《强化学习》课程作业（三）

2024 年 5 月 21 号 18：10 前提交

课本习题 练习 5.1，5.2，5.3，5.5，5.6，5.7，6.3，6.4，6.5，6.6

*** 编程题** 本题选做，提交的同学可以获得最多作业总分的奖励。考虑上次作业编程题中实现的赌徒问题，设定 $N = 190$，$p = 0.49$。

1. 用同轨策略的蒙特卡洛算法和时序差分算法 TD(0) 计算以下策略的状态价值函数：(1) One-dollar；(2) Two-dollar；(3) All-in。画出你算出的结果。

2. 设 π 是随机下注策略，即拥有 n 枚金币时以等概率投注 $1, \ldots, \min(n, N-n)$ 枚金币。在策略 π 下用离轨的蒙特卡洛算法计算以下策略的状态价值函数：(1) One-dollar；(2) Two-dollar；(3) All-in。画出你算出的结果。

3. 使用合适的控制算法，找出一个比以上策略都更好的策略。画出你找到的策略和对该策略的评估。

本题你需要提交源代码，代码运行的配置说明和各小题画出的图。