强化学习作业

（作业提交时间：2024 年 1 月 14 日 23：55，课堂派，加课码 LZ4KAQ）

第一道题：

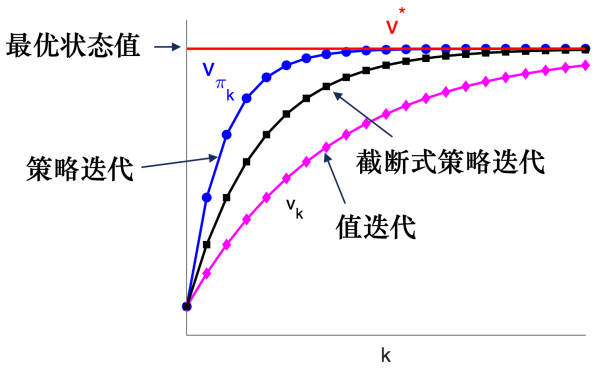
使用程序语言（自选）实现三个算法伪码：1）值迭代算法；2）策略迭代算法；3）截断式策略迭代算法。验证示例使用以下设置：

网格世界：5 × 5；每个格子有 5 个动作，a_1-a_5

奖励：$r_{\text{boundary}}=-1$，$r_{\text{forbidden}}=-10$，$r_{\text{target}}=1$，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$\|v_k-v^*\|$为 k 时刻的状态值误差，停止标准是 $\|v_k-v^*\|<0.01$。

（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=1，4，7，50，描述观测到的实验结果。

第二道题：（视讲课内容和进度待定）