RL Homework 2

20241202239

September 2024

Language: Python

Problem setup **Environment:** 4*4 grid world. **Reword:** is $r_{boundary} = r_{forbidden} = -1$, and $r_{target} = 1$. **Discount rate:** is $\gamma = 0.9$

Two policies Two policies is used,the determined policy is an array [1,1,1,0,1,2,4,0,2,4,4,3,2,3,4,2], where index is related to state, value is related to action. We have

$$r_\pi = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{1}$$

and

$$P_\pi = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

The stochastic policy is show as follow table, rows related to state, columns related to action,values is probability to take this action. The probability of stay is set to zero, and each boundary action is also set to zero to fasten the policy. We have

$$r_\pi = \begin{pmatrix} 0 \\ 0 \\ -0.33 \\ 0 \\ 0 \\ -0.5 \\ -0.33 \\ -0.33 \\ -0.75 \\ -0.33 \\ 0 \\ -0.67 \\ -0.33 \end{pmatrix} \tag{3}$$

and

$$P_\pi = \begin{bmatrix} 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0.33 & 0.33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0 & 0 & 0 & 0 & 0.33 & 0 & 0.33 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.25 & 0 & 0 & 0.25 & 0.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 & 0.33 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.75 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0 & 0.33 & 0 & 0 & 0 & 0.33 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.33 & 0.67 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.5 \end{bmatrix} \tag{4}$$

Here plot the tow policies.

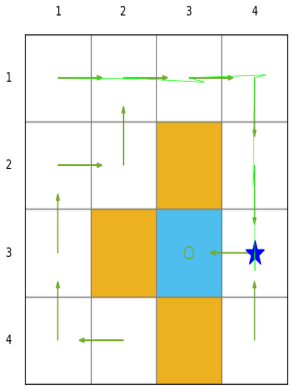

图 1: Policy 1

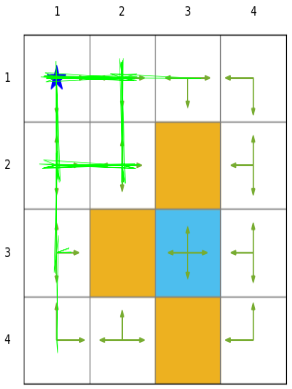

图 2: Policy 2

closed-form algorithm $v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$ is directly transformed from Bellman equation $v_\pi = r_\pi + \gamma P_\pi v_\pi$, it's useful for theoretical analysis, but not applicable for computer calculate because it involves matrix inversion operation. Here plots state value calculated from closed-form algorithm.

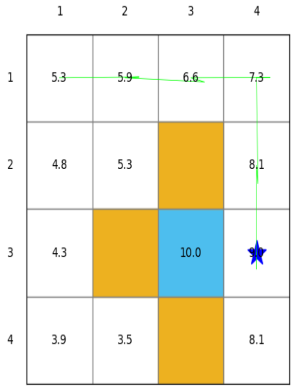

图 3: Policy 1

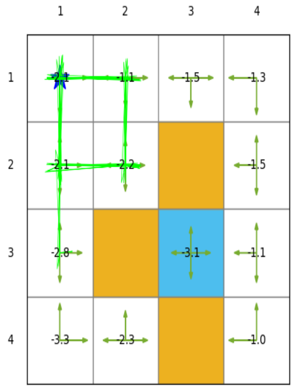

图 4: Policy 2

iterative algorithm $v_{k+1} = r_\pi + \gamma P_\pi v_k$ is using fix point iterative solution to find v_π, it is easier to calculate. Here plots state value calculated from iterative algorithm, v_0 uses all one array and 100 iteration is applied.

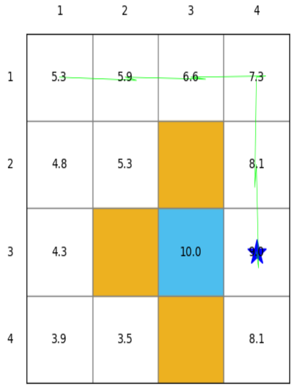

图 5: Policy 1

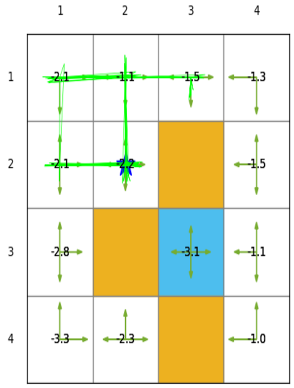

图 6: Policy 2

Observation For deterministic policy 1, state value decrease with the distance goes far. For policy 2, most state values decrease with the distance goes far, but the states near the forbidden state have lower state value. Also compare policy 1 to policy 2, policy 1 generally have higher state value.