Exercise Set 2 - Reinforcement Learning

Tabular Methods

Giulio Starace - 13010840

September 19, 2022

Homework: Coding Assignment - Monte Carlo

1. To derive the incremental update rule of the value function for a given state $V(s)$ using ordinary importance sampling, we begin with the definition of the estimate $V_n(s)$ given n sampled returns $G_1, \dots, G_n$ under this regime:

$$V_n(s) = \frac{1}{n} \sum_{k=1}^{n} W_k G_k, \tag{1}$$

where W_k is the importance sampling ratio of target policy π over behavior policy b.

$$W_k = \rho_{t_k:T(t_k)-1} = \prod_{t=t_k}^{T(t_k)-1} \frac{\pi(A_t|S_t)}{b(A_t|S_t)}. \tag{2}$$

We expand equation (1) to obtain:

$$\begin{aligned}
V_n(s) &= \frac{1}{n} \left[W_{n-1} G_{n-1} \sum_{i=1}^{n-1} W_k G_k \right] \\
&= \frac{1}{n} \left[W_{n-1} G_{n-1} (n-1) \underbrace{\left(\frac{1}{n-1} \right) \sum_{i=1}^{n-1} W_k G_k} \right] \\
&= \frac{1}{n} \left[W_{n-1} G_{n-1} (n-1) \underbrace{V_{n-1}(s)} \right] \\
&= \frac{1}{n} \left[W_{n-1} G_{n-1} + n V_{n-1}(s) - V_{n-1}(s) \right] \\
V_n(s) &= V_{n-1}(s) + \frac{1}{n} \left[W_{n-1} G_{n-1} - V_{n-1}(s) \right].
\end{aligned} \tag{3}$$

We see that equation (3) is of the form

$$V_n = V_{n-1} + \alpha * (\beta - V_{n-1}), \tag{4}$$

with $\alpha = \frac{1}{n}$ and $\beta = W_{n-1} G_{n-1}$.

■

2. Coding answers have been submitted on codegra under the group "stalwart cocky sawly". Please refer to Figures 1 and 2 for the requested figures.

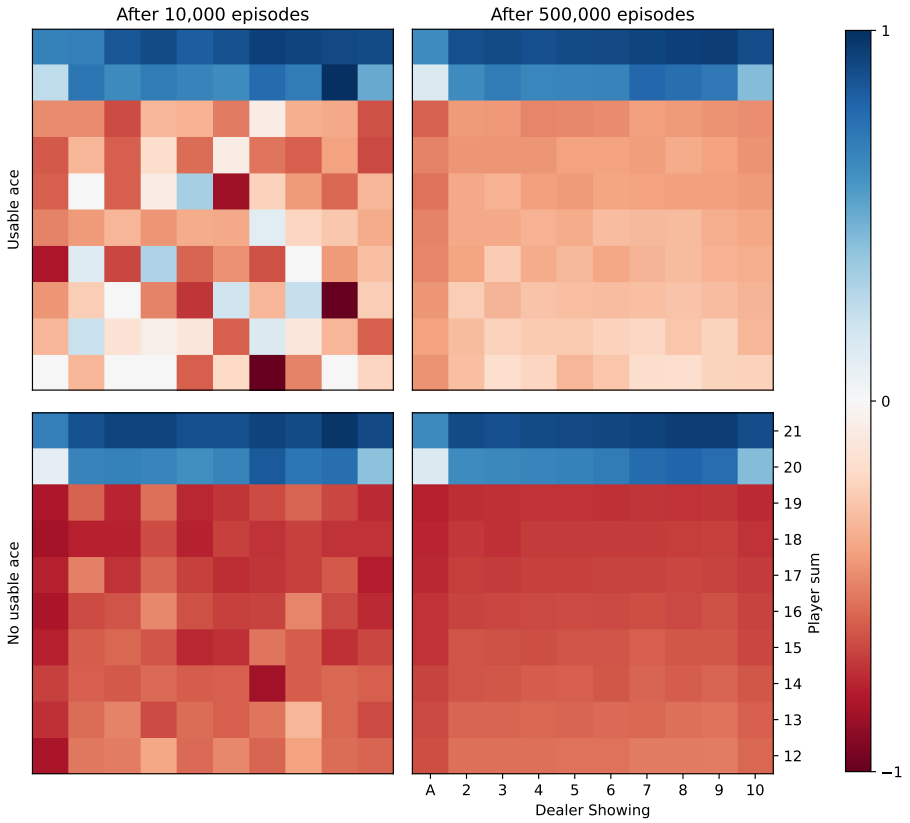

Figure 1: Blackjack value function across state configurations after 10k and 500k episodes using on-policy MC prediction. Reproduction of figure 5.1 from Sutton and Barto, using heatmaps.

3. hello world

Homework: SARSA and Q-learning

1. (a) hey
   (b) ho
2. (a) hey
   (b) ho
3. let's go

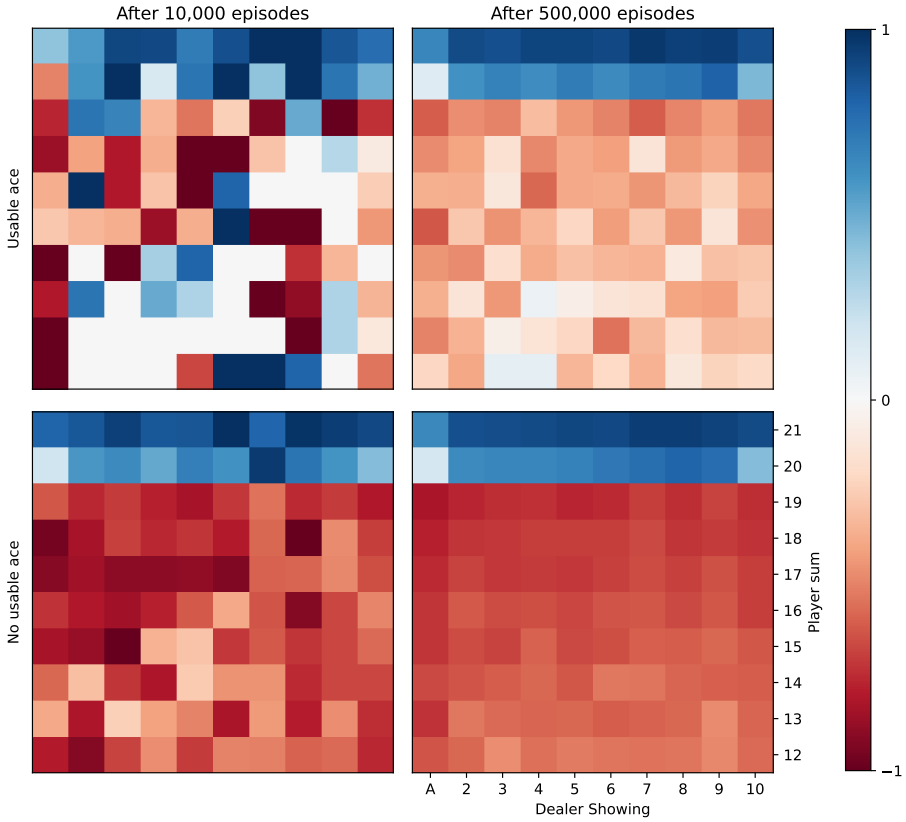

Figure 2: Blackjack value function across state configurations after 10k and 500k episodes using off-policy MC prediction with ordinary importance sampling.