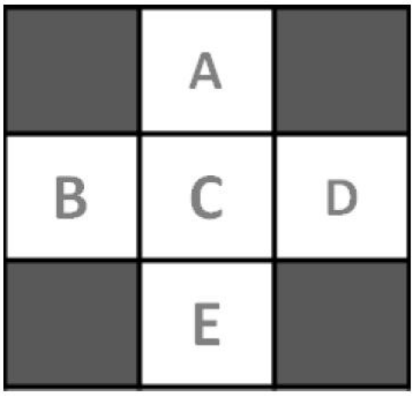
习题三：强化学习和贝叶斯网络（共 60 分）

1、网格世界中的学习（10 分）

考虑课堂中遇到的网格世界，我们想要用 TD 学习和 Q 学习来找到这些状态的值。



假设我们观察到下面的几个状态转换。

(B, East, C, 2), (C, South, E, 4), (C, East, A, 6), (B, East, C, 2)

每个状态的初始值是 0，假设$\gamma = 1$ 和$\alpha = 0.5$。

a) 基于 TD 学习，经过上面的观测后，我们学习到的值分别是什么？（5 分）

$V(B) = 3.5$
$V(C) = 4$
All other states have a value of 0.

b) 基于 Q 学习，经过上面的观测后，我们学习到的 Q 值分别是什么？（5 分）

$Q(B, East) = 3$
$Q(C, South) = 2$
$Q(C, East) = 3$
All other q-states have a value of 0.

2、基于特征的 Q-Learning 吃豆人（20 分）

我们想设计一个 Q -Learning 的吃豆人，但大型网格的状态空间太大，内存中无法容纳。为了解决这个问题，我们切换到基于特征的状态表示。

a) 我们会有两个特征，F_g和F_p，其定义如下：

$$F_g(s,a) = A(s) + B(s,a) + C(s,a)$$
$$F_p(s,a) = D(s) + 2E(s,a)$$

其中

$A(s)$ = 离状态s只有一步之遥的鬼魂数量

$D(s)$ = 离状态s只有一步之遥的食物数量

$B(s,a)$ = 吃豆人从s出发，采取动作a后，接触到的鬼魂数量

$C(s,a)$ = 吃豆人从s出发，采取动作a后，离他只有一步之遥的鬼魂数量

$E(s,a)$ = 吃豆人从s出发，采取动作a后，能够吃到的粮食数量

在这个吃豆人的游戏板上，鬼魂永远是静止的，吃豆人的动作空间是{左、右、上、下、停止}。

从上图所示的当前状态，计算{左、右、上、停止} 几个动作的特征值$F_g(s,a)$和$F_p(s,a)$。（8 分）

$$F_p(s,up) = 1 + 2(1) = 3$$
$$F_p(s,left) = 1 + 2(0) = 1$$
$$F_p(s,right) = 1 + 2(0) = 1$$
$$F_p(s,stay) = 1 + 2(0) = 1$$
$$F_g(s,up) = 2 + 0 + 0 = 2$$
$$F_g(s,left) = 2 + 1 + 1 = 4$$
$$F_g(s,right) = 2 + 1 + 1 = 4$$
$$F_g(s,stay) = 2 + 0 + 2 = 4$$

b) 经过几轮 Q 学习后，特征值的权重是$w_g = -10$ 和$w_p = 100$，计算从上图状态出发，采取{左、右、上、停止} 几个动作后，分别得到的 Q 值是什么？（4 分）

$$Q(s,a) = w_g F_g(s,a) + w_p F_p(s,a)$$

$$Q(s,up) = w_p F_p(s,up) + w_g F_g(s,up) = 100(3) + (-10)(2) = 280$$
$$Q(s,left) = w_p F_p(s,left) + w_g F_g(s,left) = 100(1) + (-10)(4) = 60$$
$$Q(s,right) = w_p F_p(s,right) + w_g F_g(s,right) = 100(1) + (-10)(4) = 60$$
$$Q(s,stay) = w_p F_p(s,stay) + w_g F_g(s,stay) = 100(1) + (-10)(4) = 60$$

c) 从上图中的状态出发，我们观察到吃豆人采取一个向上的行动，以状态s'结束（上面食物颗粒的位置）并获得奖励 $R(s,a,s')=250$。从状态 s'出发，吃豆人可用的操作只有向下和停止两种。假设折扣为 γ = 0.5，基于本次观测，计算新的 Q 值。（4 分）

$$\begin{aligned} Q_{new}(s,a) &= R(s,a,s') + \gamma * \max_{a'} Q(s',a') \\ &= 250 + 0.5 * \max\{Q(s',down), Q(s',stay)\} \\ &= 250 + 0.5 * 0 \\ &= 250 \end{aligned}$$

where

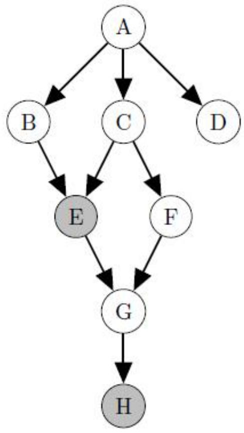
$$Q(s',down) = w_p F_p(s',down) + w_g F_g(s',down) = 100(0) + (-10)(2) = -20$$
$$Q(s',stay) = w_p F_p(s',stay) + w_g F_g(s',stay) = 100(0) + (-10)(0) = 0$$

d) 基于新的 Q 估计，假设学习比率（learning rate）$\alpha=0.5$，更新每个特征的权重。（4 分）

$$w_p = w_p + \alpha * (Q_{new}(s,a) - Q(s,a)) * F_p(s,a) = 100 + 0.5 * (250 - 280) * 3 = 55$$
$$w_q = w_q + \alpha * (Q_{new}(s,a) - Q(s,a)) * F_q(s,a) = -10 + 0.5 * (250 - 280) * 2 = -40$$

3、贝叶斯网络：推理（15 分）

假设我们有以下的贝叶斯网络，并希望通过推理以获得$P(B,D|E=e,H=h)$。



a) 对此查询$P(B,D|E=e,H=h)$，通过枚举推理生成的最大因子的行数是多少？假设所有变量都是二进制的。（3 分）

○ 2^2 ○ 2^3 ● 2^6 ○ 2^8
○ None of the above.

Since the inference by enumeration first joins all the factors in the Bayes' net, that factor will contain six (unobserved) variables. The question assumes all variables are binary, so the answer is 2^6.

b) 标记以下所有最适合计算$P(B,D|E=e,H=h)$的变量消除顺序。优越性是通过产生的因子多少之和来衡量的。假设所有变量都是二进制的。（4 分）

☐ C,A,F,G ☐ F,G,C,A ☐ A,C,F,G ■ G,F,C,A
☐ None of the above.

The sum of the sizes of factors that are generated for the variable elimination ordering G, F, C, A is $2^1+2^1+2^2+2^2$ rows, which is smaller than for any of the other variable elimination orderings. The ordering F, G, C, A is close but the sum of the sizes of factors is slightly bigger, with $2^2+2^1+2^2+2^2$ rows.

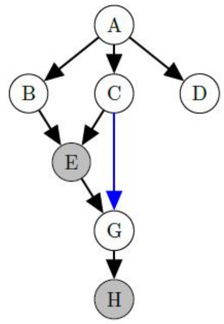
c) 假设我们决定通过变量消除来计算$P(B,D|E=e,H=h)$，并选择先消除 F。

1) 当F被消除时，产生什么中间因子，它是如何计算的？确保你很清楚哪些变量在条件栏之前，哪些变量在条件栏之后。（4 分）

$$f_1(\underline{\quad G\mid C,e \quad}) = \sum_f \underline{\quad P(f\mid C)P(G|f,e) \quad}$$

This follows from the first step of variable elimination, which is to join all factors containing F, and then marginalize over F to obtain the intermediate factor f_1.

2) 现在考虑 F 被消除之后的剩余因子可以表示的概率分布。在以下贝叶斯网络结构上绘制最少数量的有向边，让它可以表示此集合中的任何分布。如果不需要额外的定向边缘，请选择下面的选项。（4 分）

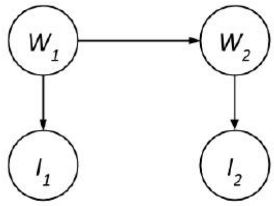


○ No additional directed edges needed

An additional edge from C to G is necessary, because the intermediate factor is of the form $f_1(G|C)$. Without this edge from C to G, the Bayes' net would not be able to express the dependence of G on C. (Note that adding an edge from G to C is not allowed, since that would introduce a cycle.)

4、采样和动态贝叶斯网络（15 分）

我们想分析人们在晴天和雨天的吃冰淇淋的习惯。假设在两天的时间里，我们考虑天气，以及一个人吃冰淇淋的时间。我们将有四个随机变量：W_1 和 W_2 代表第 1 天和第 2 天的天气，可以是下雨 R 或晴天 S，变量 I_1 和 I_2 表示该人是否在第 1 天和第 2 天吃了冰淇淋，取值 T（表示真正吃冰淇淋）或 F。我们可以将其建模为具有这些概率的如下所示的贝叶斯网络。



W_1	$P(W_1)$
S	0.6
R	0.4

W_1	W_2	$P(W_2 \mid W_1)$
S	S	0.7
S	R	0.3
R	S	0.5
R	R	0.5

W	I	$P(I \mid W)$
S	T	0.9
S	F	0.1
R	T	0.2
R	F	0.8

假设我们从上面的冰淇淋模型中产生了如下采样(W_1, I_1, W_2, I_1)：

R, F, R, F   R, F, R, F   S, F, S, T   S, T, S, T   S, T, R, F
R, F, R, T   S, T, S, T   S, T, S, T   S, T, R, F   R, F, S, T

a) 采样分配给事件$W_2 = R$的概率$P(W_2 = R)$是什么？（3 分）

Number of samples in which $W_2 = \mathtt{R}$: 5. Total number of samples: 10. Answer $5/10 = 0.5$.

b) 假设我们计算$P(W_2 \mid I_1 = T, I_2 = F)$，划掉上面的样本中会被拒绝抽样（rejection sampling）排除的样本。（3 分）

~~R, F, R, F~~   ~~R, F, R, F~~   ~~S, F, S, T~~   ~~S, T, S, T~~   S, T, R, F
~~R, F, R, T~~   ~~S, T, S, T~~   ~~S, T, S, T~~   S, T, R, F   ~~R, F, S, T~~

c) 拒绝抽样似乎会浪费很多精力，所以我们改用似然加权。假设我们在给定证据$I_1 = T, I_2 = F$的情况下生成以下六个样本：

$$(W_1, I_1, W_2, I_2) = \Big\{(\mathtt{S},\mathtt{T},\mathtt{R},\mathtt{F}), (\mathtt{R},\mathtt{T},\mathtt{R},\mathtt{F}), (\mathtt{S},\mathtt{T},\mathtt{R},\mathtt{F}), (\mathtt{S},\mathtt{T},\mathtt{S},\mathtt{F}), (\mathtt{S},\mathtt{T},\mathtt{S},\mathtt{F}), (\mathtt{R},\mathtt{T},\mathtt{S},\mathtt{F})\Big\}$$

上面第一个样本(S, T, R, F)的权重是什么？（3 分）

The weight given to a sample in likelihood weighting is

$$\prod_{\text{Evidence variables } e} \Pr(e \mid \text{Parents}(e)).$$

In this case, the evidence is $I_1 = \mathtt{T}, I_2 = \mathtt{F}$. The weight of the first sample is therefore

$$w = \Pr(I_1 = \mathtt{T} \mid W_1 = \mathtt{S}) \cdot \Pr(I_2 = \mathtt{F} \mid W_2 = \mathtt{R}) = 0.9 \cdot 0.8 = 0.72$$

d) 使用似然加权估计$P(W_2 \mid I_1 = T, I_2 = F)$。（6 分）

The sample weights are given by

(W_1, I_1, W_2, I_2)	w	(W_1, I_1, W_2, I_2)	w
S, T, R, F	0.72	S, T, S, F	0.09
R, T, R, F	0.16	S, T, S, F	0.09
S, T, R, F	0.72	R, T, S, F	0.02

To compute the probabilities, we thus normalize the weights and find

$$\widehat{P}(W_2 = \mathtt{R} | I_1 = \mathtt{T}, I_2 = \mathtt{F}) = \frac{0.72 + 0.16 + 0.72}{0.72 + 0.16 + 0.72 + 0.09 + 0.09 + 0.02} = 0.889$$

$$\widehat{P}(W_2 = \mathtt{S} | I_1 = \mathtt{T}, I_2 = \mathtt{F}) = 1 - 0.889 = 0.111.$$