

常用符号

符号	中文	英文
S 或 s	状态	state
A 或 a	动作	action
R 或 r	奖励	reward
U 或 u	回报	return
γ	折扣率	discount factor
\mathcal{S}	状态空间	state space
\mathcal{A}	动作空间	action space
$\pi(a s)$	随机策略函数	stochastic policy function
$\mu(s)$	确定策略函数	deterministic policy function
$p(s' s, a)$	状态转移函数	state-transition function
$Q_{\pi}(s, a)$	动作价值函数	action-value function
$Q_{\star}(s, a)$	最优动作价值函数	optimal action-value function
$V_{\pi}(s)$	状态价值函数	state-value function
$V_{\star}(s)$	最优状态价值函数	optimal state-value function
$D_{\pi}(s)$	优势函数	advantage function
$D_{\star}(s)$	最优优势函数	optimal advantage function
$\pi(a s; \theta)$	随机策略网络	stochastic policy network
$\mu(s; \theta)$	确定策略网络	deterministic policy network
$Q(s, a; \mathbf{w})$	深度 Q 网络	deep Q network (DQN)
$q(s, a; \mathbf{w})$	价值网络	value network

目录

1	深度学习基础	1
1.1	线性模型	1
1.2	神经网络	7
1.3	反向传播和梯度下降	10
2	概率论基础与蒙特卡洛	13
2.1	概率论基础	13
2.2	蒙特卡洛	16
3	强化学习基础	24
3.1	基本概念	24
3.2	随机性的来源	28
3.3	回报与折扣回报	30
3.4	价值函数	32
3.5	策略学习和价值学习	34
3.6	实验环境	35
4	DQN 与 Q 学习	37
4.1	DQN	37
4.2	时间差分 (TD) 算法	39
4.3	用 TD 训练 DQN	42
4.4	Q 学习算法	45
4.5	同策略 (On-policy) 与异策略 (Off-policy)	47
5	SARSA 算法	49
5.1	表格形式的 SARSA	49
5.2	神经网络形式的 SARSA	52
5.3	多步 TD 目标	54
5.4	蒙特卡洛与自举	56
6	价值学习高级技巧	59
6.1	经验回放	59
6.2	高估问题及解决方法	63
6.3	对决网络 (Dueling Network)	69
6.4	噪声网络	72

7 策略梯度方法	76
7.1 策略网络	76
7.2 策略学习的目标函数	78
7.3 策略梯度定理的证明	80
7.4 REINFORCE	86
7.5 Actor-Critic	89
8 带基线的策略梯度方法	94
8.1 策略梯度中的基线	94
8.2 带基线的 REINFORCE 算法	97
8.3 Advantage Actor-Critic (A2C)	100
8.4 证明带基线的策略梯度定理	104
9 策略学习高级技巧	105
9.1 Trust Region Policy Optimization (TRPO)	105
9.2 熵正则 (Entropy Regularization)	110
10 连续控制	114
10.1 离散控制与连续控制的区别	114
10.2 确定策略梯度 (DPG)	115
10.3 双延时确定策略梯度 (TD3)	120
10.4 随机高斯策略	124
11 对状态的不完全观测	130
11.1 不完全观测问题	130
11.2 循环神经网络 (RNN)	132
11.3 RNN 作为策略网络	134
12 并行计算	136
12.1 并行计算基础	136
12.2 同步与异步	142
12.3 并行强化学习	145
13 多智能体系统	150
13.1 多智能体系统的设定	150
13.2 多智能体系统的基本概念	152
13.3 实验环境	155
14 合作关系设定下的多智能体强化学习	160
14.1 合作关系设定下的策略学习	161
14.2 合作设定下的多智能体 A2C	162

14.3 三种架构	166
15 非合作关系设定下的多智能体强化学习	174
15.1 非合作关系设定下的策略学习	175
15.2 非合作设定下的多智能体 A2C	178
15.3 三种架构	181
15.4 连续控制与 MADDPG	185
16 注意力机制与多智能体强化学习	191
16.1 自注意力机制	191
16.2 自注意力在中心化训练中的应用	195
17 模仿学习	200
17.1 行为克隆	200
17.2 逆向强化学习	204
17.3 生成判别模仿学习 (GAIL)	206
18 AlphaGo 与蒙特卡洛树搜索	213
18.1 动作、状态、策略网络、价值网络	213
18.2 蒙特卡洛树搜索 (MCTS)	215
18.3 训练策略网络和价值网络	220
A 贝尔曼方程	225