强化学习实验作业

教师: 赵冬斌

助教: 朱圆恒

2019-5-9

1 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: 40%

注意: 严禁抄袭代码和报告!

2 问题 1: 小车爬山

一辆小车沿如图 1所示的山路行驶, 目标是能够开到右侧的山顶位置. 但是由于重力作用超过了汽车发动机的动力输出能力, 在最陡坡路段即使是油门踩死也

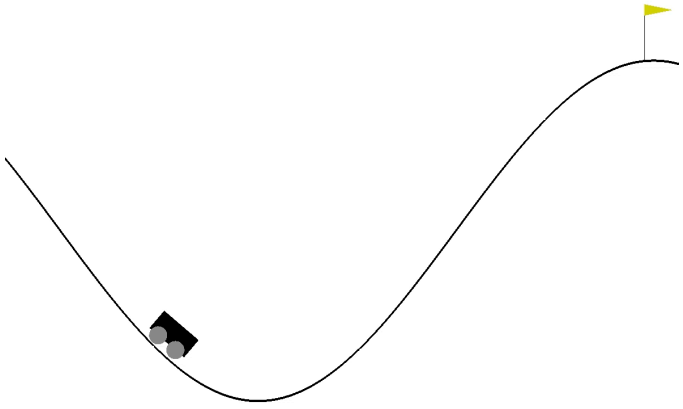
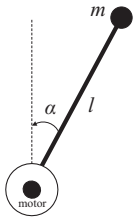


Figure 1: 小车爬山问题.

(a) 真实系统



(b) 示意图

Figure 2: 倒立摆问题.

无法提供足够的前进动力. 能够从山谷开到山顶唯一的方法就是先**向后**加速度, 远离目标点, 然后全速向前加速积累一定车速, 最终冲过最陡的路段到达顶点. 求解该控制问题要求算法能够学到为了达到目标, 需要先做出不利于目标的动作.

问题具有两个连续的状态变量, 小车的位置 p_t 和小车的速度 v_t. 两者的取值范围在

$$-1.2 \leq p \leq 0.5,\ -0.07 \leq v \leq 0.07 \tag{1}$$

山坡的几何形状满足

$$\text{高度} = \sin(3p). \tag{2}$$

每次实验小车的初始状态都是山谷最低点, 即 $p = -0.5$, $v = 0$. 动作 a_t 从离散的动作集 $\{+1, 0, -1\}$ 中选取, 分别代表向前加速, 空档滑行, 向后加速. 系统状态的变化由如下模型决定

$$\begin{cases} v_{t+1} = bound[v_t + 0.001a_t + g\cos(3p_t)] & (3) \\ p_{t+1} = bound[p_t + v_{t+1}] & (4) \end{cases}$$

其中 $g = -0.0025$ 对应重力因素, $bound$ 函数将输入变量截断在允许范围内. 如果 p_{t+1} 被截断了, v_{t+1} 也相应地被置为零. 奖励定义成每一步都为 -1. 当 $p_{t+1} > 0.5$ 时, 小车已经达到了目标点, 即达到了终止状态.

3 问题 2: 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 2). 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.

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

$$\ddot{\alpha} = \frac{1}{J}\left(mgl\sin(\alpha) - b\dot{\alpha} - \frac{K^2}{R}\dot{\alpha} + \frac{K}{R}u\right) \tag{5}$$

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 (5) 由欧拉法获得

$$\begin{cases} \alpha_{k+1} = \alpha_k + T_s\dot{\alpha}_k & (6) \\ \dot{\alpha}_{k+1} = \dot{\alpha}_k + T_s\ddot{\alpha}(\alpha_k, \dot{\alpha}_k, a_k) & (7) \end{cases}$$

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

$$\begin{aligned} \mathcal{R}(s, a) &= -s^T Q_{rew} s - R_{rew} a^2 \\ Q_{rew} &= \begin{bmatrix} 5 & 0 \\ 0 & 0.1 \end{bmatrix}, R_{rew} = 1. \end{aligned} \tag{8}$$

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.

(**TIPS:** 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.)

4 问题 3: 自选问题

不同的同学来自不同的专业和领域, 每个领域都有具有马尔科夫决策属性的问题需要解决.

欢迎同学根据自己的研究领域自选对象, 建立合适的 MDP, 使用强化学习方法求解问题的最优策略.