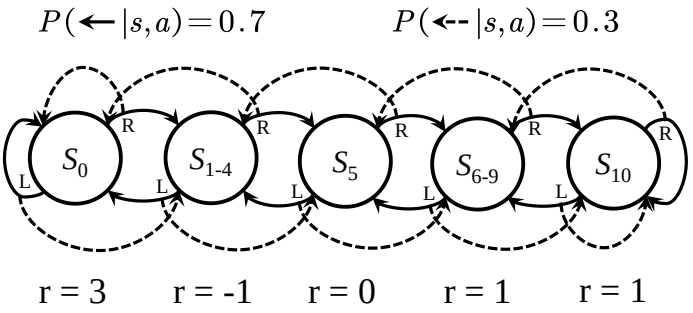
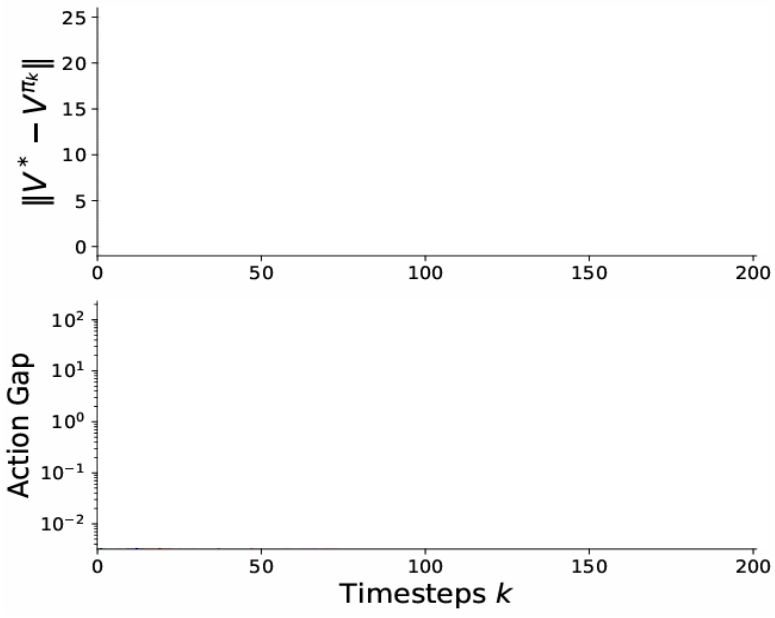
题目：11-State ChainWalk MDP



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$||V^*-V^{\pi_k}||_\infty$以及动作间隔（action gap）的变化？



提示：

1. 算子形式：

Bellman 最优算子：$\mathcal{T}^{*}Q(s,a)=r(s,a)+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'}Q(s',a')]$

优势学习算子：

$$\mathcal{T}_{\mathrm{AL}}Q(s,a)=r(s,a)+\alpha(Q(s,a)-\max_{\tilde{a}}Q(s,\tilde{a}))+\gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}[\max_{a'}Q(s',a')]$$

取值$\alpha=0.99$和$\gamma=0.99$

2. 初始 Q 值随机生成，如 10 * np.random.random();

3. V^{π_k} 策略π_k的 V 值函数，而π_k是根据第 k 次迭代 Q 值诱导的贪婪策略 $\pi_k(s)=\mathrm{argmax}_a Q_k(s,a)$;

4. 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 $mean_{s\in\mathcal{S}}(Q(s,\text{‘}L\text{’})-Q(s,\text{‘}R\text{’}))$