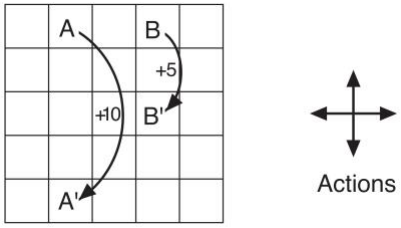
Homework #1

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., v_π(**A**) and v_π(**B**).

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

(3) What is v_c in terms of c and γ?

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

(5) Could you provide a new policy which is better than π?

Solution: (1) Each grid is marked as a state $s_k(k = 1 \dots 25)$,
the grid that lies in the i-th row and the j-th($i, j \in [1,5]$) corresponds state $s_{(i-1)*5+j}$,
and $v_\pi(A) = v_\pi(s_2), v_\pi(B) = v_\pi(s_4)$,
Then, according to

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')] \ (\gamma = 0.9)$$

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

$$\begin{cases} v_\pi(s_1) = 0.5 \cdot (-1 + \gamma v_\pi(s_1)) + 0.25 \cdot (0 + \gamma v_\pi(s_2)) + 0.25 \cdot (0 + \gamma v_\pi(s_6)) \\ v_\pi(s_2) = 4 \cdot 0.25 \cdot (10 + \gamma v_\pi(s_{22})) = 10 + 0.9 \cdot v_\pi(s_{22}) \\ v_\pi(s_3) = 0.25 \cdot (-1 + \gamma v_\pi(s_3)) + 0.25 \cdot (0 + \gamma v_\pi(s_2)) + 0.25 \cdot (0 + \gamma v_\pi(s_4)) + 0.25 \cdot (0 + \gamma v_\pi(s_8)) \\ v_\pi(s_4) = 4 \cdot 0.25 \cdot (5 + \gamma v_\pi(s_{14})) = 5 + 0.9 \cdot v_\pi(s_{14}) \\ \dots \\ v_\pi(s_{25}) = 0.5 \cdot (-1 + \gamma v_\pi(s_{25})) + 0.25 \cdot (0 + \gamma v_\pi(s_{20})) + 0.25 \cdot (0 + \gamma v_\pi(s_{24})) \end{cases}$$

$$\Rightarrow \begin{cases} 0.55 v_\pi(s_1) - 0.225 v_\pi(s_2) - 0.225 v_\pi(s_6) = -0.5 \\ v_\pi(s_2) - 0.9 \cdot v_\pi(s_{22}) = 10 \\ -0.225 v_\pi(s_2) + 0.775 v_\pi(s_3) - 0.225 v_\pi(s_4) - 0.225 v_\pi(s_8) = -0.25 \\ v_\pi(s_4) - 0.9 \cdot v_\pi(s_{14}) = 5 \\ \dots \\ -0.225 v_\pi(s_{20}) - 0.225 v_\pi(s_{24}) + 0.55 v_\pi(s_{25}) = -0.5 \end{cases}$$

Let $v_\pi(s_k) = x_k$, in a matrix form as Ax=b， A is 25 order square matrices， x=$(x_1 x_2, \ldots, x_{25})^T$

A=

$$\begin{bmatrix} 0.55 & -0.225 & 0 & 0 & 0 & -0.225 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.9 & 0 & 0 & 0 & 0 \\ 0 & -0.225 & 0.775 & -0.225 & 0 & 0 & 0 & -0.225 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & & & & & & & & & & & & & & & & & \end{bmatrix}$$

.....

b = $(-0.5, 10, -0.25, \ldots, -0.5)^T$,

⇒ x = $A^{-1}b$

x:	
x1	3.308996336
x2	8.789291863
x3	4.427619183
x4	5.322367593
x5	1.492178759
x6	1.521588069
x7	2.992317856
⇒ x8	2.250139951
x9	1.907571705
x10	0.547402706
x11	0.05082249
x12	0.73817059
x13	0.67311326
x14	0.358186215
x15	-0.403141143
x16	-0.973592304
x17	-0.43549543
x18	-0.354882267
x19	-0.585605088
x20	-1.183075081
x21	-1.85770055
x22	-1.345231264
x23	-1.229267262
x24	-1.422918148
x25	-1.975179048

⇒ x = $(3.3,8.8,4.4,5.3,\ldots,-2.0)^T$ (x *keeps one decimal place*)

∴ $v_\pi(A) = v_\pi(s_2) = x_2 = 8.8$, $v_\pi(B) = v_\pi(s_4) = x_4 = 5.3$

(2) $G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ (γ is known)

After adding a constant c，

$G_t' \doteq (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \ldots = \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c)$ (γ, c are known)

∵ $\gamma \in [0,1]$, ∴ $\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$, $G_t' \doteq \sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) = G_t + \frac{1}{1-\gamma} \cdot c$

$v_\pi'(s) = E_\pi[G_t' | S_t = s] = E_\pi[G_t + \frac{1}{1-\gamma} \cdot c | S_t = s] = E_\pi[G_t | S_t = s] + \frac{1}{1-\gamma} \cdot c = v_\pi(s) + \frac{1}{1-\gamma} \cdot c$

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

(3) By (2) know, $v_c = \frac{c}{1-\gamma} = 10c$ ($\gamma = 0.9$)

(4) 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 v_c 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.

(5) Here π 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 π 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 π' sets the probability of all out-of-bounds actions to 0, and the other actions have equal probability. i.e.,

$\pi'(upper|s_1) = \pi'(left|s_1) = 0, \pi'(right|s_1) = \pi'(down|s_1) = 0.5$

$\pi'(upper|s_2) = 0, \pi'(right|s_2) = \pi'(left|s_2) = \pi'(down|s_2) = \frac{1}{3}, \pi'(a|s_3) = \pi'(a|s_4) = \pi'(a|s_2)$

$\pi'(upper|s_5) = \pi'(right|s_5) = 0, \pi'(left|s_5) = \pi'(down|s_5) = 0.5$

$\pi'(left|s_6) = 0, \pi'(right|s_6) = \pi'(upper|s_6) = \pi'(down|s_6) = \frac{1}{3}, \pi'(a|s_{11}) = \pi'(a|s_{16}) = \pi'(a|s_6)$

$\pi'(upper|s_7) = \pi'(right|s_7) = \pi'(left|s_7) = \pi'(down|s_7) = 0.25,$

$\pi'(a|s_8) = \pi'(a|s_9) = \pi'(a|s_{12}) = \pi'(a|s_{13}) = \pi'(a|s_{14}) = \pi'(a|s_{17}) = \pi'(a|s_{18}) = \pi'(a|s_{19}) = \pi'(a|s_7)$

$\pi'(right|s_{10}) = 0, \pi'(upper|s_{10}) = \pi'(left|s_{10}) = \pi'(down|s_{10}) = \frac{1}{3}, \pi'(a|s_{15}) = \pi'(a|s_{20}) = \pi'(a|s_{10})$

$\pi'(down|s_{21}) = \pi'(left|s_{21}) = 0, \pi'(upper|s_{21}) = \pi'(right|s_{21}) = 0.5$

$\pi'(down|s_{22}) = 0, \pi'(upper|s_{22}) = \pi'(left|s_{22}) = \pi'(right|s_{22}) = \frac{1}{3}, \pi'(a|s_{23}) = \pi'(a|s_{24}) = \pi'(a|s_{22})$

$\pi'(down|s_{25}) = \pi'(right|s_{25}) = 0, \pi'(upper|s_{25}) = \pi'(left|s_{25}) = 0.5$

Thus, according to policy π', 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 A^{-1} are positive, resulting in the calculated state value function is positive.

Obviously, for the states whose $v_\pi(s) > 0$, $\rightarrow v_{\pi'}(s) > 0 > v_\pi(s)$,

and for the states whose $v_\pi(s) > 0$, because the probability of all out-of-bounds actions sets to 0, there is no negative rewards, $\rightarrow v_{\pi'}(s) > v_\pi(s)$.

The following calculated result is proved：

s:	Vπ'(s):	Vπ(s):
s1	6.810017128	3.308996336
s2	10.93267678	8.789291863
s3	6.362840585	4.427619183
s4	6.590898834	5.322367593
s5	2.638650234	1.492178759
s6	4.200694617	1.521588069
s7	4.771659662	2.992317856
s8	3.685893003	2.250139951
s9	3.20603831	1.907571705
s10	2.204601946	0.547402706
s11	2.420638601	0.05082249
s12	2.388111876	0.73817059
s13	2.041208124	0.67311326
s14	1.767665372	0.358186215
s15	1.50398461	-0.403141143
s16	1.479988843	-0.973592304
s17	1.380324174	-0.43549543
s18	1.230365856	-0.354882267
s19	1.105059497	-0.585605088
s20	1.041014715	-1.183075081
s21	1.132333369	-1.85770055
s22	1.036307532	-1.345231264
s23	0.941700896	-1.229267262
s24	0.872329601	-1.422918148
s25	0.861004942	-1.975179048

$v_{\pi'}(s) \geq v_\pi(s)$， $\forall s \in S \rightarrow \pi' > \pi$ i.e., the modified policy π' is better than the policy π.