UNIVERSITY OF ALBERTA
CMPUT 365 Winter 2022

Final Exam

April 22, 2022

Do Not Distribute

Total: 150

Duration: 180 minutes

Question 1. [20 MARKS]

For the average-reward setting, the differential action-value function is defined as:

$$q_\pi(y,b) \doteq E_\pi\left[\sum_{k=t+1}^{\infty}\left(R_k - r(\pi)\right)|S_t = y, A_t = b\right], \forall y \in \mathcal{S}, \forall b \in \mathcal{A},$$

where the average reward is

$$r(\pi) \doteq \lim_{h\to\infty} E_\pi\left[\frac{1}{h}\sum_{t=0}^{h-1} R_{t+1}\right].$$

Derive the Bellman equation for q_π:

$$q_\pi(y,b) \doteq \sum_{y',r} p(y',r|y,b)\sum_{b'}\pi(b'|y')\left[r - r(\pi) + q_\pi(y',b')\right], \forall y \in \mathcal{S}, \forall b \in \mathcal{A},.$$

Show each step and the rules used in there.

Question 2. [20 MARKS]

Give the specification of expected Sarsa for on-policy control. Specifically, provide the update rule, and describe the behavior and the target policies.

Question 3. [20 MARKS]

Give the pseudocode of expected Sarsa for on-policy control. Use the following pseudocode of Q-learning and describe the changes needed to achieve it. Consider the target policy to be ϵ-greedy.

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        Choose A from S using policy derived from Q (e.g., ε-greedy)
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
        $S \leftarrow S'$
    until S is terminal

Question 4. [20 MARKS]
Consider the following off-policy TD(0) update for v_π for transition $S, A \sim b, S', R$ with the behavior policy b:

$$V(S) \leftarrow V(S) + \alpha\rho(A|S)\left(R + \gamma V(S') - V(S)\right).$$

Here, $\rho(a|s) = \dfrac{\pi(a|s)}{b(a|s)}, \forall s, a$, and we assume that the support of the behavior policy b covers the support of the target policy π: $\pi(a|s) > 0 \implies b(a|s) > 0, \forall s, a$.

Show that the conditional expectation of the increment above is equivalent to the expected on-policy increment:

$$E_{A\sim b}\left[\rho(A|S)\left(R + \gamma V(S') - V(S)\right)|S = s\right] = E_{A\sim\pi}\left[R + \gamma V(S') - V(S)|S = s\right].$$

Note that

$$E_{A\sim b}\left[\rho(A|S)|S = s\right] = \sum_a b(a|s)\frac{\pi(a|s)}{b(a|s)} = \sum_a \pi(a|s) = 1.$$

Question 5. [20 MARKS]

Consider a neural network, where the input *column* vector $\mathbf{s}$ is mapped linearly by the input-weight matrix $\mathbf{A}$ to $\boldsymbol{\psi} = \mathbf{A}\mathbf{s}$, which is then mapped by the activation function g to the feature vector $\mathbf{x} \doteq g(\boldsymbol{\psi})$. Then the feature vector is mapped by the output-weight matrix $\mathbf{B}$ linearly to the output vector $\hat{\mathbf{y}} \doteq \mathbf{B}\mathbf{x}$.

Here, for a vector $\mathbf{c}$, the ith element is denoted by c_i, and for a matrix $\mathbf{C}$, the element at the ith row and the jth column is denoted by $C_{i,j}$.
Recall the following gradients

$$\frac{\partial \hat{y}_k}{\partial B_{k,j}} = x_j,$$
$$\frac{\partial \hat{y}_k}{\partial A_{i,j}} = B_{k,i}\frac{\partial x_i}{\partial A_{i,j}} = B_{k,i}\frac{\partial g(\psi_i)}{\partial \psi_i} s_j.$$

Now let's consider that the activation function g is sigmoid: $g(a) = \dfrac{1}{1+e^{-a}}$, the derivative of which is $\dfrac{\partial g(a)}{\partial a} = g(a)\,(1-g(a))$.

Moreover, all weights are initialized to zero, that is the initial values of the weight matrices are set as $A_{i,j} = B_{i,j} = 0, \forall i, j$. Will both weights $\mathbf{B}$ and $\mathbf{A}$ stay at zero if we use backprop many times to update them by turns? Answer yes or no, and provide arguments supporting your answer.

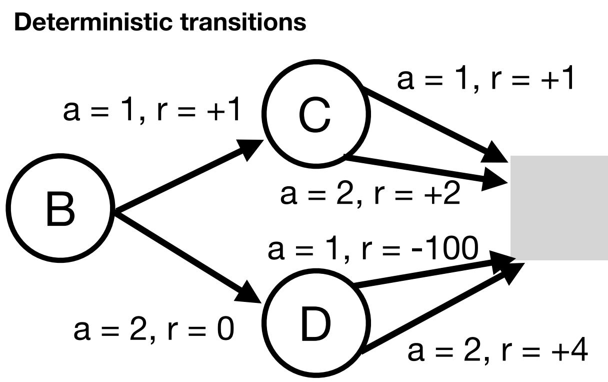
Note that for the sigmoid function, $x_j = g(\psi_j) = \dfrac{1}{1+e^{-\psi_j}}$, and $\psi_j = \sum_p A_{j,p}s_p$.

Question 6. [20 MARKS]

Consider an MDP, given by the following diagram, with three states B, C and D: $\mathcal{S} = \{B, C, D\}$, two actions: $\mathcal{A} = \{1, 2\}$, deterministic transitions, and $\gamma = 1$. Assume the action values are initialized as $Q(s, a) = 1 \ \forall s \in \mathcal{S}$ and $\forall a \in \mathcal{A}$. **Note that the values are not initialized to zero**. The agent takes actions according to an ϵ-greedy policy with $\epsilon = 0.1$. Also consider $\alpha = 0.1$.

a) Imagine the agent experienced a single episode and the following experience: $S_0 = B, A_0 = 2, R_1 = 0, S_1 = D, A_1 = 2, R_2 = 4$. What are the Sarsa updates during this episode? Start with state B, perform the Sarsa update, and then update the value of state D.

b) The agent experienced the same episode one more time. What are the Sarsa updates now during this episode? Again start with state B, perform the Sarsa update, and then update the value of state D.



Name: ____________________

Question 7. [30 MARKS]

Consider a Markov reward process consisting of a ring of three states A, B, and C, with state transitions going deterministically around the ring from A to B to C. A reward of -3 is received upon arrival in C and otherwise the reward is 1.5. What are the differential values of the three states? Use the Cesàro sum for calculating the values.

# 1	# 2	# 3	# 4	# 5	# 6	# 7	Total
/20	/20	/20	/20	/20	/20	/30	/150