《强化学习》课程作业（二）

2024 年 4 月 30 日 18:10 前提交至邮箱 gaojie@ios.ac.cn

课本习题 练习 4.1，4.2，4.3，4.4，4.5（注意，题号以第二版中文版教材为准）

计算题 考虑课本例 4.1，不同的是，我们现在面对的是 3×3 的网格图，左上角和右下角的格子是终止状态。从非终止状态到非终止状态的迁移收益均为 -1，从非终止状态到终止状态的迁移收益为 1。设 π 是等概率随机策略。

1. 用迭代策略评估算法估计 v_π，仿照课本图 4.1 画出第 1 到 3 次迭代的结果。

2. 利用贝尔曼方程建立线性方程组，求解准确的 v_π，直接写出或画出计算的结果，本小题你可以借助计算器或计算机程序辅助计算。

3. 画出第 1 问迭代过程中，每一步价值函数估计对应的贪心策略。

编程题 * 考虑课堂上讲的赌徒模型。持有金币数大于等于 N 时胜利退场，金币全部输光时失败退场。每局游戏可以押上不超过当前持有金币数量的正整数枚金币，每局游戏胜利的概率是 p，不同局的游戏结果相互独立。实现这个马尔可夫决策过程，你可以参考 github 中的代码（见第一次作业中的链接）。注意，我们的目标是让胜利退场的概率最大，因此需要根据此目标设定合适的收益和折扣率。

1. $N = 200$，$p = 0.6$ 时，用迭代策略评估算法估计以下策略的状态价值函数：(1) All-in，即持有的金币数不大于 $0.5N$ 时下注全部金币，否则下注 N 与持有的金币数的差值；(2) One-dollar，即每次下注 1 枚金币；(3) Two-dollar，即每次下注 2 枚金币（除非你只剩下 1 枚或 $N-1$ 枚金币，此时下注 1 枚）。用折线图画出估计出的三个状态价值函数，并比较这三个策略的优劣。

2. $N = 179$，$p = 0.4$ 时，用迭代策略评估算法算法估计 All-in 策略的状态价值函数，用折线图画出估计的结果。此时是不是存在比 All-in 策略更好的策略，请用策略迭代算法寻找最优策略。

本题你需要提交源代码和画出的图，以及策略比较的结论。本题选做，提交的同学可以获得最多 2 分的额外加分。