**题目：11-State ChainWalk MDP**



环境特点：

1. 智能体在每个状态下执行“向左”或“向右”动作，按照动作指令转移一个状态的概率为0.7，按动作指令相反方向转移一个状态的概率为0.3；左（右）端点状态向左（右）转移状态时保持端点位置不变；
2. 奖励跟所处状态有关，中间状态s5奖励为0，右半部分状态s6-s10奖励均为1；左半部分状态，除左端点s0状态的奖励为3，其余状态s1-s4奖励为-1；

要求：计算并画出利用bellman最优算子以及优势学习算子情况下迭代策略的性能界以及动作间隔（action gap）的变化？



提示：

1. 算子形式：

Bellman最优算子：

优势学习算子：



取值和

1. 初始Q值随机生成，如 10 \* np.random.random();
2. 策略的V值函数，而是根据第k次迭代Q值诱导的贪婪策略；
3. 真实最优策略为“在任何状态下都执行‘向左’的动作”，那么Action gap定义为