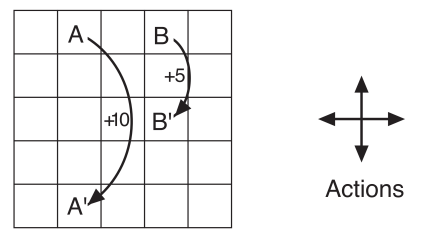
**Homework** **#1**

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?

Solution：(1) Each grid is marked as a state

the grid that lies in the i-th row and the j-th corresponds state

and

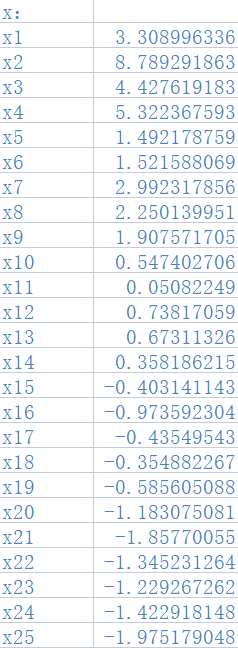
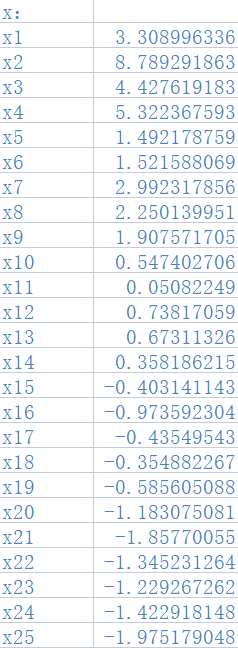
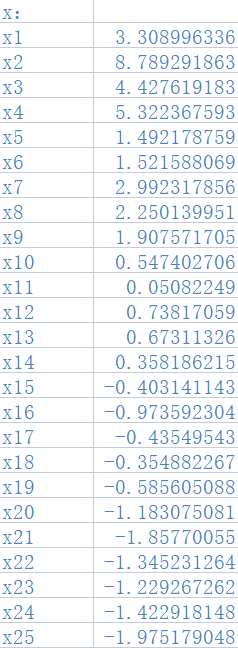
Then, according to

we can obtain a 25-variable linear equation as followed：

Let, in a matrix form as Ax=b，A is 25 order square matrices，x=

A=

,

∴

After adding a constant c，

∵ ∴ , = +

Thus，adding a constant c to all the rewards adds a constant to the values of all states

1. By (2) know,
2. Since here is a kind of continuing task not an episodic task, only the intervals between rewards are important, because when the intervals between rewards are unchanged, the signs of rewards are changed that means a constant c is added to all the rewards. According to (2), adding a constant c to all the rewards adds a constant to the values of all states and does not affect the relative values of any states under any policies, thus only the intervals between rewards matter.
3. Here is an equal probability random strategy, leading to the high probability of out of bounds at the boundary. Thus, the expected penalty value of out of bounds is high.

So, we can modify the policy by decreasing the probability of the agent choosing out-of-bounds in order to improve the policy.

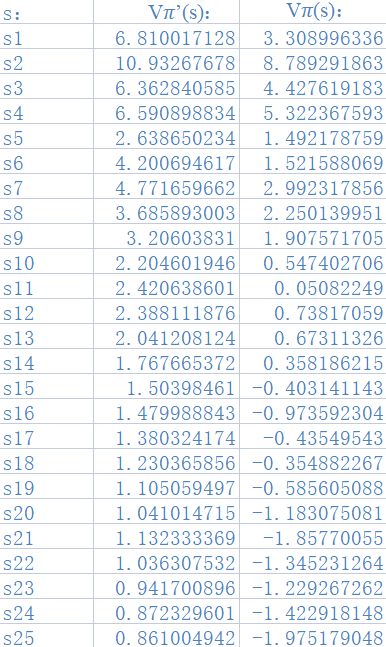
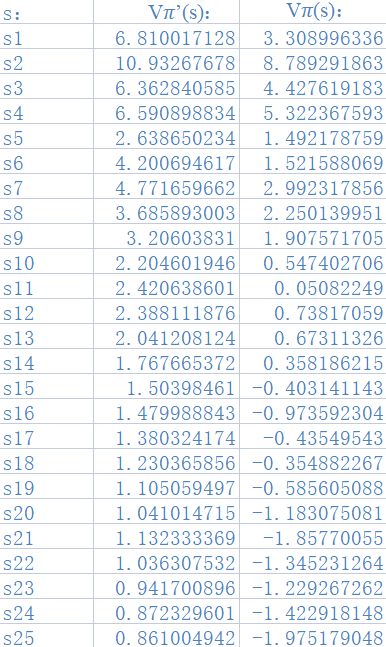
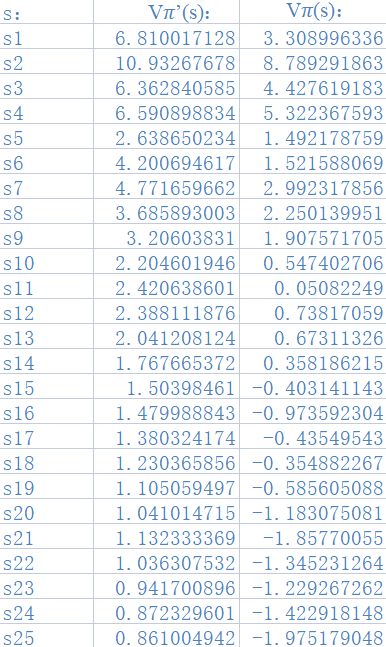
For example, same as the state marker in (1), the modified policy sets the probability of all out-of-bounds actions to 0, and the other actions have equal probability. i.e.,

Thus, according to policy , in matrix Ax=b, all elements in column matrix b is not negative. And all the principal minor sequences of A is positive, with all off diagonal elements are not positive → all elements in inverse matrix are positive, resulting in the calculated state value function is positive.

Obviously, for the states whose , → ,

and for the states whose, because the probability of all out-of-bounds actions sets to 0, there is no negative rewards, → .

The following calculated result is proved：

i.e., the modified policy is better than the policy .