Name: ____________________

Question 1. [20 MARKS]

Part 1: (10)

In this question, we ask you to give an extension of the law of total probability. The law of total probability applied to an unconditional probability $P(B)$ is given by:

$$P(B) = \sum_k P(B|A_k)P(A_k).$$

Here we are writing the probability of event B in an expanded form with conditional probabilities where the conditional events A_k is a partition of the sample space Ω, which amounts to the following three conditions:

$$\begin{aligned}
&A_i \neq \{\}\,\forall i;\ \text{each event is non-empty,}\\
&A_i \cap A_j = \{\}, i \neq j, \forall i, j;\ \text{mutually exclusive,}\\
&\cup_i A_i = \Omega;\ \text{collectively exhaustive.}
\end{aligned}$$

Now, instead of $P(B)$, if we want apply the law of total probability to conditional probability $P(B|C)$, what will be the formula? Write all the correct options and the corresponding formulas in your answer.

(a) $P(B|C) = \sum_k P(B|A_k)P(A_k)$

(b) $P(B|C) = \sum_k P(B|A_k \cap C)P(A_k)$ => only true when A and C are independent.

(c) ✓ $P(B|C) = \sum_k P(B|\underbrace{A_k \cap C}_{\text{A与C的交集.}})P(A_k|C).$ both have to given C.

The conditional events A_k are still a partition of the sample space here.

(c). ∵ $P(A,B|C) = P(A|B,C)P(B|C)$ ∴ c对.

这个逗号相当于 ∩

∵ $P(B) = \sum_j P(B|A)P(A)$. ∴ a不对.

Part 2: (10)

If two non-empty events D and F are *independent*, then their probabilities do not change if the other event is given as a condition: $P(D) = P(D|F)$ and $P(F|D) = P(F)$.

In part 1, if the events A_k in the partition are all independent of event C, which of the above options (a), (b) and (c) are correct? Write all the correct options and the corresponding formulas in your answer.

A与c independent. ∴ $P(A_k) = P(A_k|C)$

∴ (b)与(c)在此情况下相等.

∴ 这里选 (b). (c).

Question 2. [20 MARKS]

In this question, we ask you to derive a formula related to the Bellman equation for action value q_π. Recall that $q_\pi(x, c)$ is the action value of state action pair x and c under policy π defined as the expected return:

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

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

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

If $g(x, c)$ is the expected reward:

$$g(x, c) \doteq E\left[R_{t+1} | S_t = x, A_t = c\right],$$

then derive the following identity:

只是这种形式一律考虑用LOTUS，对 $E[g(X)] = \sum_{x \in X} g(x) \cdot P(X=x)$

$$q_\pi(x, c) = g(x, c) + \gamma \boxed{\sum_{x'} \underbrace{p(x'|x, c)}_{P(X=x)\text{ 这种就是 }X.}\underbrace{v_\pi(x')}_{g(x).}}, \quad \forall x, \forall c,$$

where $p(x'|x, c) = P(S_{t+1} = x'|S_t = x, A_t = c)$ is the probability of next state x' given the current state-action pair x, c.

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), and (MP). ⇒ 将 $q_\pi(x,c)$ 化成

$q_\pi(x, c) \doteq E_\pi[G_t | S_t = x, A_t = c]$

展开

$= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = x, A_t = c]$

$= E[R_{t+1} | S_t = x, A_t = c] + E_\pi[\gamma G_{t+1} | S_t = x, A_t = c]$ by

Linearity: $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$ ✓.

sum of two different R.V.

不是 E_π 的我们不需要通过 policy 来计算 R_{t+1} $= E[R_{t+1} | S_t = x, A_t = c] + \gamma E_\pi[G_{t+1} | S_t = x, A_t = c]$ by

Linearity: $\mathbf{E}[aX] = a\mathbf{E}[X]$ ✓

将 expectation 中的 constant 提出来. 因为它是常数.

等于 expected reward.

$= g(x, c) + \gamma E_\pi[G_{t+1} | S_t = x, A_t = c]$

$= g(x, c) + \gamma E_\pi[E_\pi[G_{t+1} | S_{t+1}, S_t = x, A_t = c] | S_t = x, A_t = c]$ by LOTE

复制而来.

加入新东西. 相当于 Y.

注: $E[X|Z] = E[E[X|Y, Z]|Z]$ ⇒ LOTE for conditional expectation.

加入的新东西.

This is known as the **law of total expectation** and also the tower rule.

这只求X的积分，不求Y的.

Again, the conditional expectation $E[X|Y]$ integrates only over the random variable X but not Y. On the other hand, in $E[E[X|Y]]$, the outer expectation integrates over the random variable Y but not X which is already integrated out.

所以求Y的积分，因为X的积分在E[X|Y]中已经被求出来了.

Interesting results

$$E[E[f(X)|Y]\,g(Y)] = E[E[f(X)g(Y)|Y]].$$

$$E[f(X, Y)g(Y)] = E[E[f(X, Y)|Y]\,g(Y)].$$

$= g(x, c) + \gamma E_\pi[E_\pi[G_{t+1} | S_{t+1}] | S_t = x, A_t = c]$ by MP.

只保留新东西. 因为根据 MP, we don't need the previous state and action to determine the current one!!

因为只需基于当前 state S_{t+1} 下所知道的条件对 G_{t+1} 进行计算.

expectation of an conditional expectation.

Show that $\mathbf{E}[X] = \mathbf{E}[\,\mathbf{E}[X \mid Y]\,]$ ⇒ law of total expectation.

R.V.

Daniel Asimiakwini 对所有人说 下午1:45

for the MP step, we can drop the variables in the inner expectation because we don't need the previous state and action to determine the current one?

Yes!!

$= g(x, c) + \gamma E_\pi[v_\pi(S_{t+1}) | S_t = x, A_t = c]$

不需知道 S_{t+1} 是多少.

$= g(x, c) + \gamma \sum_{x'} p(x'|x, c) v_\pi(x')$ by LOTUS.

因为 $S_t = x$, 所以 $S_{t+1} = x'$

Law of the unconscious statistician: $\mathbf{E}[\,g(X)\,] = \sum_{x \in \mathcal{X}} g(x)\,P(X=x)$

a R.V. which itself is a function of another R.V.

对 $E[Y]$, $Y = g(X)$.

$= \sum_{y \in Y} y\, P(Y=y)$

$= \sum g(x)\, P(g(X) = g(x))$

$= \sum_{x \in X} g(x)\, P(X=x)$

probability, ∴ 它的 input 必须是 event!! 就是 S_{t+1}, S_t, A_t 等于的东西.

即, 这里将 $[v_\pi(S_{t+1}) | S_t = x, A_t = c]$ 变为 $(x' \mid x, c)$

Question 3. [20 MARKS]

(10+10)

In this question, we ask you to give the Bellman optimality equation and value iteration update rule for action values. The Bellman optimality equation for v_* is given by:

$$v_*(y) = \max_b \sum_{y',w} p(y', w|y, b) \left[w + \gamma v_*(y')\right], \forall y, \tag{1}$$

where $p(y', w|y, b)$ is the joint probability of next state y' and reward w given the current state-action pair y, b. Then the value iteration method can be directly obtained based on the optimality equation by replacing the optimal state value v_* with estimate v_k for the kth iteration and replacing equality $=$ with assignment $\doteq$:

$$v_{k+1}(y) \doteq \max_b \sum_{y',w} p(y', w|y, b) \left[w + \gamma v_k(y')\right], \forall y. \tag{2}$$

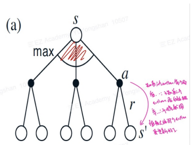
Bellman optimality equation.

Provide the Bellman optimality equation for q_* and the corresponding value iteration update rule for action value using the state-action pair y, b.

因为：

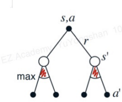
$$\begin{aligned} v_*(s) &= \max_{a\in\mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\ &= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s', r|s,a)[r + \gamma v_*(s')]. \end{aligned}$$

$V_* \to V_*$



$$\begin{aligned} q_*(s,a) &= \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \,\middle|\, S_t = s, A_t = a\right] \\ &= \sum_{s',r} p(s', r|s,a)\left[r + \gamma \underbrace{\max_{a'} q_*(s', a')}_{\text{取最大的那个}a'}\right]. \end{aligned}$$

$q_* \to q_*$



所以：$q_*(y,b) = \sum_{y',w} p(y',w|y,b)\left[w + \gamma \boxed{\max_{b'}} q_*(y',b')\right] \forall y, \forall b.$

$\rightarrow$ 与 V_* 和 V_{k+1} 的唯一不同之处.

$q_{k+1}(y,b) \doteq \sum_{y',w} p(y',w|y,b)\left[w + \gamma \boxed{\max_{b'}} q_k(y',b')\right] \forall y, \forall b.$

Question 4. [20 MARKS]

Prove that the discounted sum of rewards is always finite if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $0 < R_{\max} < \infty$:

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty; \qquad \text{for } \gamma \in [0, 1).$$

$\because |R_{t+1}| \leq R_{max}$

$\therefore R_{t+1} \leq R_{max}.$

$\because$ 在有 γ 的情况下, $\sum_{i=0}^{\infty} \gamma^i R_{t+1+i} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

根据这个条件得到.

$\leq R_{max} + \gamma R_{max} + \gamma^2 R_{max} + \cdots$

提取公因子 R_{max}.

$= R_{max}(1 + \gamma + \gamma^2 + \cdots)$

$= \boxed{\dfrac{R_{max}}{1-\gamma}}$ → 如果R一直等于 R_{max}.

$< \infty$

$\therefore \left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

- Continuing Task 有γ: [illegible]

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \qquad 0 \leq \gamma \leq 1$$

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

如果R一直等于1： $G = \dfrac{1}{1-\gamma}$

Question 5. [20 MARKS]

In this question, we ask you to amend incorrect statements. In the following, there are two incorrect statements. Write mathematical equations or expressions that will make the statements correct and briefly list the changes you made on the provided equations or expressions.

Part 1: (10)

The Bellman equation for action value can be written as:

改为 =

$$q_\pi(s,a) \doteq \sum_{s',r,s,a'} p(s',r|s,a)\pi(a'|s')\left[r + \gamma q_\pi(s',a')\right], \forall s, \forall a.$$

前面缺了求和 $\sum_{a'}$

function括号里没有的东西就必须出现在求和下面!!

∴ 求和下面也不应该有function括号里有的东西!!

Part 2: (10)

The optimal state value can be related to the optimal action value in the following way:

$$v_*(s) \le \max_a q_\pi(s,a), \forall s, \forall a.$$

Part 1. 因为 $q_\pi(s,a) \doteq \mathbb{E}_\pi\left[G_t \mid S_t = s, A_t = a\right]$

$$= \sum_{s'}\sum_{r} p(s',r \mid s,a)\left[r + \gamma \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s'\right]\right]$$

$$= \sum_{s'}\sum_{r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') \mathbb{E}_\pi\left[G_{t+1} \mid S_{t+1} = s', A_{t+1} = a'\right]\right]$$

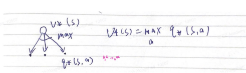
$$= \sum_{s'}\sum_{r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\right]$$

$$v_\pi(s) = \sum_{a} \pi(a \mid s) \sum_{s'}\sum_{r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

$$q_\pi(s,a) = \sum_{s'}\sum_{r} p(s',r \mid s,a)\left[r + \gamma \sum_{a'} \pi(a' \mid s') q_\pi(s',a')\right]$$

所以 $q_\pi(s,a) = \sum_{s',r} p(s',r|s,a) \sum_{a'} \pi(a'|s')\left[r + \gamma q_\pi(s',a')\right], \forall s, \forall a.$

Part 2. 因为:



所以: $v_*(s) = \max_a q_*(s,a)$

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