

**Due:** Monday 2/25/2019 at 11:59pm (submit via Gradescope).

Leave self assessment boxes blank for this due date.

**Self assessment due:** Monday 3/4/2019 at 11:59pm (submit via Gradescope)

For the self assessment, **fill in the self assessment boxes in your original submission** (you can download a PDF copy of your submission from Gradescope). For each subpart where your original answer was correct, write “correct.” Otherwise, write and explain the correct answer.

**Policy:** Can be solved in groups (acknowledge collaborators) but must be written up individually

**Submission:** Your submission should be a PDF that matches this template. Each page of the PDF should align with the corresponding page of the template (page 1 has name/collaborators, question 1 begins on page 2, etc.). **Do not reorder, split, combine, or add extra pages.** The intention is that you print out the template, write on the page in pen/pencil, and then scan or take pictures of the pages to make your submission. You may also fill out this template digitally (e.g. using a tablet.)

First name	Shenao
Last name	Zhang
SID	3034487184
Collaborators	None

# Q1. Reinforcement Learning

Imagine an unknown game which has only two states  $\{A, B\}$  and in each state the agent has two actions to choose from:  $\{\text{Up}, \text{Down}\}$ . Suppose a game agent chooses actions according to some policy  $\pi$  and generates the following sequence of actions and rewards in the unknown game:

$t$	$s_t$	$a_t$	$s_{t+1}$	$r_t$
0	A	Down	B	2
1	B	Down	B	-4
2	B	Up	B	0
3	B	Up	A	3
4	A	Up	A	-1

Unless specified otherwise, assume a discount factor  $\gamma = 0.5$  and a learning rate  $\alpha = 0.5$

(a) Recall the update function of Q-learning is:

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_{a'} Q(s_{t+1}, a'))$$

Assume that all Q-values initialized as 0. What are the following Q-values learned by running Q-learning with the above experience sequence?

$$Q(A, \text{Down}) = \underline{1}, \quad Q(B, \text{Up}) = \underline{1.75}$$

Self assessment

If your answer was correct, write “correct” above. Otherwise, write and explain the correct answer.

(b) In model-based reinforcement learning, we first estimate the transition function  $T(s, a, s')$  and the reward function  $R(s, a, s')$ . Fill in the following estimates of T and R, estimated from the experience above. Write “n/a” if not applicable or undefined.

$$\hat{T}(A, \text{Up}, A) = \underline{1}, \quad \hat{T}(A, \text{Up}, B) = \underline{0}, \quad \hat{T}(B, \text{Up}, A) = \underline{0.5}, \quad \hat{T}(B, \text{Up}, B) = \underline{0.5}$$

$$\hat{R}(A, \text{Up}, A) = \underline{-1}, \quad \hat{R}(A, \text{Up}, B) = \underline{\text{n/a}}, \quad \hat{R}(B, \text{Up}, A) = \underline{3}, \quad \hat{R}(B, \text{Up}, B) = \underline{0}$$

Self assessment

If your answer was correct, write “correct” above. Otherwise, **write and explain** the correct answer.

- (c) To decouple this question from the previous one, assume we had **a different experience** and ended up with the following estimates of the transition and reward functions:

$s$	$a$	$s'$	$\hat{T}(s, a, s')$	$\hat{R}(s, a, s')$
A	Up	A	1	10
A	Down	A	0.5	2
A	Down	B	0.5	2
B	Up	A	1	-5
B	Down	B	1	8

- (i) Give the optimal policy  $\hat{\pi}^*(s)$  and  $\hat{V}^*(s)$  for the MDP with transition function  $\hat{T}$  and reward function  $\hat{R}$ .  
*Hint: for any  $x \in \mathbb{R}$ ,  $|x| < 1$ , we have  $1 + x + x^2 + x^3 + x^4 + \dots = 1/(1 - x)$ .*

$$\hat{\pi}^*(A) = \underline{\text{Up}}, \quad \hat{\pi}^*(B) = \underline{\text{Down}}, \quad \hat{V}^*(A) = \underline{20}, \quad \hat{V}^*(B) = \underline{16}.$$

- (ii) If we repeatedly feed this new experience sequence through our Q-learning algorithm, what values will it converge to? Assume the learning rate  $\alpha_t$  is properly chosen so that convergence is guaranteed.
- ☒ the values found above,  $\hat{V}^*$
  - ☐ the optimal values,  $V^*$
  - ☐ neither  $\hat{V}^*$  nor  $V^*$
  - ☐ not enough information to determine

Self assessment

If your answer was correct, write “correct” above. Otherwise, **write and explain** the correct answer.

## Q2. Policy Evaluation

In this question, you will be working in an MDP with states  $S$ , actions  $A$ , discount factor  $\gamma$ , transition function  $T$ , and reward function  $R$ .

We have some fixed policy  $\pi : S \rightarrow A$ , which returns an action  $a = \pi(s)$  for each state  $s \in S$ . We want to learn the  $Q$  function  $Q^\pi(s, a)$  for this policy: the expected discounted reward from taking action  $a$  in state  $s$  and then continuing to act according to  $\pi$ :  $Q^\pi(s, a) = \sum_{s'} T(s, a, s') [R(s, a, s') + \gamma Q^\pi(s', \pi(s'))]$ . The policy  $\pi$  will not change while running any of the algorithms below.

(a) Can we guarantee anything about how the values  $Q^\pi$  compare to the values  $Q^*$  for an optimal policy  $\pi^*$ ?

- ☒  $Q^\pi(s, a) \leq Q^*(s, a)$  for all  $s, a$
- ☐  $Q^\pi(s, a) = Q^*(s, a)$  for all  $s, a$
- ☐  $Q^\pi(s, a) \geq Q^*(s, a)$  for all  $s, a$
- ☐ None of the above are guaranteed

Self assessment

If your answer was correct, write “correct” above. Otherwise, write and explain the correct answer.

(b) Suppose  $T$  and  $R$  are *unknown*. You will develop sample-based methods to estimate  $Q^\pi$ . You obtain a series of *samples*  $(s_1, a_1, r_1), (s_2, a_2, r_2), \dots (s_T, a_T, r_T)$  from acting according to this policy (where  $a_t = \pi(s_t)$ , for all  $t$ ).

(i) Recall the update equation for the Temporal Difference algorithm, performed on each sample in sequence:

$$V(s_t) \leftarrow (1 - \alpha)V(s_t) + \alpha(r_t + \gamma V(s_{t+1}))$$

which approximates the expected discounted reward  $V^\pi(s)$  for following policy  $\pi$  from each state  $s$ , for a learning rate  $\alpha$ .

Fill in the blank below to create a similar update equation which will approximate  $Q^\pi$  using the samples. You can use any of the terms  $Q, s_t, s_{t+1}, a_t, a_{t+1}, r_t, r_{t+1}, \gamma, \alpha, \pi$  in your equation, as well as  $\sum$  and  $\max$  with any index variables (i.e. you could write  $\max_a$ , or  $\sum_a$  and then use  $a$  somewhere else), but no other terms.

$$Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha [r_t + \gamma \max_a Q(s_{t+1}, a_{t+1})]$$

(ii) Now, we will approximate  $Q^\pi$  using a linear function:  $Q(s, a) = \sum_{i=1}^d w_i f_i(s, a)$  for weights  $w_1, \dots, w_d$  and feature functions  $f_1(s, a), \dots, f_d(s, a)$ .

To decouple this part from the previous part, use  $Q_{\text{samp}}$  for the value in the blank in part (i) (i.e.  $Q(s_t, a_t) \leftarrow (1 - \alpha)Q(s_t, a_t) + \alpha Q_{\text{samp}}$ ).

Which of the following is the correct sample-based update for each  $w_i$ ?

- ☐  $w_i \leftarrow w_i + \alpha [Q(s_t, a_t) - Q_{\text{samp}}]$
- ☐  $w_i \leftarrow w_i - \alpha [Q(s_t, a_t) - Q_{\text{samp}}]$

- ☒  $w_i \leftarrow w_i + \alpha[Q(s_t, a_t) - Q_{samp}]f_i(s_t, a_t)$
- ☐  $w_i \leftarrow w_i - \alpha[Q(s_t, a_t) - Q_{samp}]f_i(s_t, a_t)$
- ☐  $w_i \leftarrow w_i + \alpha[Q(s_t, a_t) - Q_{samp}]w_i$
- ☐  $w_i \leftarrow w_i - \alpha[Q(s_t, a_t) - Q_{samp}]w_i$

(iii) The algorithms in the previous parts (part i and ii) are:

- ☐ model-based      ☒ model-free

Self assessment

-----  
If your answer was correct, write “**correct**” above. Otherwise, **write and explain** the correct answer.