

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

《深度学习》2021-02-09 尚未校对，仅供预览。
如发现错误，请告知作者 shusen.wang@stevens.edu

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	25
3.1	基本概念	25
3.2	随机性的来源	29
3.3	回报与折扣回报	31
3.4	价值函数	33
3.5	策略学习和价值学习	35
3.6	实验环境	36
4	DQN 与 Q 学习	39
4.1	DQN	39
4.2	时间差分 (TD) 算法	41
4.3	用 TD 训练 DQN	44
4.4	Q 学习算法	47
4.5	同策略 (On-policy) 与异策略 (Off-policy)	49
5	SARSA 算法	51
5.1	表格形式的 SARSA	51
5.2	神经网络形式的 SARSA	54
5.3	多步 TD 目标	56
5.4	蒙特卡洛与自举	58
6	价值学习高级技巧	63
6.1	经验回放	63
6.2	高估问题及解决方法	67
6.3	对决网络 (Dueling Network)	73
6.4	噪声网络	77

7 策略梯度方法	81
7.1 策略网络	81
7.2 策略学习的目标函数	83
7.3 策略梯度定理的证明	85
7.4 REINFORCE	91
7.5 Actor-Critic	94
8 带基线的策略梯度方法	101
8.1 策略梯度中的基线	101
8.2 带基线的 REINFORCE 算法	104
8.3 Advantage Actor-Critic (A2C)	107
8.4 证明带基线的策略梯度定理	111
9 策略学习高级技巧	113
9.1 Trust Region Policy Optimization (TRPO)	113
9.2 熵正则 (Entropy Regularization)	118
10 连续控制	123
10.1 离散控制与连续控制的区别	123
10.2 确定策略梯度 (DPG)	124
10.3 双延时确定策略梯度 (TD3)	129
10.4 随机高斯策略	133
11 对状态的不完全观测	139
11.1 不完全观测问题	139
11.2 循环神经网络 (RNN)	141
11.3 RNN 作为策略网络	143
12 并行计算	145
12.1 并行计算基础	145
12.2 同步与异步	151
12.3 并行强化学习	154
13 多智能体系统	159
13.1 多智能体系统的设定	159
13.2 多智能体系统的基本概念	161
13.3 实验环境	164
14 合作关系设定下的多智能体强化学习	169
14.1 合作关系设定下的策略学习	170
14.2 合作设定下的多智能体 A2C	171

14.3 三种架构	175
15 非合作关系设定下的多智能体强化学习	183
15.1 非合作关系设定下的策略学习	184
15.2 非合作设定下的多智能体 A2C	187
15.3 三种架构	190
15.4 连续控制与 MADDPG	194
16 注意力机制与多智能体强化学习	201
16.1 自注意力机制	201
16.2 自注意力在中心化训练中的应用	205
17 模仿学习	211
17.1 行为克隆	211
17.2 逆向强化学习	215
17.3 生成判别模仿学习 (GAIL)	217
18 AlphaGo 与蒙特卡洛树搜索	225
18.1 动作、状态、策略网络、价值网络	225
18.2 蒙特卡洛树搜索 (MCTS)	227
18.3 训练策略网络和价值网络	232
A 贝尔曼方程	237

《深度强化学习》2021-02-09 尚未校对，仅供预览。
如发现错误，请告知作者 shusen.wang@stevens.edu

《深度学习》 2021 - 02 - 09 尚未校对，仅供预览。
如发现错误，请告知作者 shusen.wang@stevens.edu