强化学习第三次作业

提交作业时间：2024.12.5 22:00 前

作业描述：使用程序语言（建议 C++）实现以下算法。2 道大题

各题的提交文档

1）Word内容，包括：a. 设计思想；b.伪码描述（强化学习核心算法）；
c. 时间复杂度分析；d. 结果（截图）分析。

2）源码 zip 压缩包。

1. 设置：$X \in \mathbb{R}^2$ 表示平面中的随机位置。 其分布在以原点为中心、边长为 30 的正方形区域内是均匀分布。

   a) 根据期望公式求解均值$\mathbb{E}[X]$。

   b) 在上述正方形区域内随机采集样本：400 个独立同分布样本 $\{x_i\}_{i=1}^{400}$ 。

   c) 使用随机梯度下降方法估计$\mathbb{E}[X]$，初值 w_1=(50, 50)，$\alpha_k = 1/k$。输出可视化结果：400 个样本点的平面分布，随机梯度下降方法的轨迹(参照图 1 样例)；迭代次数与估计误差（均值估计值与期望真值差的绝对值）(参照图 2 样例)。适当调整 α_k （比如取一个常数 0.005； 取 c_k/k（其中 c_k 有界）等），并在对应的同张图（图 1、图 2）中输出结果。

   d) 使用小批量梯度下降(MBGD)方法估计$\mathbb{E}[X]$, 初值 w_1=(50, 50), $\alpha_k = 1/k$, m 分别取 1, 10， 50, 100。输出可视化结果: 400 个样本点的平面分布，随机梯度下降方法的轨迹(参考图 3 样例)；迭代次数与估计误差（均值

估计值与期望真值差的绝对值）(参考图 3 样例)。

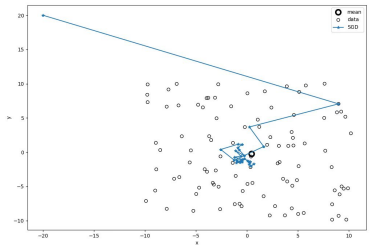


图 1. 求解收敛轨迹

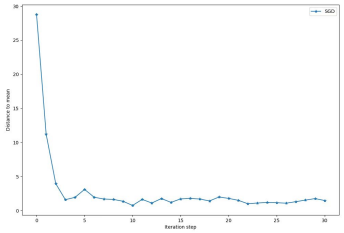


图 2. 误差

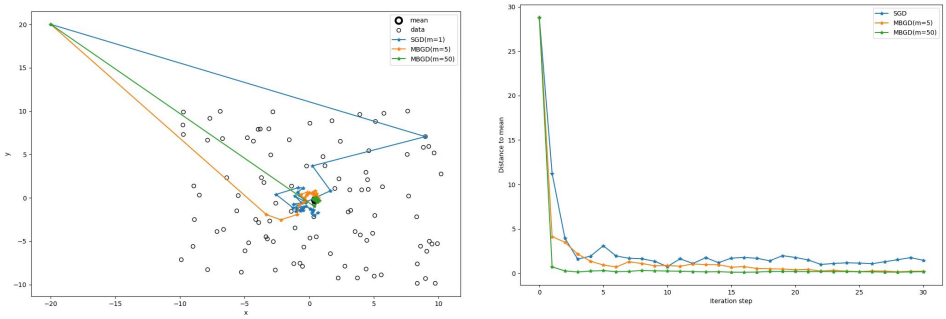


图 3：不同梯度下降算法均值估计

2. 任务描述：

a) 使用 Q-learning 的时序差分学习方法进行策略搜索（off-policy 版本），为 5×5 的网格世界所有状态找到最优策略。

b) 奖励设置为 $r_{\text{boundary}} = r_{\text{forbidden}} = -1$，$r_{\text{target}} = 1$，$r_{\text{otherstep}} = 0$。折扣率为 $\gamma = 0.9$。 学习率为 α=0.1。

c) 行为策略, ε=1(图 4a)、ε=0.5(图 4b)、ε=0.1(图 4c, 4d); 各生成一个 episode,

长度为 10^5 步。①绘制对应的 episode 的轨迹图（参考图 5 所示）；②计算状态值误差（|正确值-估计值|），绘制状态值误差图结果输出，其中横轴为 episode 的步数，纵轴为状态值误差值。给定正确答案（Ground truth）：最优策略和相应的最优状态值（如图 6 所示）。

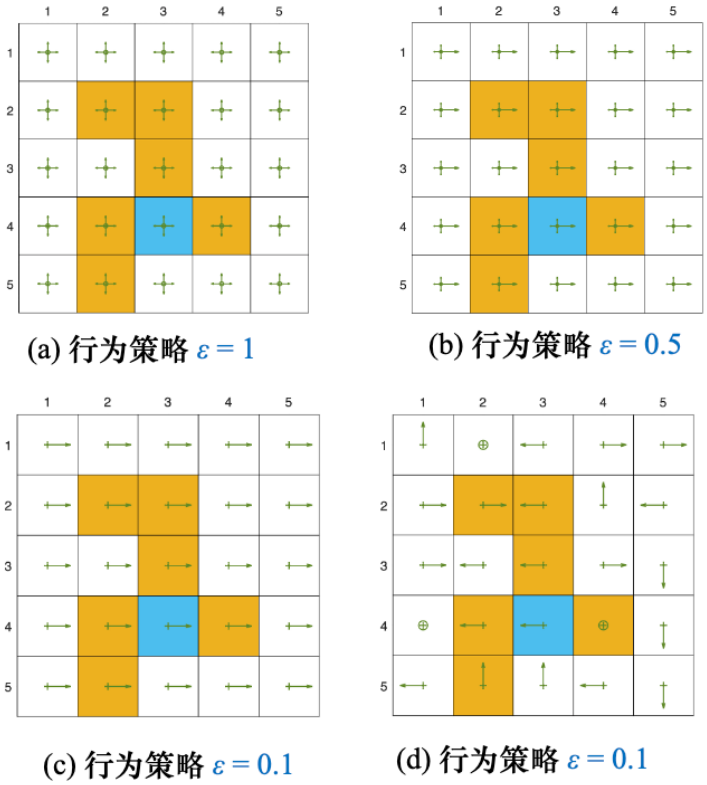


图 4：行为策略

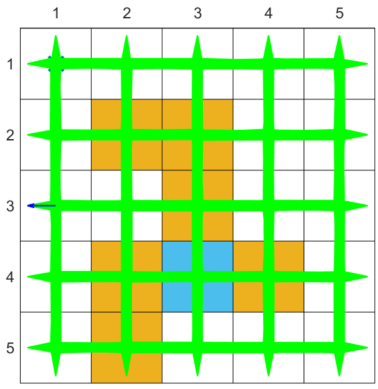
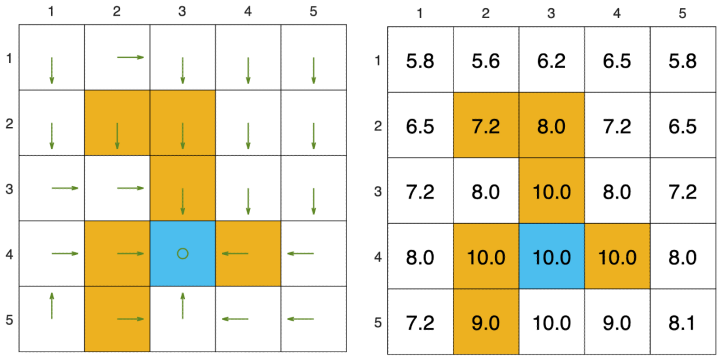


图 5：episode 的轨迹图示例



(a) 最优策略　　(b) 最优状态值

图 6：正确答案（Ground truth）