Date / /

Ex - 04

Monte - Carlo Methods.

1) Incremental implementation of M-C methods.
First Visit MC prediction, for estimating
$V \approx v_\pi$ I/P : a policy π to be evaluated.
Initialize:
$V(s) \in R$, arbitrary, for all $s \in S$
where $N(s) \leftarrow 0$.

∴ Loop Forever (for every episode)
where $G \leftarrow 0$

∴ Loop for each step of episode, $t = T-1, T-2 \dots 0$
$G \leftarrow \gamma G + R_{t+1}$

Unless S_t appears in $S_0, S_1 \dots S_{t-1}$
$N(S_t) \leftarrow N(S_t) + 1$
$V(S_t) \leftarrow V(S_t) + (G - V(S_t)) / N(S_t)$

2) First visit Vs every-visit:

a)

In the BlackJack task suppose every visit MC used instead of first-visit. The each episode is constantly changing which will appear only once. Even if every visit MC is used. Since the state appears only once it gets the same result as using first-visit.

b) MDP with a single nonterminal state and a single action that transitions back to the non-terminal state with probability p and terminal state 1-p.

$T = 10$
$\gamma = 1$ } (given)

$G_t = R_t + G_1 = 10$
$G_1 = R_2 + G_2 = 9$
$G_2 = R_3 + G_3 = 8$
$\vdots$
$G_9 = R_{10} + G_{10} = 1$
$G_{10} =$

For First-visit
$V(s) = 10$

For every visit
$V(s) = \frac{G_{10} + G_9 + \cdots + G_0}{10}$

$= \frac{0 + 1 \cdots + 10}{10}$

$V(s) = 5.5$

2(c)

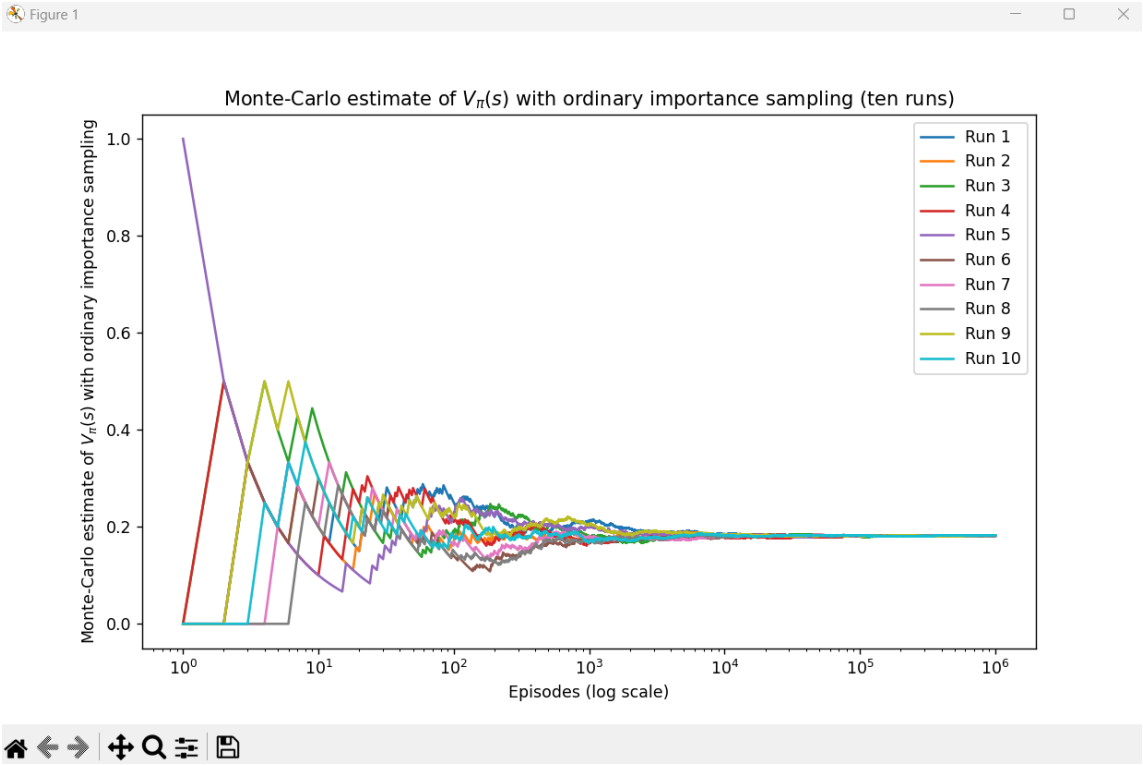
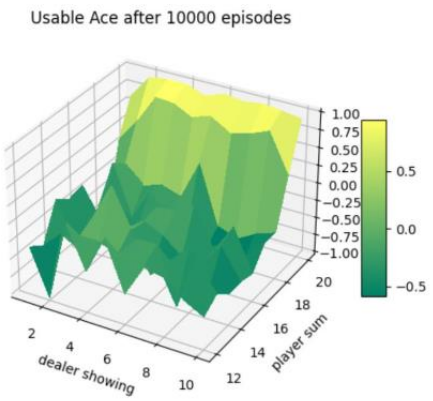
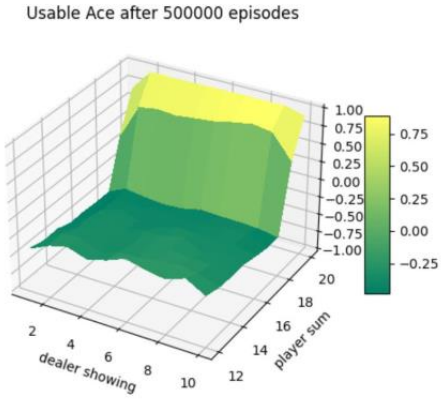
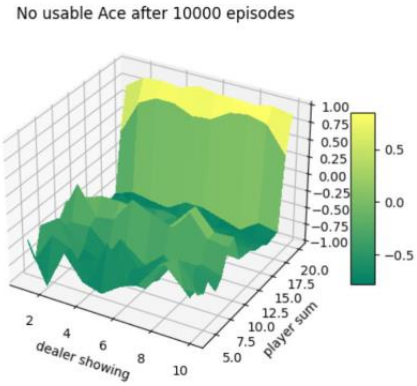


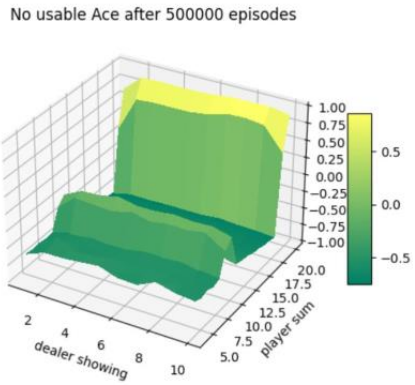
Figure 1
Monte-Carlo estimate of $V_\pi(s)$ with ordinary importance sampling (ten runs)
Monte-Carlo estimate of $V_\pi(s)$ with ordinary importance sampling
Episodes (log scale)
Run 1
Run 2
Run 3
Run 4
Run 5
Run 6
Run 7
Run 8
Run 9
Run 10
1.0
0.8
0.6
0.4
0.2
0.0
10^0
10^1
10^2
10^3
10^4
10^5
10^6

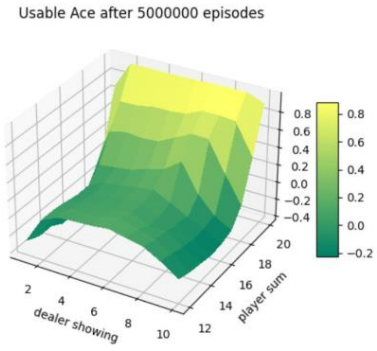
3) a), b)

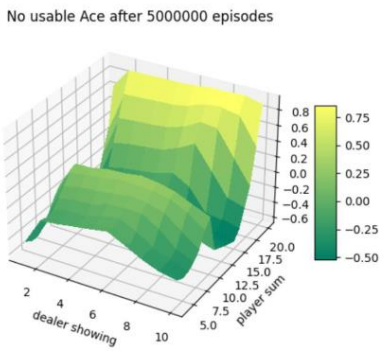

Usable Ace after 10000 episodes
dealer showing
player sum

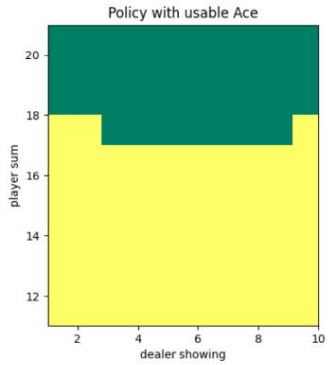

Usable Ace after 500000 episodes
dealer showing
player sum

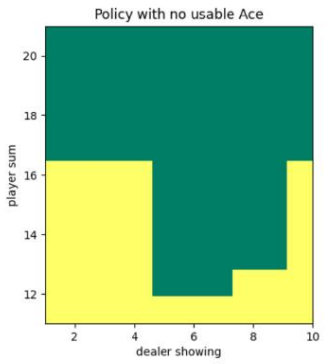

No usable Ace after 10000 episodes
dealer showing
player sum


No usable Ace after 500000 episodes
dealer showing
player sum


Usable Ace after 5000000 episodes
dealer showing
player sum

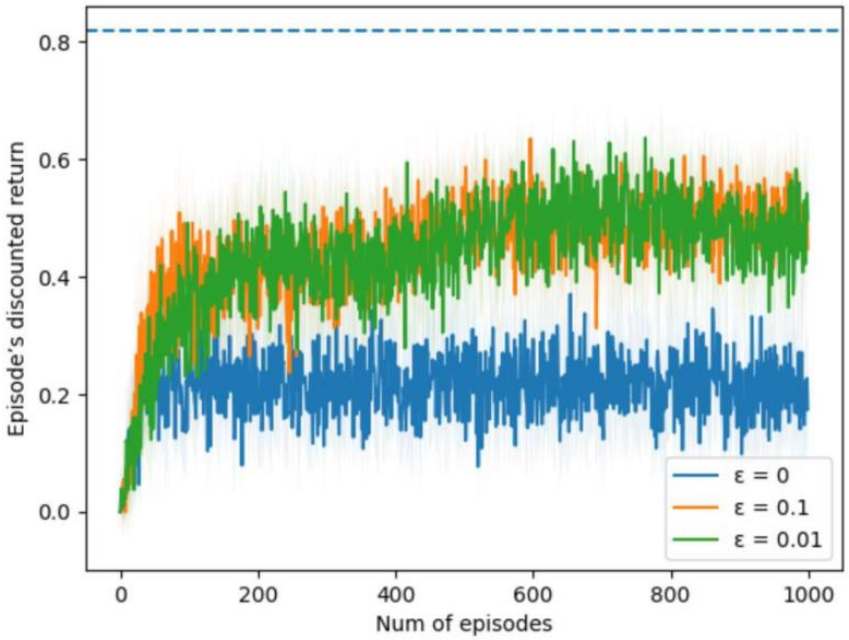

No usable Ace after 5000000 episodes
dealer showing
player sum


Policy with usable Ace
player sum
dealer showing


Policy with no usable Ace
player sum
dealer showing

4)

b)


0.8
0.6
0.4
0.2
0.0
Episode's discounted return
0
200
400
600
800
1000
Num of episodes
ε = 0
ε = 0.1
ε = 0.01

DOMS | Page No.
Date / /

4) Four Rooms, revisited.

c) Importance of doing exploring starts in M-C ES.

when $\epsilon = 0$ ~~Gobi~~ without exploring start point will always choose the state with the 1st +ve result & refuse to make a new attempt. However with Exploining Start this is prevented. It will have chances to try other states. Thus this will still be useful for improving the policy.