1.) Input : a Policy π to be evaluated

Initialize :

$V(s) = 0$ for all $s \in S$

$N(s) = 0$ for all $s \in S$

Loopforever (for each episode):

Generate an episode following π: $S_0, A_0, R_1, S_1, A_1, R_2 - - -$
S_{T-1}, A_{T-1}, R_T

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \ldots 0$:

$G \leftarrow \gamma G + R_{t+1}$

$N(S_t) \leftarrow N(S_t) + 1$

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}\left(G - V(S_t)\right)$$

2.a) => A State in Blackjack is the players total sum, the dealers visible card and whether there is a usable ace.

Considering there are no splitting or doubling downs according to special rules.

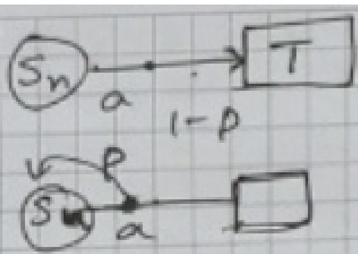
It can be considered that there wont be much of a difference because the episodes are ended in ~~the~~ game of blackjack when the player wins or losses or got a draw.

lets consider a episode where the user has (4, 2, usable Ace) and if he hits and got a 6 the total could be 21

another (K, 2, usable Ace) and he strikes the total could be 21

which is not usable as the episode is already happend and the game in won which will not impact the policy improvement as required.

2b)

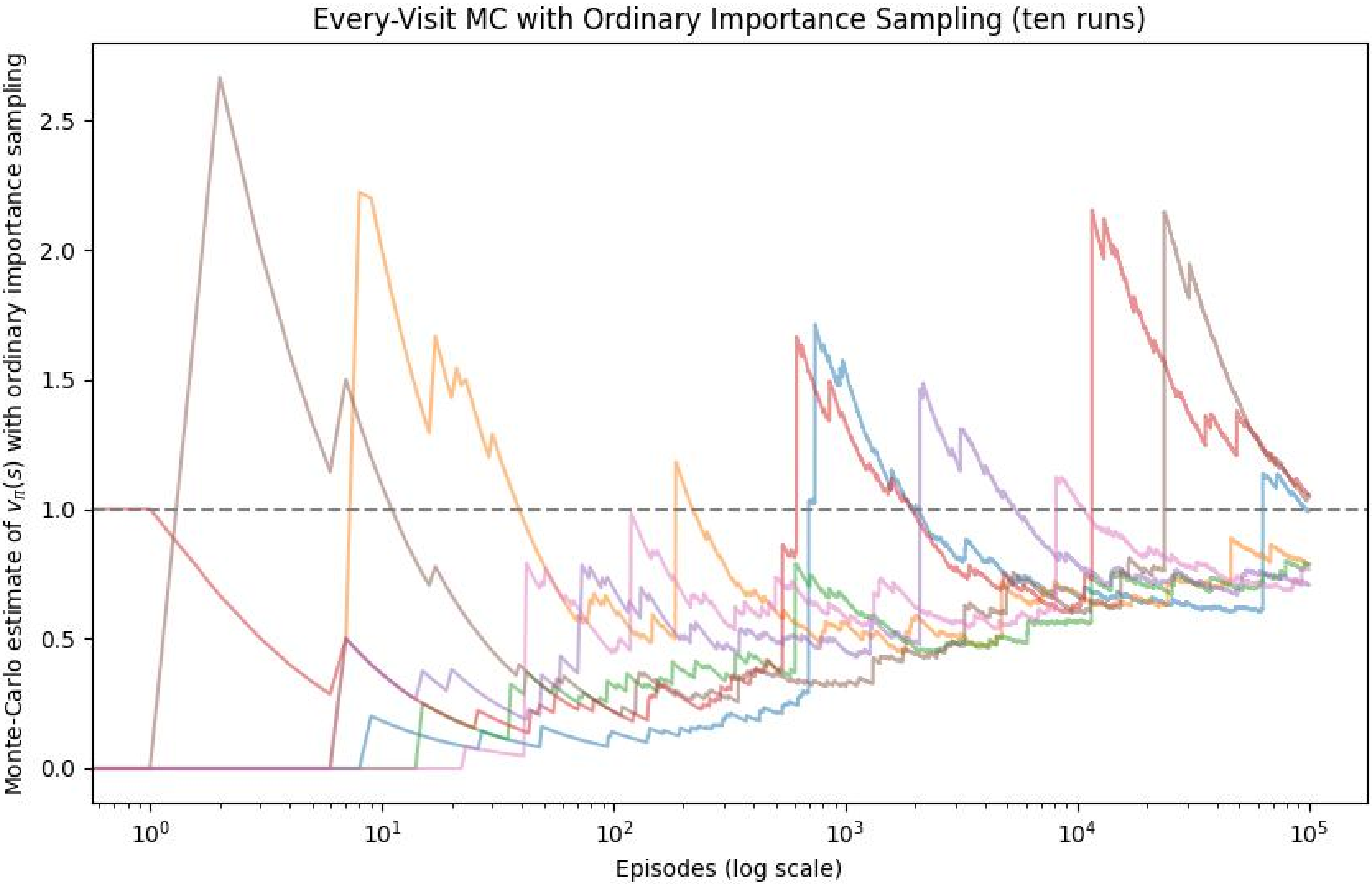


The first visit only updates the returns for the non-terminal state only once

$$Q_{FVMC}(s, a) = \text{average}(10) = 10$$

But in case of every visit-MC ~~at~~ updates the non-terminal state every time it visits with an increased reward value.

$$\therefore \quad Q_{EVMC}(s, a) = \text{average}(1+2+3 \cdots + 10) = 5.5$$


Every-Visit MC with Ordinary Importance Sampling (ten runs)
Monte-Carlo estimate of $v_\pi(s)$ with ordinary importance sampling
0.0
0.5
1.0
1.5
2.0
2.5
10^0
10^1
10^2
10^3
10^4
10^5
Episodes (log scale)