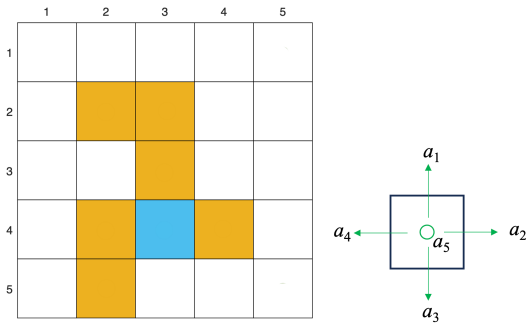
强化学习第一次作业

提交作业时间：2024.11.21 22:00 前

作业描述：使用程序语言（建议 C++）实现以下算法。共两道大题

1. 使用程序语言实现三个算法伪码：1）值迭代算法；2）策略迭代算法；3）截断式策略迭代算法。验证示例使用以下设置：

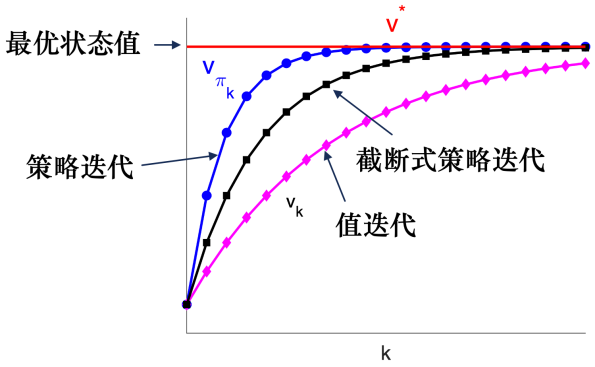
网格世界：5 × 5；每个格子有 5 个动作，a_1-a_5



奖励：$r_{\text{boundary}} = -1$，$r_{\text{forbidden}} = -10$，$r_{\text{target}} = 1$，$r_{\text{otherstep}} = 1$，折扣率 $\gamma = 0.9$。初始策略是从所有状态出发所有动作都采取 a_5

定义$\|v_k - v_{k\text{-}1}\|$为 k 时刻的状态值迭代误差，停止标准是 $\|v_k - v_{k\text{-}1}\| < 0.001$。

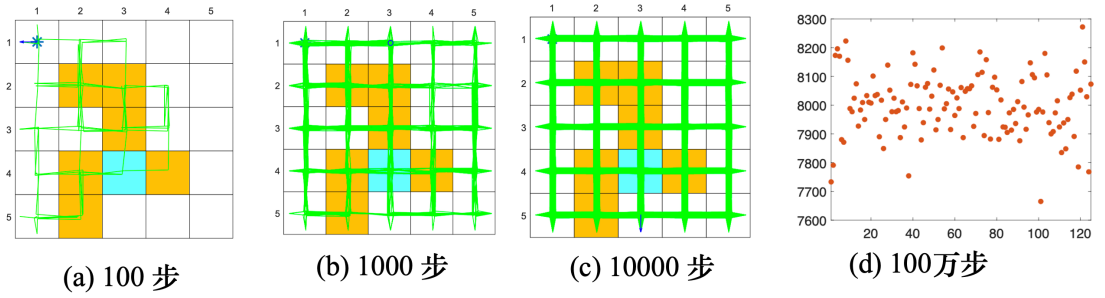
（1）比较三种算法的收敛速度（收敛迭代次数，参考下图；选择某个状态的状态值打印出类似的示意图）

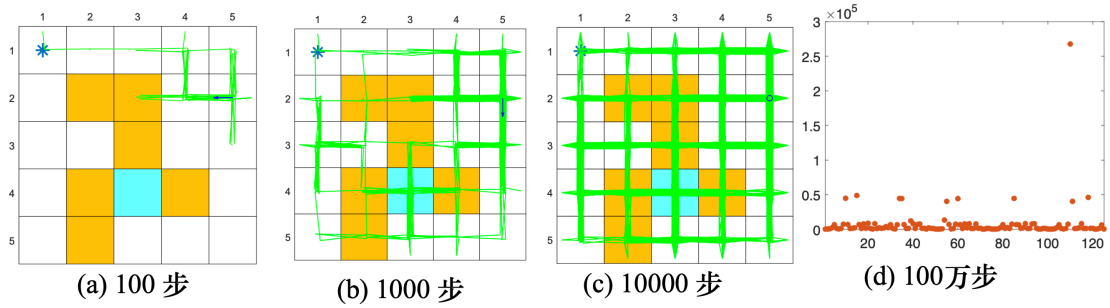


（2）对于截断式策略迭代-x 算法，给出 x=1，5，9，56，描述观测到的实验结果，并绘制结果图，其中横轴表示迭代次数，纵轴表示状态值误差（建议先用贝尔曼最优方程的迭代算法，算出 v^*）。

2. MC ε-贪心策略算法

（1）分析 ε=1、ε=0.5 时，单个 episode 可以访问的“状态-动作对”情况：episode 的长度分别为 100 步、1000 步、10000 步，100 万步的情况，并用以下类似的方式可视化展示（系统环境也参照下图）。





（2）分析 ε-贪心策略算法的最优性与探索性。网格世界：5 × 5；每个格子有 5 个动作，a_1-a_5，设置 $r_{\text{boundary}} = -1, r_{\text{forbidden}} = -10, r_{\text{target}} = 1$，$r_{\text{otherstep}} = 0$，$\gamma = 0.9$。

根据所讲授的 MC ε-贪心策略算法伪码，a. 使用程序语言实现该算法的伪码，b. 分析 ε=0、ε=0.1、ε=0.2、ε=0.5 时，最优的 ε-贪心策略及其状态值，并在网格世界图分别表示出对应的最优的 ε-贪心策略及其状态值，打印输出。