



Reinforcement Learning: An Introduction

强化学习导论第二版疑问

作者：吕昀璿

组织：UESTC

时间：March 24, 2020



Victory won't come to us unless we go to it. — M. Moore

目录

1	介绍	1
1.1	强化学习	1
1.2	例子	1
1.3	强化学习的要素	1
1.4	局限和范围	1
1.5	拓展例子：井字游戏	1
1.6	总结	1
1.7	强化学习的早期历史	1
	第一部分 表格解决方法	2
2	多臂赌博机	3
2.1	k 臂赌博机问题	3
2.2	动作值方法	3
2.3	10 臂试验	3
2.4	渐增实现	3
2.5	非平稳问题	3
2.6	乐观初始值	3
2.7	置信上限动作选择	3
2.8	梯度赌博机算法	3
2.9	关联搜索（上下文赌博机）	3
2.10	总结	3
3	有限马尔可夫决策过程	4
3.1	Agent-环境接口	4
3.2	目标和奖励	4
3.3	回报和 episode	4
3.4	回合和连续任务的统一符号	4
3.5	策略和值函数	4
3.6	最优策略和最优值函数	4
3.7	最优和近似	4
3.8	总结	4

4	动态规划	5
4.1	策略评估	5
4.2	策略改进	5
4.3	策略迭代	5
4.4	值迭代	5
4.5	异步动态规划	5
4.6	广义策略迭代	5
4.7	动态规划效率	5
4.8	总结	5
5	蒙特卡罗方法	6
5.1	蒙特卡洛预测	6
5.2	动作值的蒙特卡洛估计	6
5.3	蒙特卡洛控制	6
5.4	无探索起点的蒙特卡洛控制	6
5.5	重要性采样的 off-policy 预测	6
5.6	渐增实现	6
5.7	Off-policy 蒙特卡洛控制	6
5.8	* 折扣的重要性采样	6
5.9	* 每决策的重要性采样	6
5.10	总结	6
6	时间差分学习	7
6.1	TD 预测	7
6.2	TD 预测方法的优势	7
6.3	TD(0) 的最优性	7
6.4	Sarsa: on-policy TD 控制	7
6.5	Q-learning: off-policy TD 控制	7
6.6	期望 Sarsa	7
6.7	最大化偏差与 Double Learning	7
6.8	游戏、后期状态和其他特殊情况	7
6.9	总结	7
7	n 步自举	8
7.1	n 步 TD 预测	8
7.2	n 步 Sarsa	8
7.3	n 步 off-policy 学习	8
7.4	* 控制变量的每决策方法	8
7.5	无重要性采样的 off-policy 学习: n 步树备份算法	8

7.6 * 一种统一算法: $Q(\sigma)$	8
8 表格法进行规划和学习	9



第一章 介绍

1.1 强化学习

1.2 例子

1.3 强化学习的要素

1.4 局限和范围

1.5 拓展例子：井字游戏

1.6 总结

1.7 强化学习的早期历史

第一部分

表格解决方法

第二章 多臂赌博机

2.1 k 臂赌博机问题

2.2 动作值方法

2.3 10 臂试验

2.4 渐增实现

2.5 非平稳问题

$$(1 - \alpha)^n + \alpha \sum_{i=1}^n (1 - \alpha)^{n-i} \quad (2.1)$$

$$= (1 - \alpha)^n + \alpha \frac{1 - (1 - \alpha)^n}{1 - (1 - \alpha)} \quad (2.2)$$

$$= (1 - \alpha)^n + 1 - (1 - \alpha)^n \quad (2.3)$$

$$= 1 \quad (2.4)$$

2.6 乐观初始值

2.7 置信上限动作选择

2.8 梯度赌博机算法

2.9 关联搜索（上下文赌博机）

2.10 总结

第三章 有限马尔可夫决策过程

3.1 Agent-环境接口

3.2 目标和奖励

3.3 回报和 episode

3.4 回合和连续任务的统一符号

3.5 策略和值函数

3.6 最优策略和最优值函数

3.7 最优和近似

3.8 总结

第四章 动态规划



- 4.1 策略评估
- 4.2 策略改进
- 4.3 策略迭代
- 4.4 值迭代
- 4.5 异步动态规划
- 4.6 广义策略迭代
- 4.7 动态规划效率
- 4.8 总结

第五章 蒙特卡罗方法

5.1 蒙特卡洛预测

5.2 动作值的蒙特卡洛估计

5.3 蒙特卡洛控制

5.4 无探索起点的蒙特卡洛控制

5.5 重要性采样的 off-policy 预测

5.6 渐增实现

5.7 Off-policy 蒙特卡洛控制

5.8 * 折扣的重要性采样

5.9 * 每决策的重要性采样

5.10 总结

第 六 章 时间差分学习

6.1 TD 预测

6.2 TD 预测方法的优势

6.3 TD(0) 的最优性

6.4 Sarsa: on-policy TD 控制

6.5 Q-learning: off-policy TD 控制

6.6 期望 Sarsa

6.7 最大化偏差与 Double Learning

6.8 游戏、后期状态和其他特殊情况

6.9 总结

第七章 n 步自举

7.1 n 步 TD 预测

7.2 n 步 Sarsa

7.3 n 步 off-policy 学习

7.4 * 控制变量的每决策方法

7.5 无重要性采样的 off-policy 学习: n 步树备份算法

7.6 * 一种统一算法: $Q(\sigma)$

7.7 总结

第 八 章 表格法进行规划和学习

