RL-Week-6

Practice Assignment

truncated reward — correc^n quantity

1) Consider following equation for reward at time step t: — 1 step reward — 1 point

$G_t^{(1)} = r_{t+1} + \gamma V_t(s_{t+1})$ $\quad G_t^{(1)} = r_{t+1} + \gamma V_t(s_{t+1})$

Choose the correct statement(s) from following:

- [ ] It is a 1-step return. → 1-step truncated & corrected reward
- [ ] It is truncated because only the next step reward is available and no rewards are known/sampled from the episode.
- [ ] $\gamma V_t(s_{t+1})$ is the correction quantity because the return is truncated.
- [ ] None of the above

2) What is the advantage of computing n^{th} step reward vs 1-step reward? — TD(n) — 1 point

- [ ] At the beginning of the training the agent, the value estimates could be very far away from the true values. The incorrect estimates will be given a smaller (exactly γ^n) weightage. — TD(λ) → TD(n step returns) → the wrong estimates are penalized with the discounting factor
- [ ] It will be more accurate.
- [ ] It will take lesser computation. — much larger computa^n
- [ ] None of these

3) What is the disadvantage of computing n^{th} — TD(n) → the final update happens after n-steps — 1 point

- [ ] The update can only happen after taking n-steps.
- [ ] The variance could be very large due to longer due to more number of samples drawn — variance ↑
- [ ] None of these.

sol^n → mini-batch TD: it updates after each batch completes the episode.

$G_t^{\lambda} = r_{t+1} + \gamma V_t(s_{t+1})$

4) In computing G_t^{λ}, which of the following return estimates are used, assuming $0 < \lambda \leq 1$? — 1 point

- ○ Only G_t^1
- ○ 1-step return, 2-step return..... n-th step return, i.e. all of them. — $0.05 G_t^1 + 0.02 G_t^2 + 0.1 G_t^3 + \dots + 0.9 G_t^n$
- ○ 1-step return, 3-step return..... $(2k+1)$-th step return, i.e. only alternates.
- ○ None of these.

5) What values of λ, for the lambda return G_t^{λ}, gives 1-step TD return and Monte Carlo return ? — 1 point

- [ ] Lambda return G_t^{λ}, becomes Monte Carlo return if $\lambda = 1$
- [ ] Lambda return G_t^{λ}, becomes 1-step TD return if $\lambda = 1$
- [ ] Lambda return G_t^{λ}, becomes Monte Carlo return if $\lambda = 0$
- [ ] Lambda return G_t^{λ}, becomes 1-step TD return if $\lambda = 0$

MC → T → leng. of episode → $\lambda = 1$

G_t^{λ} → 1 step return

$\lambda = 0 \Rightarrow G_t^{(1)}$ → TD(0) → 1 step TD return

$G_t^{\lambda} = (1-\lambda)\sum_{n=1}^{T-t-1} \lambda^{n-1} G_t^{(n)} + \lambda^{T-t-1} G_t$ — in MC there is no weighted avg.

G_t → MC

Common data for Q.6 and Q.7 questions

Consider following pseudo code for TDλ algorithm:

Algorithm 1 TD (λ) algorithm

1: Initialize $V(s)$ randomly and $e(s) = 0, \forall s \in S$
2: **repeat**(for each episode):
3: Initialize s
4: **repeat**(for each step in the episode):
5: $a \leftarrow$ action given by π for s
6: Take action a, receive reward r and observe next state s'
7: $\delta \leftarrow r + \gamma V(s') - V(s)$
8: $e(s) \leftarrow e(s) + 1$
9: **for** do$\forall s \in S$
10: $V(s) \leftarrow V(s) + \alpha\delta e(s)$
11: $e(s) \leftarrow e(s) + \gamma\lambda e(s)$
12: **end for**
13: $s \leftarrow s'$
14: **until** s is terminal

6) Which of the following statements are correct about the above algorithm? *1 point*

- [ ] $e(s) \leftarrow e(s) + 1$ appears before updating $V(s)$ because the algorithm increases the eligibility trace only for the state s which appears in the episode, and later updates the eligibility trace for all the states.
- [ ] The algorithm is incorrect.
- [ ] The eligibility traces are decayed incorrectly.
- [ ] The eligibility traces are decayed after the update to pre-calculate for the next update/step.
- [ ] None of the above

7) Write the step number that has to be changed in the above algorithm if accumulating traces have to be changed to replacing traces.

14) Which of the following options represent correct update rules for double Q-learning? *1 point*

Assume that the current state is S, action taken in state S is A, next state is S' and reward received is R. The updates for Q_1 and Q_2 are executed with 0.5 probability each.

- $Q_1(S,A) \leftarrow Q_1(S,A) + \alpha(R + \gamma Q_2(S', argmax_a Q_1(S',a)) - Q_1(S,A))$
  $Q_2(S,A) \leftarrow Q_2(S,A) + \alpha(R + \gamma Q_1(S', argmax_a Q_2(S',a)) - Q_2(S,A))$
- $Q_1(S,A) \leftarrow Q_1(S,A) + \alpha(R + \gamma Q_1(S', argmax_a Q_2(S',a)) - Q_1(S,A))$
  $Q_2(S,A) \leftarrow Q_2(S,A) + \alpha(R + \gamma Q_1(S', argmax_a Q_2(S',a)) - Q_2(S,A))$
- $Q_1(S,A) \leftarrow Q_1(S,A) + \alpha(R + \gamma Q_2(S', argmax_a Q_1(S',a)) - Q_1(S,A))$
  $Q_2(S,A) \leftarrow Q_2(S,A) + \alpha(R + \gamma Q_1(S', argmax_a Q_1(S',a)) - Q_2(S,A))$
- $Q_1(S,A) \leftarrow Q_1(S,A) + \alpha(R + \gamma Q_2(S', max_a Q_1(S',a)) - Q_1(S,A))$
  $Q_2(S,A) \leftarrow Q_2(S,A) + \alpha(R + \gamma Q_1(S', max_a Q_2(S',a)) - Q_2(S,A))$

Graded Assignment

1) Which of the following is the correct 5-step truncated corrected return starting from time t?

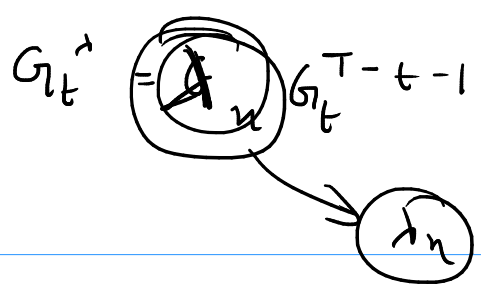
- $G_t^{(5)} = r_{t+1} + \gamma V_t(s_{t+1})$
- $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma^5 V_t(s_{t+1})$
- $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma^5 V_t(s_{t+5})$
- $G_t^{(5)} = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \gamma^4 r_{t+5} + \gamma V_t(s_{t+5})$
- $G_t^{(5)} = \sum_{i=t+1}^{T} [\gamma^{i-1} r_i]$, where T is the terminal state index.

2) Which of the following is the correct formula for computing G_t^λ? *1 point*

- $G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^n$
- $G_t^\lambda = (-1+\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^n$
- $G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n} G_t^n$
- $G_t^\lambda = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n+1} G_t^n$

3) Which of the following is the correct coefficient of G_t^{T-t-1} for computing G_t^{λ}? *1 point*

- $(1-\lambda)\lambda^{T-t-1}$
- $(1-\lambda)\lambda^{T-t}$
- λ^{T-t-1}
- $(1-\lambda)\lambda^{T-1}$
- None of these



4) Which of the following is the correct definition of accumulating eligibility traces? *1 point*

$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s), & \text{if } s \neq s_t. \\ \gamma\lambda e_{t-1}(s) + 1, & \text{if } s = s_t. \end{cases}$$

$$e_t(s) = \begin{cases} \lambda e_{t-1}(s), & \text{if } s \neq s_t. \\ \lambda e_{t-1}(s) + 1, & \text{if } s = s_t. \end{cases}$$

$$e_t(s) = \begin{cases} \gamma e_{t-1}(s), & \text{if } s \neq s_t. \\ \gamma e_{t-1}(s) + 1, & \text{if } s = s_t. \end{cases}$$

$$e_t(s) = \begin{cases} \lambda e_{t-1}(s), & \text{if } s \neq s_t. \\ \lambda e_{t-1}(s) \cdot 10, & \text{if } s = s_t. \end{cases}$$

5) Which of the following are correct about the notion of eligibility traces? *1 point*

- It is the backward view of the TD(λ
- It is the forward view of the TD(λ
- The motivation behind eligibility traces is, if an agent receives a reward/punishment at any time step, then which decisions in the past are eligible to get credit for it.

Yes, the answer is correct.
Score: 1

Accepted Answers:

It is the backward view of the TD(λ
The motivation behind eligibility traces is, if an agent receives a reward/punishment at any time step, then which decisions in the past are eligible to get credit for it.

6) Which of the following statement(s) are correct about the TD(λ) algorithm? *1 point*

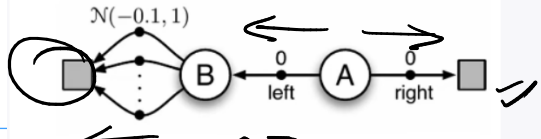
- For $\lambda = 1$, the algorithm will behave like SARSA.
- For $\lambda = 1$, the algorithm will behave like Q-learning.
- For $\lambda = 1$, the algorithm will behave like Double Q-learning.
- For $\lambda = 1$, the algorithm will behave like Monte Carlo and will update incrementally instead of waiting for the trajectory to end.

7) Consider following statements: *1 point*

Assertion: Double Q-learning reduces maximization bias as compared to Q-learning.
Reason: Double Q-learning uses one third of the samples to estimate the action values and remaining samples to choose the action with max estimates. This reduces the probability of overestimating action values by both sets of samples.

- Assertion and Reason are both true and Reason is a correct explanation of Assertion.
- Assertion and Reason are both true and Reason is not a correct explanation of Assertion.
- Assertion is true but Reason is false
- Assertion is false but Reason is true.

Consider following diagram for Q.8 to Q.10 couple of questions:



$\mathbb{E}[R_{t+1}] = 0.1$

Refer to the lecture for the description of the above MDP.

$Q(A, right) = 0$

$Q(A, left) = Q(A, left) + \alpha(R_{t+1} + \gamma Q(B, \cdot) - Q(A, left))$

$= 0 + 1(0 + 0.9 \times -0.1 - 0)$

$= 1\ (0.9x - 0.1)$

$= -0.09$

8) After Q-learning converges (assume $\gamma = 0.9$): *1 point*

- $Q(A, left) = -0.1$ and $Q(A, right) = 0$
- $Q(A, left) = -0.09$ and $Q(A, right) = 0$
- $Q(A, left) = 0.01$ and $Q(A, right) = 0$
- $Q(A, left) > 0$ and $Q(A, right) = 0.1$
- $Q(A, left) = 1$ and $Q(A, right) = 0$
- None of these

9) Consider following statements: *1 point*

Assertion: Q-learning exhibits maximization bias.

Reason: Q-learning, for updating Q estimates, chooses the action with the highest estimate. This leads to overestimation of action values.

→ how to overcome this? ⇒ double q learning

⇓

with 1/2 probab. we choose from q_1 & q_2.

- Assertion and Reason are both true and Reason is a correct explanation of Assertion.
- Assertion and Reason are both true and Reason is not a correct explanation of Assertion.
- Assertion is true but Reason is false.
- Assertion is false but Reason is true.