UNIVERSITY OF ALBERTA
CMPUT 365 Winter 2022

Midterm Exam

Do Not Distribute

Duration: 50 minutes

Question 1. [15 MARKS]
In this question, we want you to produce identities related to conditional expectation and the law of total expectation. The law of total expectation is given by $E[X] = E[E[X|Y]]$ and can be derived as follows:

$$\begin{aligned} E[X] &= \sum_x xP(X = x) \\ &= \sum_x x \sum_y P(X = x|Y = y)p(Y = y); \text{ according to the law of total probability} \\ &= \sum_y \left(\sum_x xP(X = x|Y = y) \right) P(Y = y) \\ &= \sum_y E[X|Y = y]P(Y = y) \\ &= E[E[X|Y]]. \end{aligned}$$

Which of the following are true? Write all the correct options (just the letters).

(a) $E[X] = E[E[X|Y = y]]$

(b) $E[X|Y = y] = \sum_y yP(X = x|Y = y)$

(c) $E[X|Y = y] = \sum_x xP(X = x|Y = y)$

(d) $E[X|Y] = \sum_y E[X|Y = y]P(Y = y)$

(e) $E[E[X|Y]] = \sum_y E[X|Y = y]P(Y = y)$

(h) $E[Z] = E[E[Z|X]]$.

Question 2. [25 MARKS]

In this question, we ask you to derive a formula that relates the state value v_π to the action value q_π. Recall that $q_\pi(y, b)$ is the action value of state action pair y and b under policy π defined as the expected return:

$$q_\pi(y, b) \doteq E_\pi \left[G_t | S_t = y, A_t = b \right],$$

and $v_\pi(y)$ is the state value of state y under policy π defined as the expected return:

$$v_\pi(y) \doteq E_\pi \left[G_t | S_t = y \right].$$

If $g_\pi(y)$ is the expected reward:

$$g_\pi(y) \doteq E_\pi \left[R_{t+1} | S_t = y \right],$$

then derive the following identity:

$$v_\pi(y) = g_\pi(y) + \gamma \sum_{y'} P_\pi(S_{t+1} = y' | S_t = y) \sum_{b'} \pi(b'|y') q_\pi(y', b'), \quad \forall y,$$

where $P_\pi(S_{t+1} = y'|S_t = y)$ is the probability of next state $S_{t+1} = y'$ given the current state $S_t = y$ and $A_t \sim \pi$.

Use the linearity of expectation (LE), the law of total expectation (LOTE), the law of the unconscious statistician (LOTUS) and the Markov property (MP) in your derivation. For each step where you use one of these rules, write the name of the rule beside that step as (LE), (LOTE), (LOTUS), or (MP).

Question 3. [25 MARKS]

Suppose $\gamma = 0.8$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 5s. What are G_1 and G_0? Show the calculations and the final numbers.

Question 4. [15 MARKS]

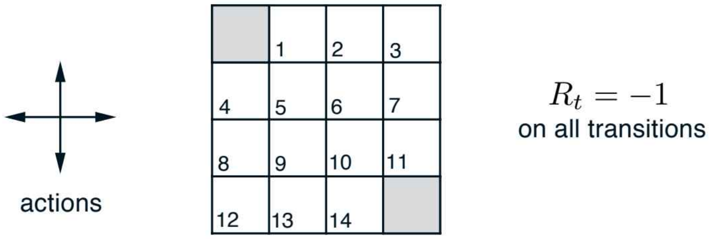
Consider the 4x4 gridworld below, where actions that would take the agent off the grid leave the state unchanged. States are given in the first table. Actions are up, down, left, and right, and rewards are -1 on all transitions. The task is episodic with $\gamma = 0.4$ and the terminal states are the shaded blocks. Using the precomputed state values v_π given in the last table for the equiprobable policy π,

A. what is $q_\pi(11, \text{down})$?

B. What is $q_\pi(7, \text{down})$?

C. What is $q_\pi(13, \text{up})$?

Show calculations and final numbers.



$k = \infty$

0.0	-14.	-20.	-22.
-14.	-18.	-20.	-20.
-20.	-20.	-18.	-14.
-22.	-20.	-14.	0.0

Name: ____________________

Question 5. [20 MARKS]

Write a complete pseudocode for iterative policy evaluation estimating the action-value function under policy μ: $Q \approx q_\mu$.

# 1	# 2	# 3	# 4	# 5	Total
/15	/25	/25	/15	/20	/100