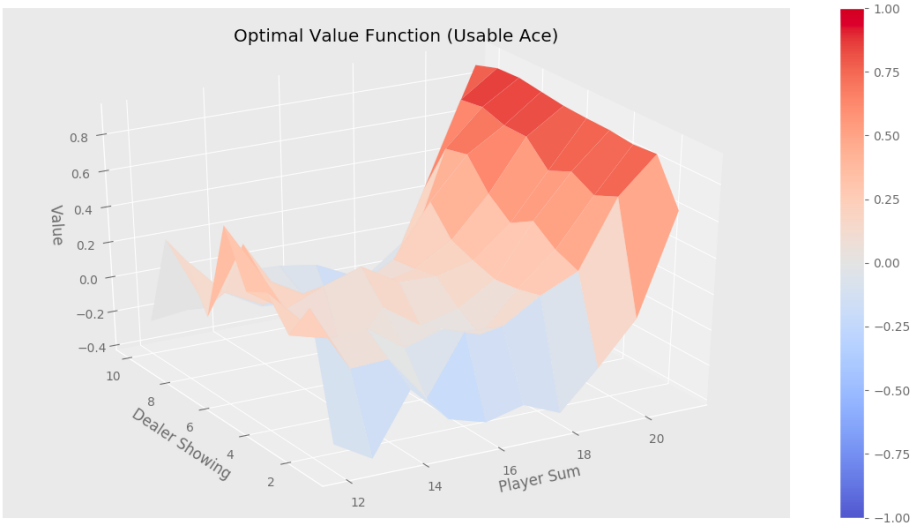


2.6 Variance Reduction in Off-Policy MC with Fixed/Arbitrary Baselines

Here we include a constant parameter for the baseline function to be subtracted from the Q(s,a) estimates. This is using a baseline value of 0.9 which is arbitrarily chosen. Results show that there is high variance in the V estimates, and good values are not found.

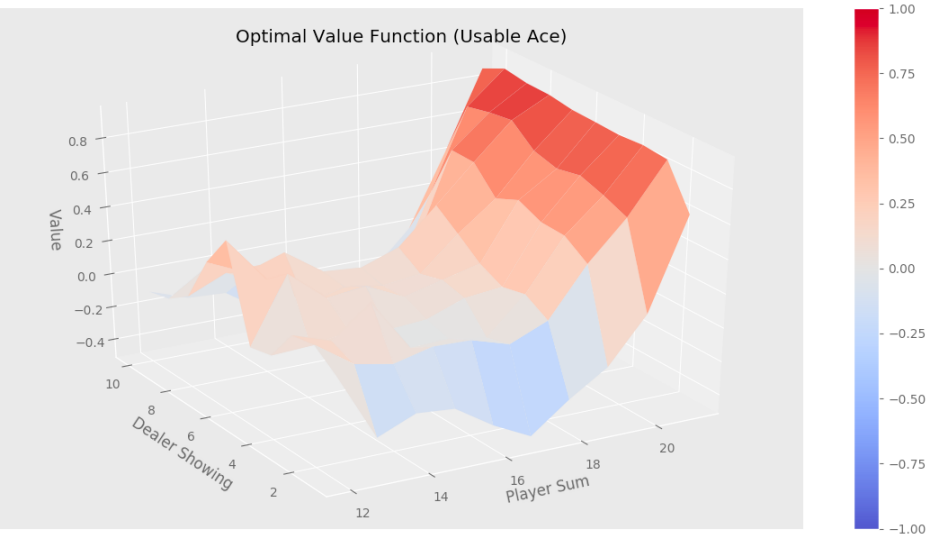


We then experiment with a much lower baseline value for the variance reduction technique. However, results here suggest that with low baselines, better V estimates can be found compared to before.



2.7 Variance Reduction in Off-Policy MC Advantage Functions

We then estimate the advantage value function instead of the action-value function. A(s,a) uses V(s) directly as a baseline method to reduce the variance in Q(s,a) estimates. The Value functions are then derived from A(s,a) directly for obtaining results as below.



2.8 Exploratory Analysis

In Off-policy MC, the Q(s,a) update equation is based on the importance sampling ratio as below

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{W}{C(s_t, a_t)}[G - Q(s_t, a_t)] \tag{1}$$

We derive an alternative Q(s,a) update from this. Even though the focus on MC methods is to trace back in the episode and estimate value functions based on average rewards. In the case of off-policy methods, we introduced importance sampling due to the difference in the policy the agent is following π, and the policy μ typically used for exploration. One alternative to this might be to analyse what happens when Q(s,a) is not updated towards the mean return G, but instead we use TD(0) in this case while using importance sampling. In other words, the update equation will look as follows:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{W}{C(s_t, a_t)}[R_{t+1} + \gamma Q(s', a') - Q(s_t, a_t)] \tag{2}$$

where essentially we compute the TD error, but still use importance sampling for estimated value functions. However, this is no longer a MC method, but instead we are perhaps just using a variant of TD learning using importance sampling(?).

3 Conclusion

All the code for our experimental results can be found accompanying this report. We used separate scripts for each of the demonstrations above for wider audience.

In this work, we demonstrated variants of using Monte-Carlo methods on the Blackjack game playing environment. We compared and analysed differences in on-policy and off-policy MC methods first, and then demonstrated the significance of using different behaviour off policies. Furthermore, we provided examples of the significance of variance reduction techniques (using baselines or control variates) for use in Monte-Carlo methods.