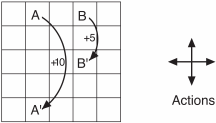
**Homework** **#1**

Please submit your homework **before** **23:00,** **April** **6,** **2022**. All delayed submissions will not be accepted.

**Problem** **1** (Gridworld)**.** Figure shows a rectangular gridworld representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are possible: **north**, **south**, **east**, and **west**, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of − 1. Other actions result in a reward of 0, except those that move the agent out of the special states **A** and **B**. From state **A**, all four actions yield a reward of +10 and take the agent to **A**′ . From state **B**, all actions yield a reward of +5 and take the agent to **B**′ . Suppose the agent selects all four actions with equal probability in all states. This policy is denoted as π . Let the discounted factor γ be 0.9.



(1) Under policy π, please compute the value of states **A** and **B**, i.e., vπ(**A**) and vπ(**B**).

(2) Prove that adding a constant c to all the rewards adds a constant vc to the values of all states, and thus does not affect the relative values of any states under any policies.

(3) What is vc in terms of c and γ?

(4) Are the signs of rewards important here, or only the intervals between rewards?

1. Could you provide a new policy which is better than π?

解：(1) 设给5\*5大小的网格中每个格子标记为状态, 第i行第j列的格子对应状态,

其中

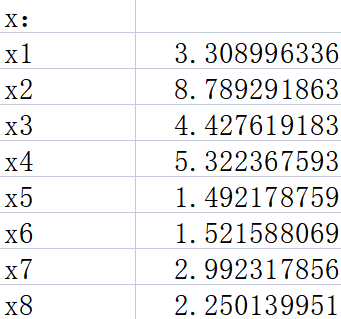
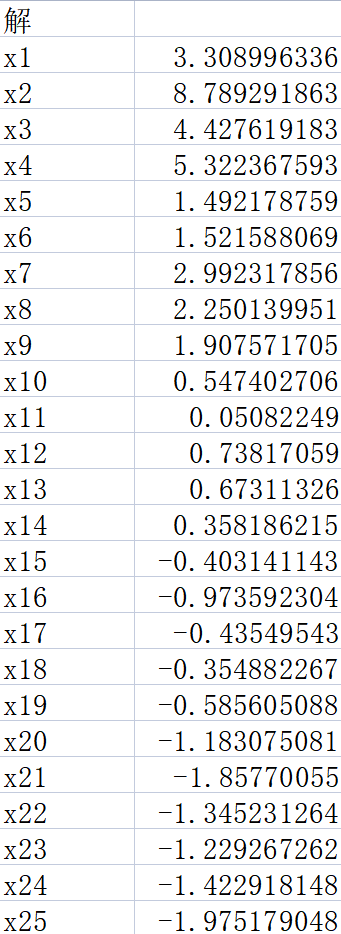
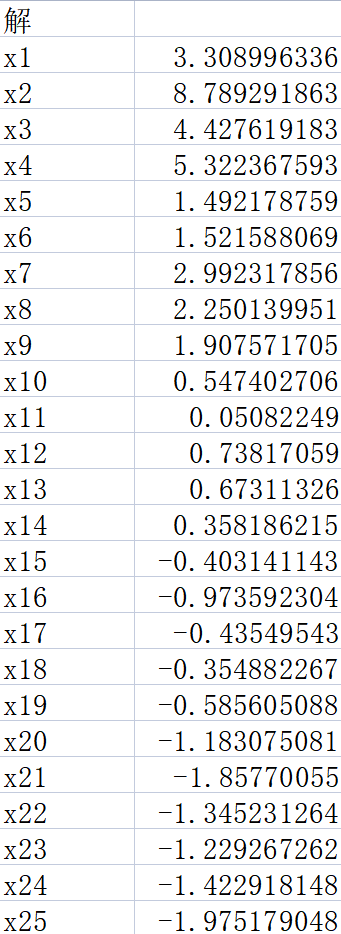
由 得25元一次方程如下：

令，写成矩阵形式为Ax=b，A为25\*25方阵，x=

A=

b=,

x=

x=

所以

加上一个常数c后，

因为 所以 , = +

由此可见，对每个收益值都加上一个常数c后就等于对每个状态价值都加上一个常数

1. 由(2)知，
2. 是rewards的符号重要，对于连续性任务而言，若原本负收益的动作变为正收益，会影响到智能体的动作选择策略，就可能与任务的最终目标背道而驰
3. 是一种等概率随机策略，这就导致边界处出界的概率较高，这样出界的期望惩罚值较高，可以修改策略在

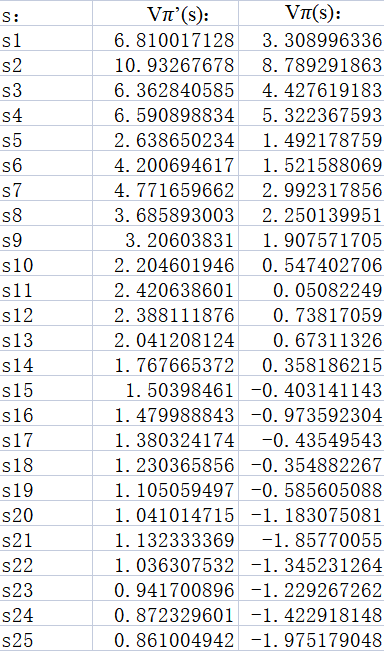
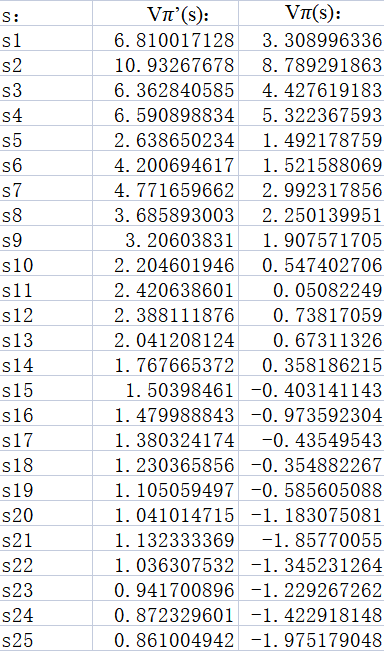
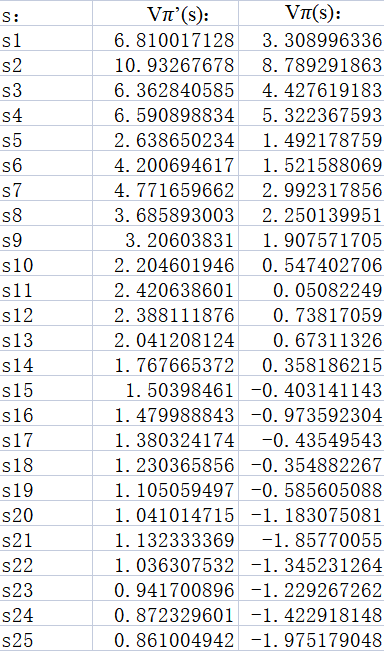
边界处时选择走到网格内格子的动作概率高于出界动作的概率，这样的策略总体期望收益会有所提升。

举个例子：同理(1)中的状态标记，策略将所有出界动作的概率设为0，其余非出界动作等概率。即，

从而这样的策略构建的Ax=b方程，b列矩阵的元素均非负，又n阶方阵A的顺序主子式全为正，且非对角元全为负，可得它的逆矩阵的每个元素全为正，这样计算得到的状态价值函数均为正，显然，

对于的s来说，,

对于的s来说，由于出界动作的概率为0，不会有负收益的惩罚，

计算结果  

故