《强化学习》课程作业（一）

2024 年 4 月 23 日 18:10 前提交至邮箱gaojie@ios.ac.cn

课本习题 练习 2.2，2.3，3.8，3.9，3.14，3.15，3.22，3.24，3.26（注意，题号以第二版中文版教材为准）

计算题 考虑一个二臂赌博机，该赌博机有两个模式：正常模式 s_0和作弊模式 s_1。正常模式游玩一局后，有 0.2 的概率切换到作弊模式，切换到作弊模式时机器发出声音提示，作弊模式游玩一局后以概率 1 切换回正常模式。正常模式下，使用臂 1 以概率 0.4 获得收益 1，以概率 0.6 获得收益 -1，使用臂 2 以概率 0.3 获得收益 2，以概率 0.7 获得收益 -1。作弊模式下，使用臂 1 以概率 0.5 获得收益 4，以概率 0.5 获得收益 -1，使用臂 2 以概率 1 获得收益 3。

1. 画出该马尔可夫决策过程的状态转移图。

2. 计算 $r(s_0, a)$，$r(s_0, a, s_0)$ 和 $r(s_0, a, s_1)$，其中 a 指使用臂 1 的行为。根据你的计算过程，写出一般情况下，收益函数 $r(s, a)$ 和 $r(s, a, s')$ 的关系。

3. 设 π 为只使用臂 2 的策略，考虑持续性任务，折扣率 $\gamma = 0.8$，计算 $v_\pi(s_0)$。

编程题 * 考虑课本例 3.5 的网格问题。其他条件不变，状态 A 不再能迁移到 A′，而是四种动作都会以 0.7 的概率迁移到到 A′左边相邻的状态，以 0.3 的概率迁移到到 A′右边相邻的状态，其收益分别为 7 和 13。

1. 实现这个马尔可夫决策过程，你可以参考这里的代码：
`https://github.com/ShangtongZhang/reinforcement-learning-an-introduction`

2. 如果在每个状态，都以相同的概率执行四种动作，编程求出该策略下的状态价值函数，画成课本图 3.2 右图的形式。

3. 通过设计并在程序中尝试其他策略，寻找你认为的最优策略，求出该策略下的状态价值函数，将你找出的最优策略和状态价值函数画成课本图 3.5 的形式。

你需要提交源代码和在第 2 问和第 3 问中画出的图。本题选做，提交的同学可以获得最多 2 分的额外加分。