Name: Alfred
Student No.: 2020-12788

Exercise 5

Data Science & Reinforcement Learning

Spring 2023

1. Recall the n-step value function is defined as

$$V_{t+n}(S_t) = V_{t+n-1}(S_t) + \alpha \left[G_{t:t+n} - V_{t+n-1}(S_t)\right], \quad 0 \le t \le T$$

where $G_{t:t+n}$ is is the n-step return

$$G_{t:t+n} := R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})$$

$V(S_t) = V(S_{t+1}) = \cdots = V$

for $0 \le t < T-n$, and $G_{t:t+n} := G_t$ for $t \ge T-n$. Show that the $\boxed{n\text{-step TD error}}$, $G_{t:t+n} - V_{t+n-1}(S_t)$, can be written as a sum of TD errors if the value estimates don't change from step to step.

$V(S_t)$

$$\sum_{k=t}^{\min(t+n,T)} \delta_t = (R_{t+1} + \gamma \cancel{V(S_{t+1})} - V(S_t)) + (R_{t+2} + \gamma \cancel{V(S_{t+2})} - \cancel{V(S_{t+1})}) + \cdots$$

$$= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots + \gamma^n V(S_{t+n}) - V(S_t)$$

$$= G_{t:t+n} - V_{t+n-1}(S_t)$$

2. Recall the n-step return of Sarsa defined as

$$G_{t:t+n} := R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n})$$

for $0 \le t < T-n$, and $G_{t:t+n} := G_t$ for $t \ge T-n$. Prove that the n-step return of Sarsa can be written exactly in terms of a novel TD error, as

$$G_{t:t+n} = Q_{t-1}(S_t, A_t) + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t} \left[R_{k+1} + \gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k)\right]$$

$$\text{RHS} = \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t} R_{k+1} + \sum_{k=t}^{\min(t+n,T)-1} \gamma^{k-t} (\gamma Q_k(S_{k+1}, A_{k+1}) - Q_{k-1}(S_k, A_k))$$

$$\begin{cases} R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} & t < t+n \\ G_T & T \ge t+n \end{cases}$$

$\gamma Q_t(S_{t+1}, A_{t+1}) - Q_{t-1}(S_t, A_t)$
$+ \gamma^2 Q_{t+2}(S_{t+2}, A_{t+2}) - \gamma Q_t(S_{t+1}, A_{t+1}) + \cdots$

$$\begin{cases} \gamma^n Q_{t+n-1}(S_{t+n}, A_{t+n}) - Q_{t-1}(S_t, A_t) & \text{for } (t < T-n) \\ 0 - Q_{t-1}(S_t, A_t) & (t \ge T-n) \end{cases}$$

$(\because Q_{\text{terminal}} = 0)$

① then expand sides.

3. An agent observes the following two episodes from an MDP,

$$S_0 = 0, A_0 = 1, R_1 = 1, S_1 = 1, A_1 = 1, R_2 = 1$$
$$S_0 = 0, A_0 = 0, R_1 = 0, S_1 = 0, A_1 = 1, R_2 = 1, S_2 = 1, A_2 = 1, R_3 = 1$$

and updates its deterministic model accordingly. What would the model output for the following queries:

(a) Model($S = 0, A = 0$): ~~S=initial , R=0.~~ S=0, R=0

(b) Model($S = 0, A = 1$): ~~S= initial , R=1.~~ S=1, R=1

(c) Model($S = 1, A = 0$): ~~S= , R=~~ initial values

(d) Model($S = 1, A = 1$): ~~S=1 , R=~~ R=1, S= terminal

4. Why did the Dyna agent with exploration bonus, Dyna-Q+, perform better in the first phase as well as in the second phase of the blocking experiment in the figure below.

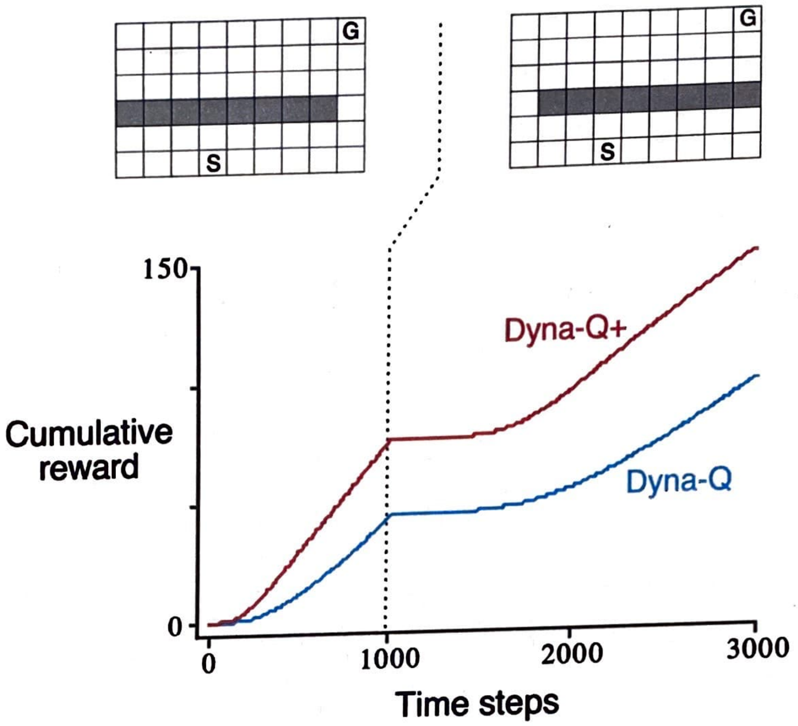


Figure 1: Average performance of Dyna agents on a blocking task. The left environment was used for the first 1000 steps, the right environment for the rest. Dyna-Q+ is Dyna-Q with an exploration bonus that encourages exploration

because there's a bonus reward

Fist half: early beginnings after Dyna-Q+ tend to give higher R,

for second phase Dyna-Q's poorer than Dyna-Q+ because

it explores more, finds optimal policy faster as well as

model's difference too. Phase 1 $\wedge$ 2 (after 1000 time steps)

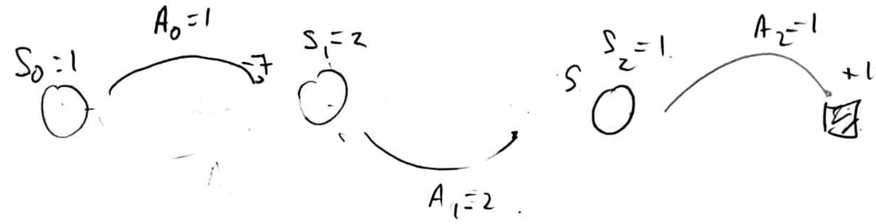
the dynamics change too.

5. Consider an MDP with three states $S = \{1, 2, 3\}$, where each state has two possible actions $\mathcal{A} = \{1, 2\}$ and a discount rate $\gamma = 0.5$. Suppose estimates of $Q(S, A)$ are initialized to 0 and you observed the following episode according to an unknown behaviour policy where S_3 is the terminal state.

$$S_0 = 1, A_0 = 1, R_1 = -7, S_1 = 2, A_1 = 2, R_2 = 5, S_2 = 1, A_2 = 1, R_3 = 10$$

(a) Suppose you used Q-learning with the above trajectory to estimate $Q(S, A)$, what are your new estimates for $Q(S = 1, A = 1)$ using $\alpha = 0.1$?

(b) Suppose in the planning loop, after search control, we would like to update $Q(S = 1, A = 1)$ with Q-planning. What are the possible outputs of Model$(S = 1, A = 1)$?



a) $Q^{new}(S=1, A=1) = 0 + 0.1(-7 + (0.5)(0) - 0)$ 1st iteration

$Q^{new}(S=1, A=1) = \underline{-0.7}$

2nd iteration.

$Q(1,1) \leftarrow Q(1,1) + 0.1(10 + 0.5 \times 0 - Q(1,1)) = 1 - 0.63 = \underline{0.37}$

b)

R, S = (-7, 2) V (10, termine) both true.

or if stochastic → {(-7, 2), (10, termine)} true too
50% 50%