强化学习第四次作业

提交作业时间：2024.12.19 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. 考虑一个 5×5 网格世界（如图 1 所示）。 给定一个策略 π：对于任何 s, a, $\pi(a|s)=0.2$（如图 1 的箭头所示），目标是估计该策略的状态值（即策略评估问题）。设置 $r_{\text{forbidden}}=r_{\text{boundary}}=-1$、$r_{\text{target}}=1$ 、$r_{\text{otherstep}}=0$ 和 $\gamma=0.9$。使用贝尔曼方程的迭代算法求解得到的结果作为正确答案（Ground truth），与后面近似结果进行比较计算状态值误差。按照给定策略生成了 500 个 episodes；每个 episode 有 500 步，从服从均匀分布的“状态-动作对”中，随机选择一个“状态-动作对”开始，每个状态使用二维平面的点表示，动作可以是一维。

   a) 函数表示模型选择线性函数，使用 TD-Linear 算法，近似给定策略 π 的真实状态值。特征向量 $\phi(s)$可以是多项式或傅立叶函数。

   b) 如果是多项式，给出 $\phi(s)=[1,x,y]^{\mathrm{T}}\in\mathbb{R}^3$、$\phi(s)=[1,x,y,x^2,y^2,xy]^{\mathrm{T}}\in\mathbb{R}^6$、$\phi(s)=[1,x,y,x^2,y^2,xy,x^3,y^3,x^2y,xy^2]^{\mathrm{T}}\in\mathbb{R}^{10}$ 的结果，包括近似的状态值，以及状态值误差；如果是傅立叶函数，给出 q=1 且 $\phi(s)\in\mathbb{R}^4$、q=2 且 $\phi(s)$

$\in \mathbb{R}^9$、q=3 且 $\phi(s) \in \mathbb{R}^{16}$ 的比较结果，包括近似的状态值，以及状态值误差。

c) 以 5×5 表格的方式输出所有状态的状态值，并以图形方式输出 3D 状态值表示（如图 2 所示）。

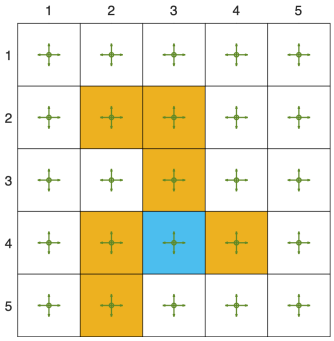
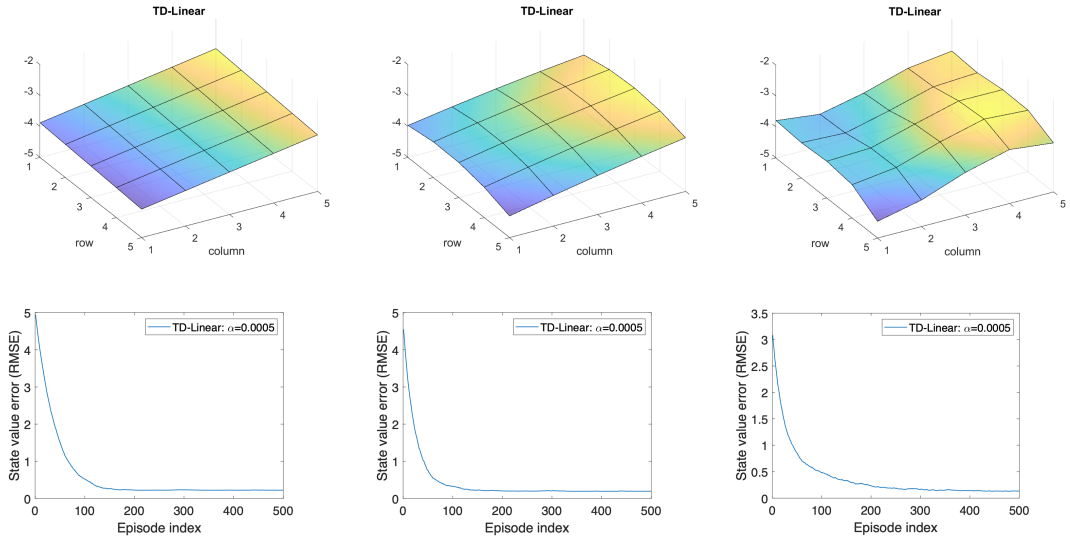


图 1. 给定策略



(a) $\phi(s) \in \mathbb{R}^3$ (b) $\phi(s) \in \mathbb{R}^6$ (c) $\phi(s) \in \mathbb{R}^{10}$

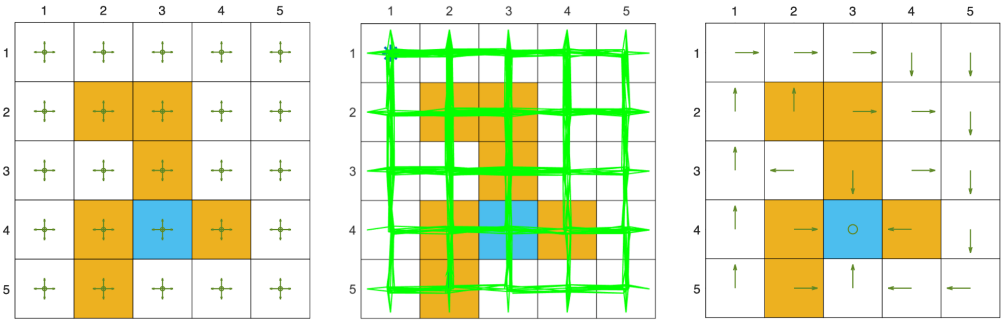
图 2.使用多项式特征得到的 TD-线性（TD-Linear）估计结果

2. 任务描述：考虑一个 5×5 网格世界，设置 $r_{\text{forbidden}} = r_{\text{boundary}} = -10$、$r_{\text{target}} = 1$ 、$r_{\text{otherstep}} = 0$ 和 $\gamma = 0.9$（如图 1 所示）。使用“深度 Q-learning”方法（off-policy 版本），学习每个“状态-动作对”的最优动作值，当获得最优动作值，计算获得

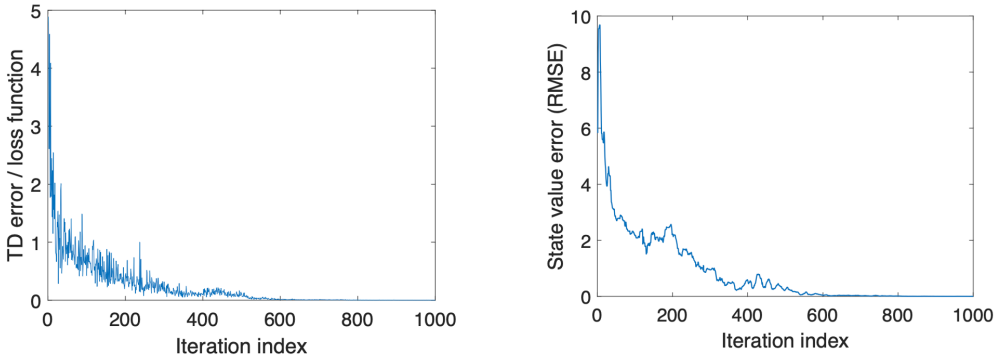
最优贪心策略。使用一个单独的 episode 来训练网络。该 episode 是由图 3(a)所示的探索性行为策略生成。此 episode 只有 1,000 步（即回放缓冲区中有 1000 样本)。小批量规模为 100，即每次获取样本时，从重放缓冲区中均匀抽取 100 个样本。神经网络的结构设计可参考如下：一个隐藏层的浅层神经网络被用作$\hat{q}(s, a, w)$ 的非线性函数表示模型，隐藏层有 100 个神经元。神经网络有 3 个输入和 1 个输出，其中前 2 个输入是状态的归一化行索引和列索引，第 3 个输入是归一化的动作索引，网络的输出是估计的动作值。

a) 在 5×5 网格世界中，输出有 1000 步 episode 的轨迹（如图 3(b)所示）。在 5×5 网格世界中，输出获得的最终策略（如图 3（c）所示）。输出 **TD 误差/损失函数**（如图 3（d））、状态估计误差(如图 3（d））。

b) 如果只使用 100 个步的单个 episode，输出类似 a）的相应结果。



(*a*) 行为策略　(*b*) 1个episode 有1000 步　(*c*) 获得的最终策略



(*d*) TD 误差/损失函数　(*e*) 状态估计误差

图 3：使用深度 Q-learning 进行最优策略学习