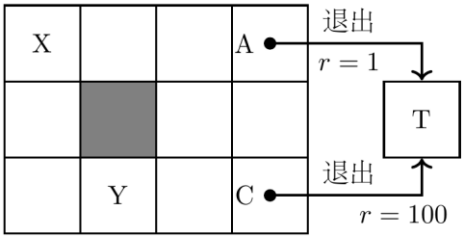
作业 3，人工智能导论课 （2023 春季学期）

MDP，增强学习（RL）

1.在下图所示的一个网格游戏里，灰色的方格表示墙，在所有其他的状态里，智能体都有 4 种基本的行动：上，下，左，右。如果一个行动导致一个无效的状态（比如跃出了网格或被墙阻挡）则保持原来的状态。在状态 A 和 C 中，智能体还有一个额外的动作“退出”，执行退出会收获相应的奖励值r，并结束游戏到达终局状态T。所有其他的状态转移都是确定性的，并且没有奖励值。假设所有初始状态值为 0，折扣率$\gamma = 1/2$。



请用值迭代算法计算这里的状态 X 和 Y 的最优状态值，即$V^*(X)$和$V^*(Y)$？在这个游戏里，该算法在多少次迭代后会收敛？或是不会收敛？

2.在骰子游戏中，玩家可以投掷一个骰子，该骰子有 6 个面分别是数字 1 到 6，并且任何一个面作为结果出现的概率是相同的。每次投掷需花费 1 元钱。玩家在开始游戏时必须先要投掷一次骰子，然后接下来该玩家有两个行动选择：

- 停止：终止该游戏，并领取骰子显示数字的相应数量的金钱，比如骰子显示 6，则领取 6 元钱。
- 继续：继续投掷一次骰子，并花费 1 元钱。

你用所学的马科夫决策过程（MDP）对这个问题建模分析。玩家初始时在开始状态，只有一个行动，即继续骰子。当玩家采取停止行动时，则退出游戏。

1） 如果你采用策略迭代算法来求解这个 MDP 问题，下面的表格里给出了各个状态及其初始策略π。请评估该策略在每个状态上的值，假设$\gamma = 1$。

s	1	2	3	4	5	6
$\pi(s)$	继续	继续	停止	停止	停止	停止
$\mathbf{V}^{\pi}(\mathbf{s})$						

2） 接下来你需要通过计算后的状态值，更新刚才的策略π，计算策略π'。注意，如果“继续”与“停止”都是合理的行动，那么把这两个行动都写下来。这里依然假设$\gamma = 1$。

s	1	2	3	4	5	6
$\pi(s)$	继续	继续	停止	停止	停止	停止
$\pi'(s)$						

3) 策略$\pi(s)$是否是最优策略？请解释为什么？

3. 增强学习算法

假设一个简单的吃豆子游戏是建立在一个未知的 MDP 模型上，这里有 3 个状态A, B, C,两个动作{停止，前进}。给定下列样本序列，请回答以下的问题，假设$\gamma = 1, \alpha = 0.5$。

1). 让我们利用下面的样本，运行 Q-learning 算法，计算以下 Q 状态值：

s	行动	s’	r
A	前进	B	2
C	停止	A	0
B	停止	A	-2
B	前进	C	-6
C	前进	A	2
A	前进	A	-2

利用 Q-learning 算法估计以下的 Q 状态值是多少？所有 Q 状态值初始化为 0.

a) $Q(C,$停止$)=$

b) $Q(C,$前进$)=$

2). 现在，让我们用基于特征的表达方式来表示 Q 状态值。假设我们使用以下两种特征函数：

- $f_1(s,a) = 1$

$$f_2(s,a) = \begin{cases} 1 & a = \text{前进} \\ -1 & a = \text{停止} \end{cases}$$

初始化的权值均为 0，在观察到以下样本后，计算更新后的权值：

s	行动	s’	r
A	前进	B	4
B	停止	A	0

在第一个样本之后，权值是多少？

a) $w_1 =$

b) $w_2 =$

继续获得第二个样本后，权值变为多少？

c) $w_1 =$

d) $w_2 =$