RL-Week-10

PA

1) In the context of REINFORCE with baseline, consider the following statements: **1 point**

The baseline should not be a function of the −−(1)−−.

The learning rate should not be a function of the −−(2)−−.

Choose the most appropriate option.

- ( ) (1) - reward, (2) action
- (x) (1) - action, (2) - reward
- ( ) (1) - action, (2) - action
- ( ) (1) - reward, (2) - reward

baseline ✗ actions
learning rate ✗ reward

2) In the basic actor-critic setup, can we use the action value function instead of the state value function as a baseline? **1 point**

- ( ) Yes
- (x) No

3) The journey from the REINFORCE (with baseline) update rule to the actor critic update rule for the weights of the policy can be accomplished in a sequence of steps:

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \left[G_t - \hat{v}(S_t, \mathbf{w}_t)\right] \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)} \qquad (1)$$

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \left[G_t - b_t(S_t)\right] \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)} \qquad (2)$$

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \delta_t \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)} \qquad (3)$$

$$\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t)\right] \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)} \qquad (4)$$

$G_t - b(t) \longrightarrow 2$

$b = \hat{v}(s_t, w_t) \rightarrow 1$

expand it $\rightarrow 4$

but that is $\delta_t \rightarrow 3$

TD error

Order the steps in the correct sequence. Enter your answer as a four digit number.

2143 ✓

1 point

Common Data for Questions 4 to 7

Consider a shared network design that is used to represent both the actor and the critic. The network has three hidden layers, all of which are fully connected. The last hidden layer has 64 neurons. Five actions are possible from each state.

4) How many neurons would be required in the output layer of this shared network?

6

networks → 3 hidden layer
FC
last hidden has 64.
5 actions are possible

actor → 5 neurons
critic → 1 neuron
has integer output

5) What would be the activation function over the neurons in the output layer corresponding to the actor? **1 point**

- ( ) tanh
- ( ) sigmoid
- (x) softmax
- ( ) identity

for actor (states) → choose the probability distribution over the possible actions, ∴, softmax

6) What would be the activation function over the neurons in the output layer corresponding to the critic? **1 point**

- ( ) tanh
- ( ) sigmoid
- ( ) softmax
- (x) identity

Critic has 1 layer, it will be linear (identical) in nature.

7) Consider one update for both the actor and the critic using a single transition (s, a, s', r). Among all the parameters in the network, how many of them will be updated exactly once? ***1 point***

- 384
- 320
- 64
- Insufficient data

6 possible x 64 neurons in last hidden layer

384

8) In actor critic methods, the learning rate used to update the parameters of the critic should be ––– the learning rate used to update the parameters of the actor. ***1 point***

- the same as
- lower than
- higher than

η critic parameter > η actor param

9) Is the following statement true or false? ***1 point***

While gathering experience in any of the threads in A3C, the agent needs to have access to both the actor and the critic networks.

q

- True
- False

GA

1) Consider the policy gradient theorem: ***1 point***

$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a \mid s, \boldsymbol{\theta})$

Which of the following is the MC policy gradient update rule without baseline?

- $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t - \alpha G_t \dfrac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$ x
- $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha \dfrac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$ x
- $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha G_t \nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)$ x
- $\boldsymbol{\theta}_{t+1} := \boldsymbol{\theta}_t + \alpha G_t \dfrac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta}_t)}{\pi(A_t \mid S_t, \boldsymbol{\theta}_t)}$

2) Consider the four expectations given below:

$$E_\pi \left[\sum_a \pi(a \mid s, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a \mid S_t, \boldsymbol{\theta})}{\pi(a \mid S_t, \boldsymbol{\theta})} \right] \quad - (1)$$

$$E_\pi \left[G_t \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} \right] \quad - (2)$$

$$E_\pi \left[\sum_a q_\pi(S_t, a) \nabla \pi(a \mid S_t, \boldsymbol{\theta}) \right] \quad - (3)$$

$$E_\pi \left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t \mid S_t, \boldsymbol{\theta})}{\pi(A_t \mid S_t, \boldsymbol{\theta})} \right] \quad - (4)$$

policy → 3
↳ divide → 1
↓
reductions → 4
↓
G_t → 2

What is the sequence of expectations that take us from the policy gradient theorem to the MC policy gradient update? Enter your answer as a four digit number. For example, if you think the sequence is $(1) \to (2) \to (3) \to (4)$, then enter 1234 as your answer.

Consider the MC policy gradient algorithm for the full RL problem. An agent can take one of three actions from any state: a_1, a_2, a_3. The following are the values of some relevant quantities at time step t:

$$\begin{aligned} S_t &= s \\ A_t &= a_1 \\ \pi(a_1 \mid s, \boldsymbol{\theta}_t) &= 0.01 \\ G_t &= 10 \\ \alpha &= 0.1 \end{aligned}$$

Note that a return of 10 is considered to be a large return in this problem.

3) Is the following statement true or false? ***1 point***

For the policy π defined at time step t using the parameters $\boldsymbol{\theta}_t$, the action a_1 taken by the agent at state s is a surprising and highly improbable choice.

- True
- False