强化学习第一次实验作业

教师: 赵冬斌　朱圆恒　张启超

实验作业任务

从下面给出的问题中选取至少一个作为实验对象, 使用强化学习课程中学到的强化学习算法, 完成问题的控制目标.

算法代码语言不限, 但建议使用 MATLAB 和 Python.

将实验过程整理成完整的报告, 内容包括但不限于方法描述, 研究思路, 研究内容, 实验结果, 分析讨论, 方法改进等.

在规定的时间内提交报告和源代码, 完成实验作业.

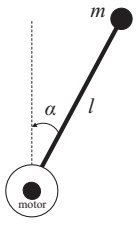
实验作业成绩占总成绩比重: 20%

注意: 严禁抄袭代码和报告!

问题 1: 倒立摆

倒立摆是将一个物体固定在一个圆盘的非中心点位置, 由直流电机驱动将其在垂直平面内进行旋转控制的系统 (图 1). 由于输入电压是受限的, 因此电机并不能提供足够的动力直接将摆杆推完一圈. 相反, 需要来回摆动收集足够的能量, 然后才能将摆杆推起并稳定在最高点.

(a) 真实系统



(b) 示意图

Figure 1: 倒立摆问题.

Table 1: 倒立摆系统参数

变量	取值	单位	含义
m	0.055	kg	重量
g	9.81	$\mathrm{m/s^2}$	重力加速度
l	0.042	m	重心到转子的距离
J	$1.91 \cdot 10^{-4}$	$\mathrm{kg \cdot m^2}$	转动惯量
b	$3 \cdot 10^{-6}$	$\mathrm{Nm \cdot s/rad}$	粘滞阻尼
K	0.0536	Nm/A	转矩常数
R	9.5	Ω	转子电阻

倒立摆系统连续时间动力学模型是

$$\ddot{\alpha} = \frac{1}{J}\left(mgl\sin(\alpha) - b\dot{\alpha} - \frac{K^2}{R}\dot{\alpha} + \frac{K}{R}u\right) \tag{1}$$

表 1 给出了所有参数的含义和取值. 系统状态包含摆杆的角度和角速度, 即 $s = [\alpha, \dot{\alpha}]^T$. 角度 α 取值范围在 $[-\pi, \pi)$ rad 之间. 其中 $\alpha = -\pi$ 对应摆杆垂直指向下, $\alpha = 0$ 对应摆杆垂直指向上. 速度 $\dot{\alpha}$ 被限制在 $[-15\pi, 15\pi]$rad/s 范围内. 控制动作 (电压)u 被限制在 $[-3, 3]$V 范围内. 采样时间 T_s 选取 0.005s, 离散时间动力学 f 可以根据 (1) 由欧拉法获得

$$\begin{cases} \alpha_{k+1} = \alpha_k + T_s\dot{\alpha}_k & (2) \\ \dot{\alpha}_{k+1} = \dot{\alpha}_k + T_s\ddot{\alpha}(\alpha_k, \dot{\alpha}_k, a_k) & (3) \end{cases}$$

控制目标是将摆杆从最低点 $s = [\pi, 0]^T$ 摆起并稳定在最高点 $s = [0, 0]^T$. 奖励函数定义成如下二次型形式

$$\begin{aligned} \mathcal{R}(s, a) &= -s^T Q_{rew} s - R_{rew} a^2 \\ Q_{rew} &= \begin{bmatrix} 5 & 0 \\ 0 & 0.1 \end{bmatrix}, R_{rew} = 1. \end{aligned} \tag{4}$$

折扣因子选取 $\gamma = 0.98$. 选取较高折扣因子的目的是为了提高目标点 (顶点) 附近奖励在初始时刻状态价值的重要性, 这样最优策略能够以成功将摆杆摆起并稳定作为最终目标.

(**Tip:** 可以将动作空间离散化成 $\{-3, 0, 3\}$ 三个动作, 以这三个动作作为动作集学习最优策略.)

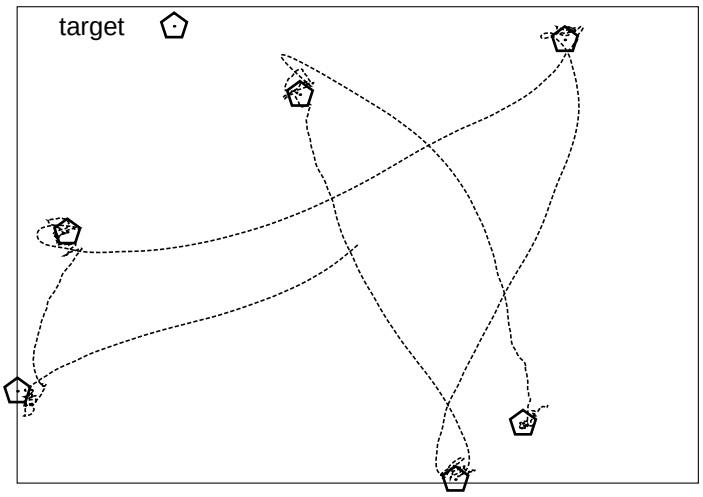


Figure 2: 冰壶游戏中冰壶在场地内控制轨迹示意图

问题 2：冰壶游戏

冰壶游戏是要控制一个半径为 1, 质量为 1 的冰壶, 在一个长宽是 100×100 的正方形球场内移动. 不考虑冰壶的自转. 当冰壶和球场的边界碰撞时, 碰撞前后冰壶的速度会乘上回弹系数 0.9, 移动方向和边界呈反射关系.

我们需要分别操纵 x 轴和 y 轴的两个力控制冰壶的移动: 在 x 轴的正或反方向施加 5 单位的力; 在 y 轴的正或反方向施加 5 单位的力. 这样一共会有 4 种不同的控制动作. 动作可以每 1/10 秒变换一次; 但在仿真冰壶运动动力学时, 仿真时间间隔是 1/100 秒. 除了我们施加的控制动作, 冰壶会受到空气阻力, 大小等于 $0.005 \times \text{speed}^2$. 假设冰壶和地面没有摩擦力.

在每个决策时刻 (1/10 秒), 环境反馈的奖励等于 $-d$, 其中 d 是冰壶和任意给定的目标点之间的距离. 为了保证学到的策略能够控制冰壶从任意初始位置上移动到任意目标点, 每隔 30 秒就把冰壶状态重置到球场内的一个随机点上, 同时 x 轴和 y 轴的速度也随机重置在 [-10, 10] 范围内. 与此同时, 目标点也被随机重置.

(**Tip**: 把每隔 30 秒当成一次轨迹，把问题定义成 episodic MDPs 问题, episodic length $= 30/0.1 = 300$ steps. $\gamma = 1$ 或 $\gamma = 0.9$)