作业 6

3. 由网格图的对称性，可令

$$v_\pi(1)=v_\pi(3)=v_\pi(5)=v_\pi(7)=a,$$

$$v_\pi(2)=v_\pi(6)=b,\quad v_\pi(4)=c.$$

则

$$v_\pi(1)=-1+\frac{1}{4}(a+b+c)=a,$$

$$v_\pi(4)=-1+\frac{1}{4}(4a)=c,$$

$$v_\pi(6)=-1+\frac{1}{4}(2b+2a)=b.$$

解得 $a=-7$, $b=-9$, $c=-8$.

因此

$$q_\pi(4,\text{left})=-1+v_\pi(3)=-8,$$

$$q_\pi(7,\text{right})=-1+v_\pi(8)=-1+0=-1.$$

4. (1) 令 $v_\pi=[v_\pi(A)\ v_\pi(B)\ v_\pi(C)\ v_\pi(D)]^T$,

$$r_\pi=[-1\ \ -1\ \ -1\ \ 0]^T,$$

$$P_\pi=\begin{bmatrix}0 & 0.5 & 0.5 & 0\\ 0 & 0.5 & 0 & 0.5\\ 0.5 & 0 & 0 & 0.5\\ 0 & 0 & 0 & 1\end{bmatrix},$$

解方程组

$$V_\pi = r_\pi + \gamma P_\pi V_\pi$$

得

$$V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

$$= \begin{bmatrix} 1.0666 & 0.3555 & 0.2666 & 0.3111 \\ 0 & 1.3333 & 0 & 0.6666 \\ 0.2666 & 0.0888 & 1.0666 & 0.5777 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} -1.6889 \\ -1.3333 \\ -1.4222 \\ 0 \end{bmatrix}.$$

(2) 若问题规模较大，但模型已知，则可以采用动态规划方法求解，即迭代计算

$$V_{k+1} = r_\pi + \gamma P_\pi V_k.$$

直至收敛.

如果模型未知，则需要使用蒙特卡洛预测算法，根据给定的策略，产生出一个观测片段，然后根据定义计算 t 时刻开始的累积回报，重复 n 次后取均值

$$v(s_t) = \frac{1}{n} \sum_{i=1}^{n} g_i.$$

其中，首次访问蒙特卡洛时，要检查是否为首次

当现，后面每次访问时，则不再检查。