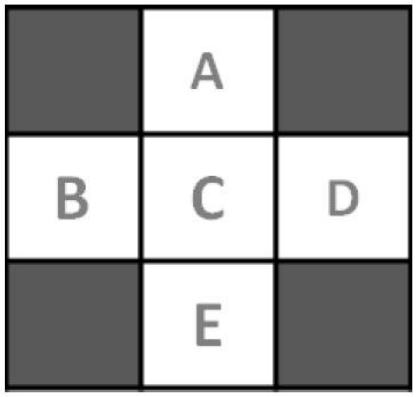
习题三：强化学习和贝叶斯网络（共 60 分）

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）

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）

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）

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）

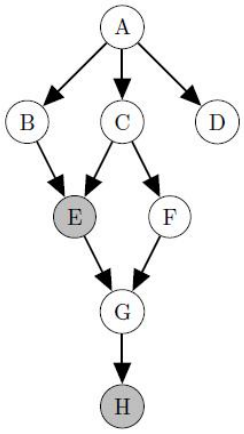
$$Q(s,a) = w_g F_g(s,a) + w_p F_p(s,a)$$

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 $\gamma = 0.5$，基于本次观测，计算新的 Q 值。（4 分）

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）

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.

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）

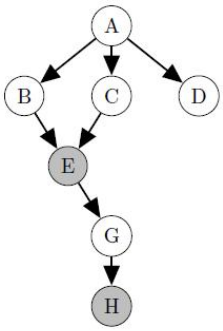
☐ C, A, F, G ☐ F, G, C, A ☐ A, C, F, G ☐ G, F, C, A
☐ None of the above.

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）

$f_1($_____ | _____$) = \sum_f$ _____

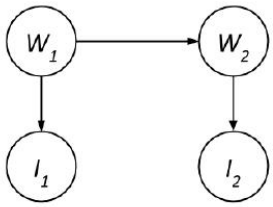
2) 现在考虑 F 被消除之后的剩余因子可以表示的概率分布。在以下贝叶斯网络结构上绘制最少数量的有向边，让它可以表示此集合中的任何分布。如果不需要额外的定向边缘，请选择下面的选项。（4 分）



○ No additional directed edges needed

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

W	I	$P(I\|W)$
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）

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

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：

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

上面第一个样本(S, T, R, F)的权重是什么？（3 分）

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