CS 4180/5180 — Exercise 4: Monte-Carlo Methods

Teoman Kaman

Q1 (RL2e 5.2, 5.5, 5.8) — First-visit vs every-visit (Written + Code/Plot)

(a)

Answer
No, it would not change much. In Blackjack, in a single episode you almost never visit the same state (player sum, dealer showing, usable ace), because the player sum typically increases with the "hit" action. So first-visit and every-visit MC are essentially the same for this problem (mostly the same states).

(b)

Answer:
For the single episode of length 10 with return 10 and $\gamma = 1$:

- **First-visit MC:** the state is first visited at $t = 0$, so
$$V(s) = G_0 = 10.$$
- **Every-visit MC:** average the return from every visit ($t = 0, 1, \ldots, 9$). Here $G_0 = 10, G_1 = 9, \ldots, G_9 = 1$, so
$$V(s) = \frac{10 + 9 + \cdots + 1}{10} = \frac{55}{10} = 5.5.$$

(c) (Extra credit)

textbfAnswer: Yes, the variance would still be infinite. In Example 5.5, ordinary importance sampling can have infinite variance when trajectories contain loops, because the importance-sampling ratios can grow exponentially with episode length, producing a heavy-tailed distribution whose second moment diverges. This is exactly the mechanism used in the book to show that the importance-sampling-scaled returns have infinite variance in this one-state looping MDP. Using every-visit MC does not eliminate these extreme ratios: every episode still includes the initial visit to s (at $t = 0$), whose ordinary-IS ratio has the same heavy-tailed behaviour as in the first-visit case; every-visit merely adds additional (later) visits and thus cannot remove the divergent second moment. Therefore, the ordinary-IS every-visit estimator also has infinite variance.

Code/Plot (Example 5.5, reproduce Fig. 5.4)

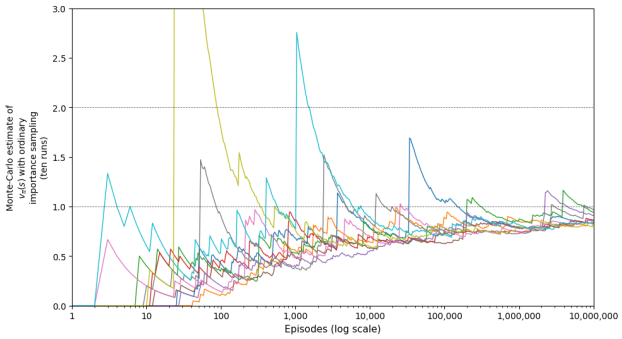

Figure 1: Q1 Code/Plot: Reproduction of RL2e Figure 5.4 (Example 5.5). R

Q2 — Blackjack (Code/Plot)

(a) First-visit MC prediction (sticks on 20 or 21) — reproduce Fig. 5.1

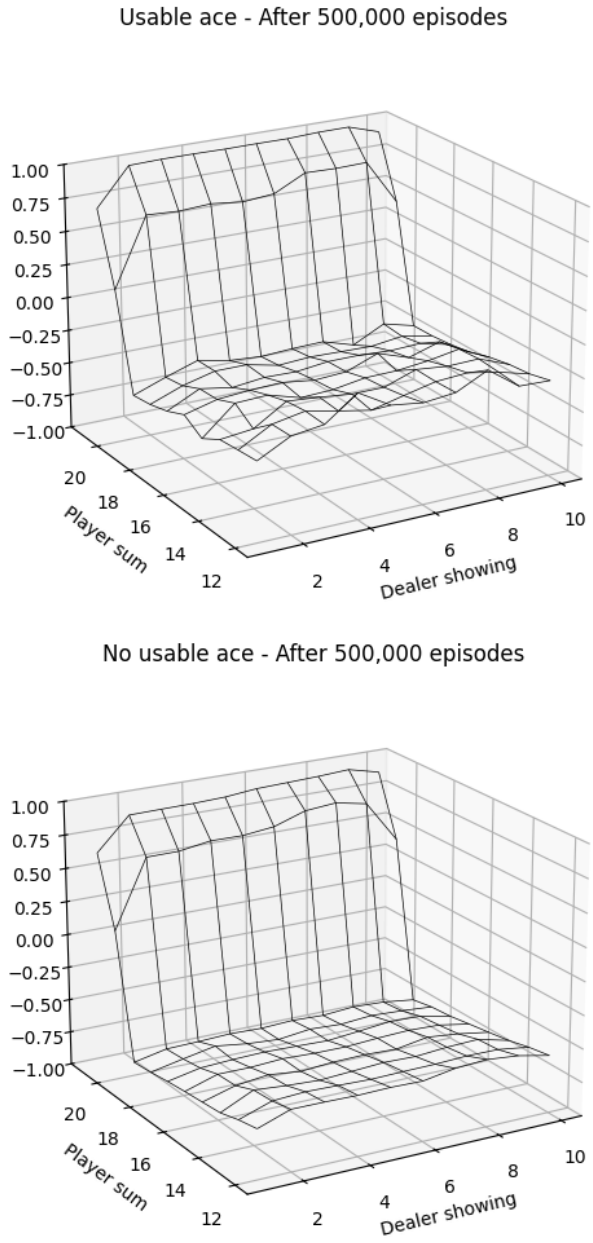

Figure 2: Q2(a): Reproduction of RL2e Figure 5.1.

(b) MC ES control — reproduce Fig. 5.2

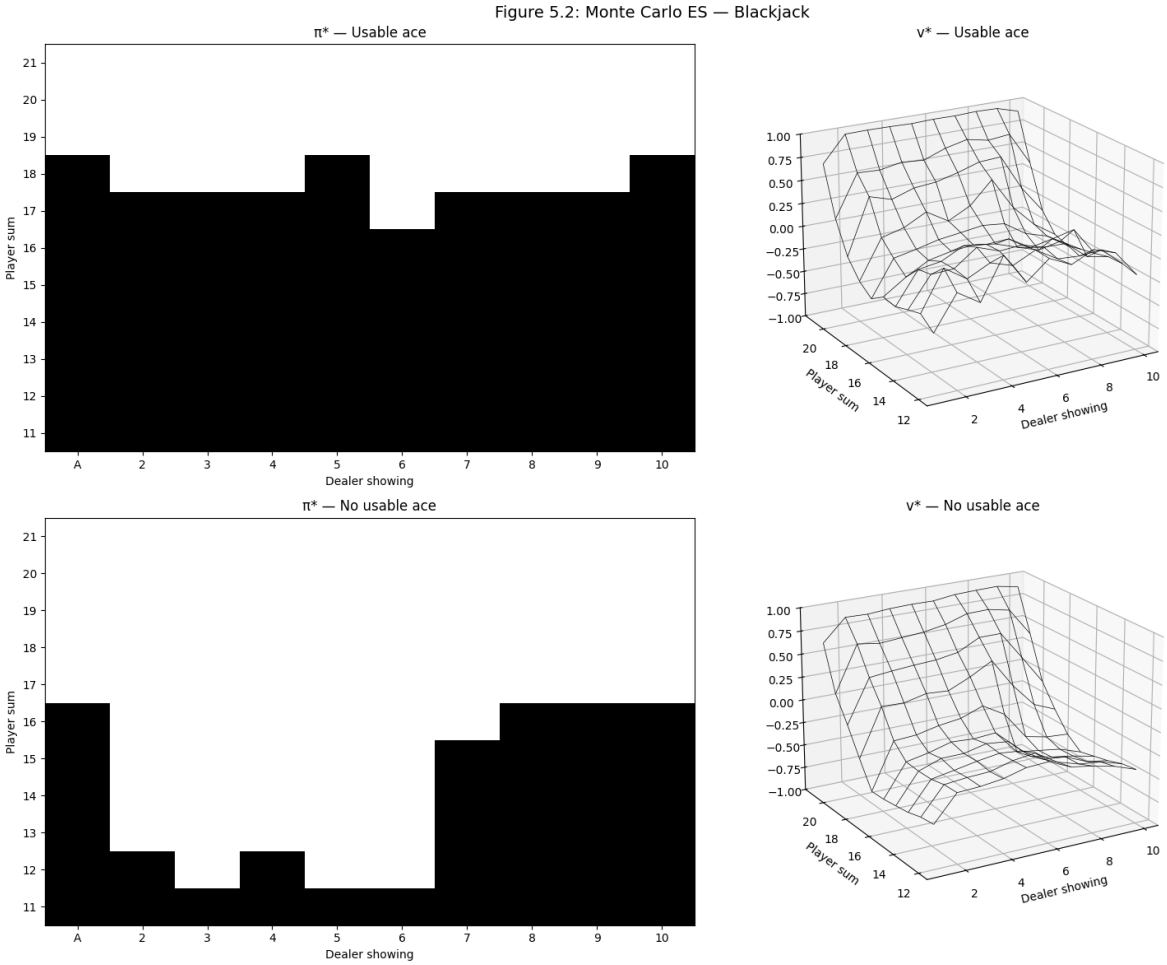

Figure 3: Q2(b): Reproduction of RL2e Figure 5.2.

Q3 — Four Rooms, re-visited (Code/Plot + Written)

(a) On-policy first-visit MC control (ϵ-soft) + learning curves

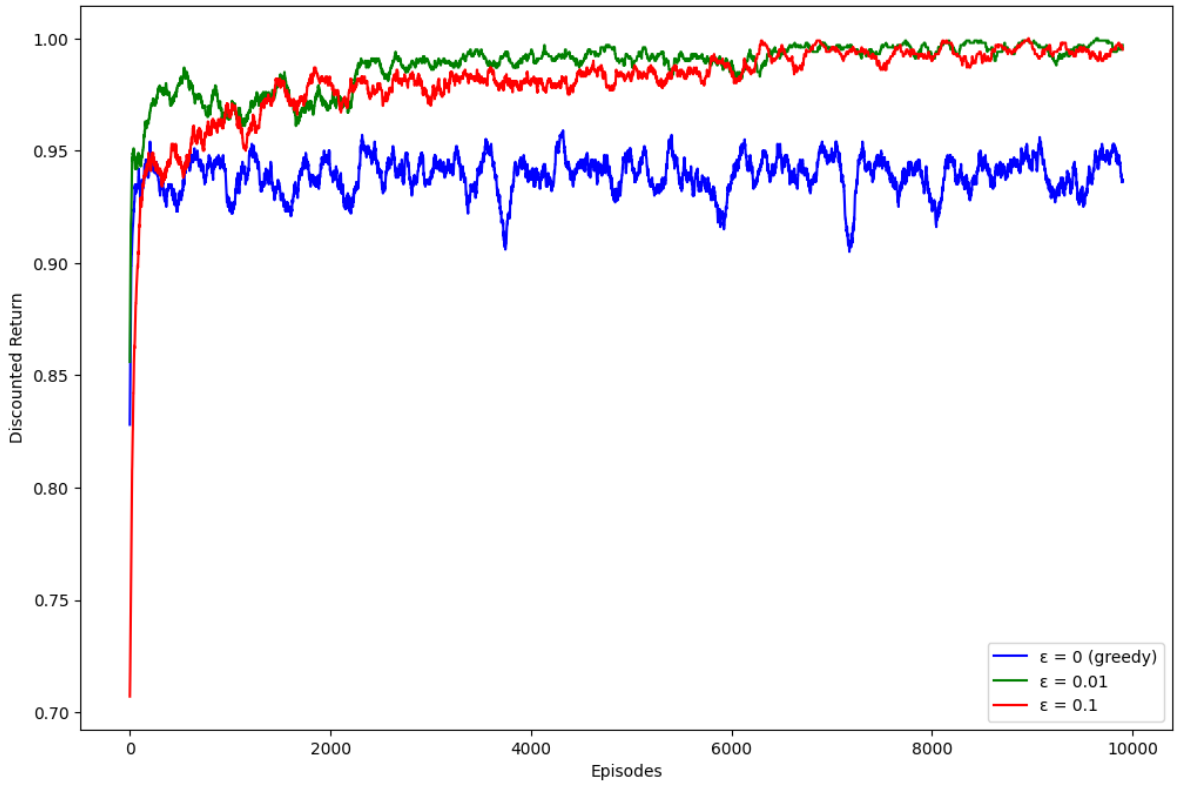

Figure 4: Q3(a): Learning curves for $\epsilon \in \{0.1, 0.01, 0\}$ with confidence bands.

(b) Written

Answer :
Without any exploration, $\epsilon = 0$ will pick only greedy actions. Since every new episode starts from $(0, 0)$, the agent might not visit all states, thus it cannot update Q-values everywhere. So it gets stuck in some path greedily and might miss optimal paths.
Monte-Carlo exploring starts fixes this by giving non-zero probability to each state-action pair, which guarantees that every state-action pair will be visited in the limit.
To sum up: $\epsilon = 0$ without exploring starts leads to poor learning because the agent is stuck in greedy trajectories; exploring starts solve this by giving each state-action pair a non-zero probability to start, so the agent can try different conditions.

Q4 (RL2e 5.10, 5.11) — Off-policy methods (Written)

(a) Derive Eq. (5.8) from Eq. (5.7)

Derivation:
Start from the weighted estimate

$$V_n \doteq \frac{\sum_{k=1}^{n-1} w_k G_k}{\sum_{k=1}^{n-1} w_k}.$$

Define the cumulative weight

$$C_n \doteq \sum_{k=1}^{n} w_k, \quad C_0 = 0, \quad C_n = C_{n-1} + w_n.$$

Let the numerator be

$$N_n \doteq \sum_{k=1}^{n} w_k G_k.$$

Then

$$V_n = \frac{N_{n-1}}{C_{n-1}}, \qquad V_{n+1} = \frac{N_n}{C_n} = \frac{N_{n-1} + w_n G_n}{C_{n-1} + w_n}.$$

Use $N_{n-1} = C_{n-1} V_n$:

$$V_{n+1} = \frac{C_{n-1} V_n + w_n G_n}{C_{n-1} + w_n}.$$

Add and subtract $w_n V_n$ in the numerator:

$$V_{n+1} = \frac{(C_{n-1} + w_n) V_n + w_n (G_n - V_n)}{C_{n-1} + w_n} = V_n + \frac{w_n}{C_n}(G_n - V_n),$$

which is Eq. (5.8).

(b)

Answer:
Because the target policy is deterministic greedy, $\pi(a|s) = 1$ only for the greedy action and 0 otherwise. So along trajectories where the action matches the greedy target policy, the importance ratio is

$$\frac{\pi(A_t|S_t)}{b(A_t|S_t)} = \frac{1}{b(A_t|S_t)}.$$

If A_t is not the greedy action, then $\pi(A_t|S_t) = 0$ and the product becomes 0, so the algorithm stops updating for that episode.

Q5 (RL2e 5.12) — Racetrack (Code/Plot + Written)

(a) On-policy first-visit MC control ($\epsilon = 0.1$) + learning curves (both tracks)

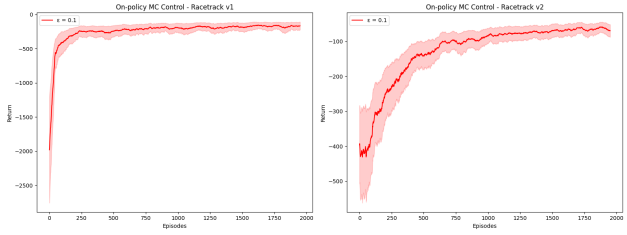

Figure 5: Q5(a): Track 1 and Track 2 learning curve with confidence bands. R

(b) Off-policy MC control + learning curves + rollouts (both tracks)

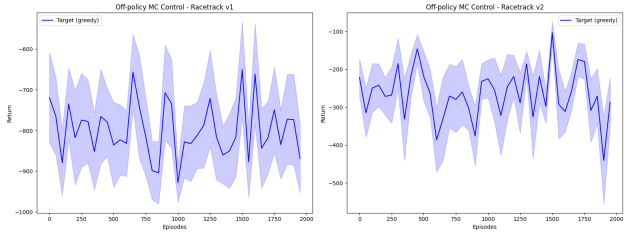

Figure 6: Q5(b): Track 1 (target policy) learning curve.

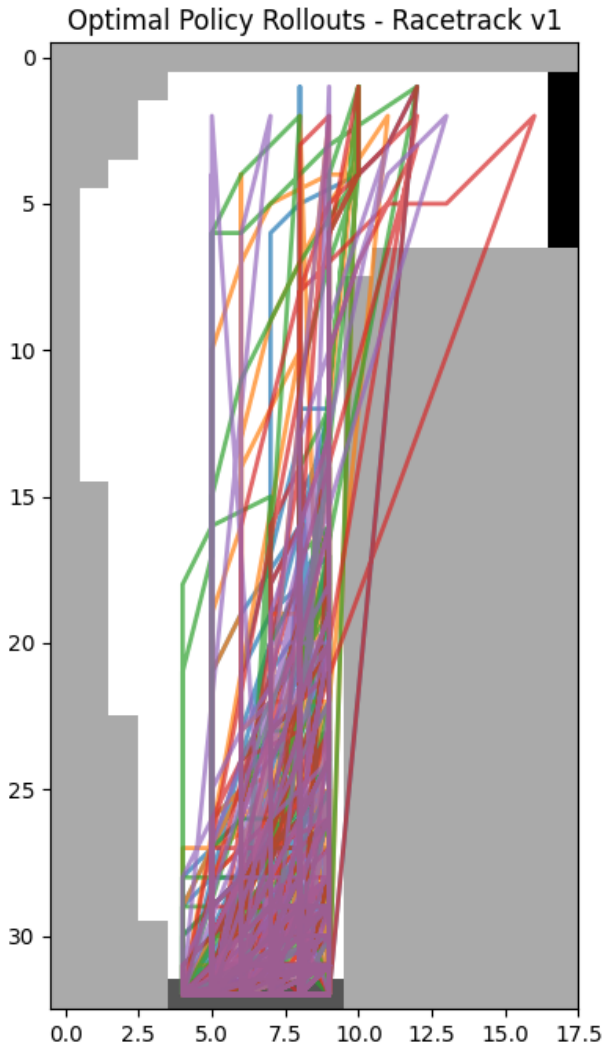

Figure 7: Q5(b): Several rollouts on Track 1.

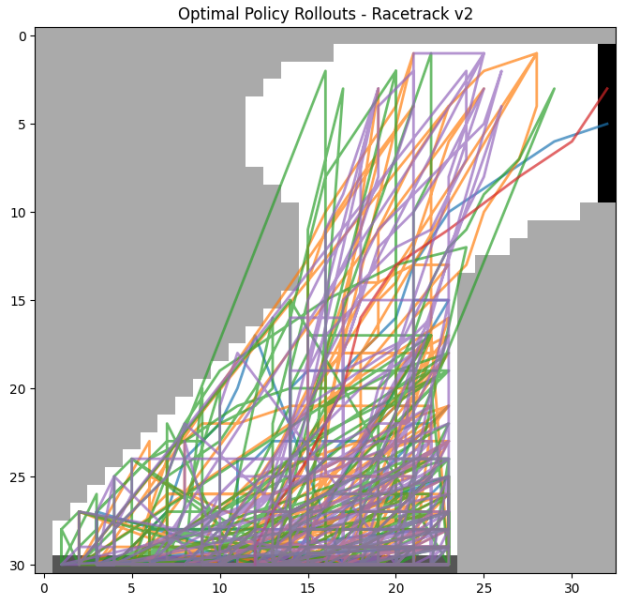

Figure 8: Q5(b): Several rollouts on Track 2. .

(c) Written

Answer:
There are clear differences between the on-policy and off-policy Monte-Carlo control methods. The on-policy method shows steady and consistent improvement over time. The learning curves increase smoothly and the variance decreases as training continues. This is because the same -soft policy is used for both behavior and evaluation, so the updates are stable.

In contrast, the off-policy method shows much higher variability. It relies on importance sampling, where the update weight is a product of probability ratios over the entire episode. Since racetrack episodes can be long, these weights can become very large or zero. Large weights increase variance, while zero weights mean some episodes do not contribute to learning. As a result, the learning process is more unstable.

There are also differences between the two racetracks. Racetrack v1 appears more difficult because it has tighter turns and fewer safe paths, which causes more crashes and slower improvement. Racetrack v2 allows smoother trajectories and multiple possible routes to the finish line, leading to better final performance.