



10.4 单智能体强化学习经典算法



本章目录

单智能体强化学习经典算法

第十章 集群协同与博弈对抗

- 10.1 集群协同基础与控制架构
- 10.2 一致性与编队控制
- 10.3 智能博弈与强化学习基础
- 10.4 单智能体强化学习经典算法**
- 10.5 多智能体强化学习与规模扩展
- 10.6 集群关键技术与挑战
- 10.7 RflySim集群框架与接口
- 10.8 本章实验案例与综合验证
- 10.9 习题与课堂讨论

本节 4 个教学单元

10.4.1 Q-Learning

教材对应：第10.1.3.4节第1项

10.4.2 深度Q网络DQN

教材对应：第10.1.3.4节第2项

10.4.3 Actor-Critic与DDPG

教材对应：第10.1.3.4节第3–4项

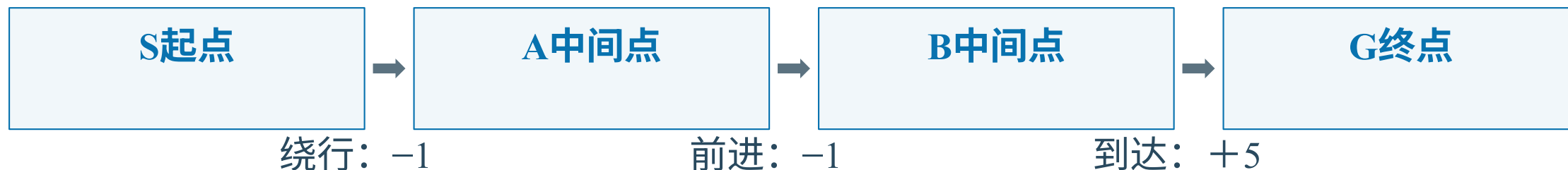
10.4.4 改进算法与选型

教材对应：第10.1.3.4节第5项



10.4.1 Q-Learning: 任务与经验

从已知模型到采样学习



- **教学模型：** S有直接／绕行两个动作，A与B仅能前进，G终止；折扣为0.9。
- **经验：** 一条状态、动作、奖励、下一状态与终止标志，来自实际访问的转移。
- **对照答案：** S直接动作价值为2，绕行为2.15；本节检查估计如何接近这些值。



10.4 单智能体强化学习经典算法



10.4.1 Q-Learning: 式10.30-10.31

Q-Learning用TD误差修正一个表项

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \delta_t \quad (10.30)$$

$$\delta_t = r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \quad (10.31)$$

$$y_t = r_{t+1} + \gamma(1 - d_{t+1}) \max_{a'} Q(s_{t+1}, a')$$

$d_{t+1} = 1$ 表示真正终止，此时目标仅为本次奖励； $0 < \alpha \leq 1$ 。



10.4.1 Q-Learning: 可复算更新 一次Q更新的三个数



经验与估计	给定值
S绕行的旧Q值	1
A的最大Q值	3.5
折扣／学习率	0.9／0.5

目标值

$$y = -1 + 0.9 \times 3.5 = 2.15$$

TD误差

$$\delta = 2.15 - 1 = 1.15$$

更新后的Q值

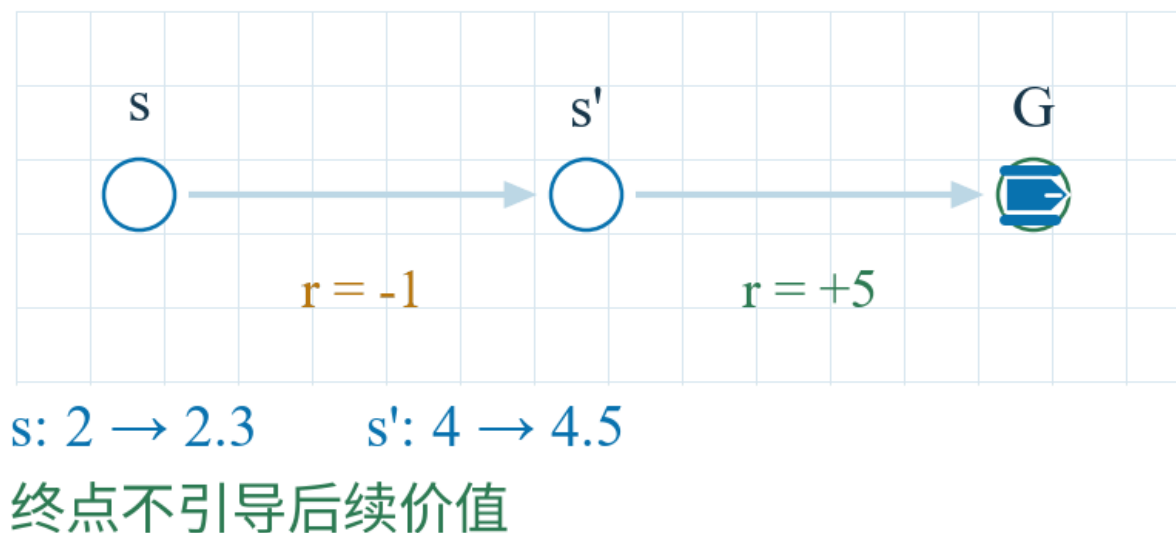
$$1 + 0.5 \times 1.15 = 1.575$$

只更新本次访问的S绕行表项；其他表项保持原值。

10.4.1 Q-Learning: 终止目标

普通转移与终止样本的更新

终止转移



$y = 5$ (done)

旧估计 4

更新后 4.5



$\alpha = 0.5, \gamma = 0.9$

图：普通转移与终止样本的Q值更新动画

独立一步算例，非四状态导航轨迹；真正终止时不引导后续价值。

10.4.1 Q-Learning: 异策略与探索

探索策略可以不同于学习目标

概念	作用	例子
行为策略	产生实际经验	探索时偶尔选当前较差动作
目标策略	定义正在学习的行为	目标选择下一状态最高Q
异策略off-policy	行为与目标策略可以不同	历史转移可用于更新
同策略on-policy	学习数据对应当前或近期采样策略	PPO使用本轮收集的数据批次

两个动作、唯一贪婪动作且 $\varepsilon = 0.2$ ：贪婪动作概率为 $0.8 + 0.2/2 = 0.9$ 。

10.4.1 Q-Learning: 导航Q表收敛过程

完整采样循环与学习结果

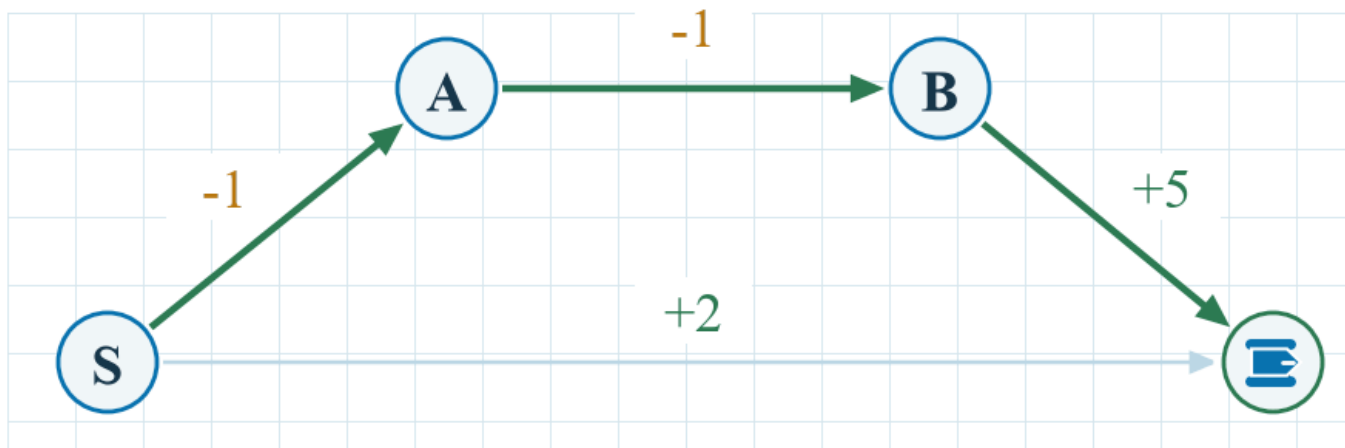
采样轮次	S绕行	A前进	B到G	S直接
0	0	0	0	0
1	-0.5000	-0.5000	2.5000	1.0000
4	-0.2906	2.1562	4.6875	1.8750
8	1.6002	3.3457	4.9805	1.9922
16	2.1418	3.4988	4.9999	2.0000
32	2.1500	3.5000	5.0000	2.0000

每轮顺序：S→A、A→B、B→G、S→G。全零初值，学习率0.5，折扣0.9。

10.4.1 Q-Learning: 采样与路径选择

持续采样，让起点逐渐看见绕行的收益

固定顺序采样 $S \rightarrow A \rightarrow B \rightarrow G / S \rightarrow G$



两条路径都持续访问 $\alpha = 0.5, \gamma = 0.9$

采样轮次 32

起点两个动作的Q值

绕行 2.1500

直达 2.0000

起点当前选择:

绕行

图：固定顺序覆盖两条路径；展示第1、4、8、16、32轮，未展示轮次也参与计算
第8轮仍偏向直达，第16轮转向绕行；当前较差的动作也要获得样本。

10.4.1 Q-Learning: 适用范围

收敛条件与终止处理

- **表格适用：**状态动作可枚举；每个表项存储一个独立估计，高维连续状态不适合逐项建表。
- **随机收敛条件：**有限平稳MDP、有界奖励、充分访问，以及适当衰减的学习率。

$$\sum_k \alpha_k(s, a) = \infty, \quad \sum_k \alpha_k^2(s, a) < \infty$$

真实终止：target=本次奖励。仅时间截断：用截断前的末状态估计后续价值。

本章目录

单智能体强化学习经典算法

第十章 集群协同与博弈对抗

- 10.1 集群协同基础与控制架构
- 10.2 一致性与编队控制
- 10.3 智能博弈与强化学习基础
- 10.4 单智能体强化学习经典算法**
- 10.5 多智能体强化学习与规模扩展
- 10.6 集群关键技术与挑战
- 10.7 RflySim集群框架与接口
- 10.8 本章实验案例与综合验证
- 10.9 习题与课堂讨论

本节 4 个教学单元

10.4.1 Q-Learning

教材对应：第10.1.3.4节第1项

10.4.2 深度Q网络DQN

教材对应：第10.1.3.4节第2项

10.4.3 Actor-Critic与DDPG

教材对应：第10.1.3.4节第3–4项

10.4.4 改进算法与选型

教材对应：第10.1.3.4节第5项

10.4.2 深度Q网络DQN: 图10.6

Q表与神经网络怎样估计价值

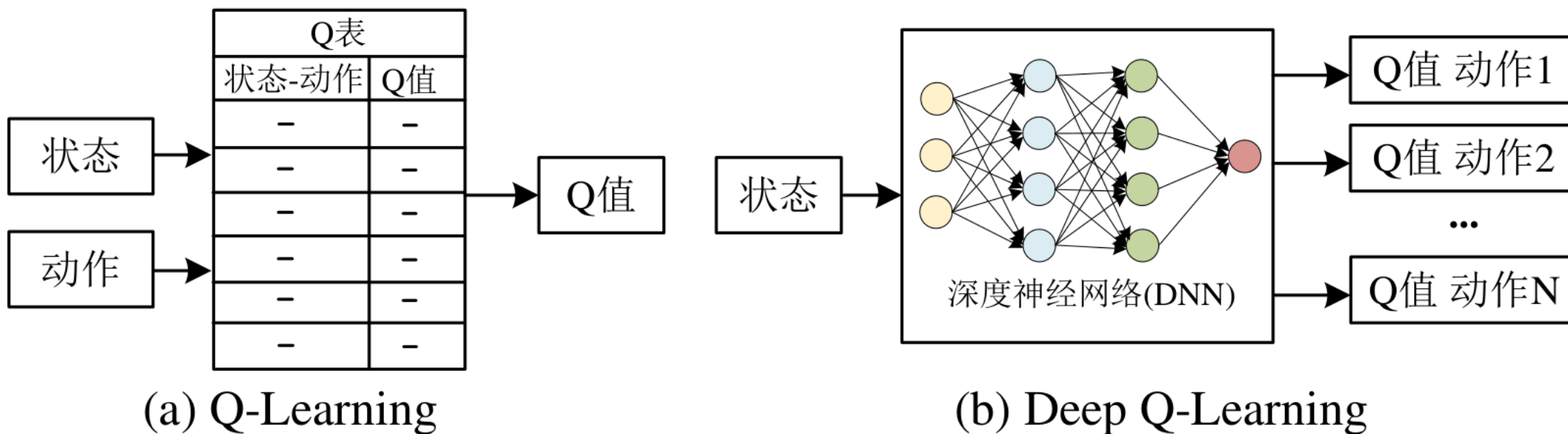


图10.6 Q-Learning和DQN在Q值估计的区别

DQN用共享参数近似Q值；输入可高维，标准输出对应有限离散动作。

10.4.2 深度Q网络DQN: 状态、输出与选择

网络输出Q值，再由比较得到动作



$$Q(s, \cdot; \theta) = [1.2, 2.4, 0.8]$$

动作	左	前	右
Q值	1.2	2.4	0.8
贪婪选择	否	是	否

10.4.2 深度Q网络DQN：稳定训练机制

回放池与目标网络各管一件事



$$D: (s, a, r, s', d)$$

- **回放池：**减少相邻样本相关性，重复利用已发生的经验；没有覆盖的动作仍缺少证据。
- **目标网络：**参数暂时保持或缓慢变化，让主网络拟合的目标不立即跟着自己变化。



10.4 单智能体强化学习经典算法



10.4.2 深度Q网络DQN: 式10.32-10.33

DQN的目标与平方误差

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s') \sim D} [(y - Q(s, a; \theta))^2] \quad (10.32)$$

$$y = r + \gamma \max_{a'} Q(s', a'; \theta^-) \quad (10.33)$$

$$y = r + \gamma(1 - d) \max_{a'} Q(s', a'; \theta^-)$$

式10.33为非终止情形，最后一式用于混合批次；目标在本次更新中固定。





10.4 单智能体强化学习经典算法



10.4.2 深度Q网络DQN：式10.34与一步计算

梯度下降把预测推向目标

$$\theta \leftarrow \theta - \alpha \nabla_{\theta} (y - Q(s, a; \theta))^2 \quad (10.34)$$

$$Q = \theta = 1, \quad y = 2.15, \quad \mathcal{L} = 1.3225$$

$$\nabla_{\theta} \mathcal{L} = 2(\theta - y) = -2.3$$

$$\theta_{\text{new}} = 1 - 0.1(-2.3) = 1.23$$

教学假设：只有一个标量参数；一般神经网络按链式法则计算参数梯度。



10.4.2 深度Q网络DQN：回放与目标冻结

随机回放旧样本，目标网络分段冻结

经验池

e_1 $S \rightarrow A$	e_2 $A \rightarrow B$	e_3 $B \rightarrow G$
e_4 $S \rightarrow G$	e_5 $A \rightarrow B$	e_6 $S \rightarrow A$

已完成 8

$k = 8$

均匀抽样序列

1 5 4 3 3 6 1 5

参数快照

主网 θ

$S \rightarrow A$	$A \rightarrow B$	$B \rightarrow G$	$S \rightarrow G$
1.503	3.241	4.360	1.200

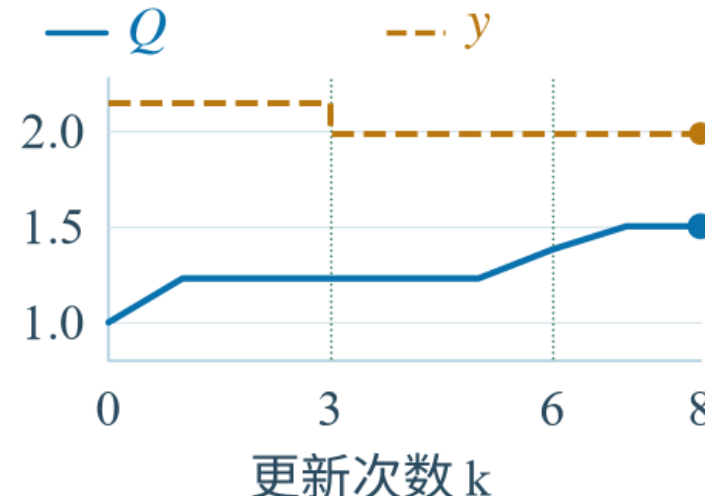
目标冻结 θ^-

1.382	3.320	4.360	1.200
-------	-------	-------	-------

第3、6步后复制

第6步: y 不变

固定探针 $S \rightarrow A$



$Q = 1.503$

$y = 1.988$

图：线性Q单样本算例（折扣0.9，学习率0.1）；目标网络每3步硬同步
 初始Q按S绕行、A前进、B到G、S直接为1、3.5、4、1；G终止。
 六条样本的奖励依次为-1、-1、5、2、-1、-1；蓝线仅跟踪S绕行。

10.4.2 深度Q网络DQN：完整算法流程

DQN训练与执行的数据流



阶段	需要的对象	动作如何选择
训练	主网络、目标网络、回放池、环境	主网络Q值+探索策略
评估／执行	固定主网络、当前状态、控制接口	通常取最大Q值对应动作

目标网络按间隔硬复制；训练停止依据独立评估指标、预算与稳定性。

10.4.2 深度Q网络DQN: Double、Dueling与优先回放

DQN改进分别针对不同问题

改进	针对的问题	主要机制
Double DQN	最大值估计容易偏高	主网络选下一动作，目标网络评价该动作
Dueling DQN	状态好坏与动作差异纠缠	状态价值分支与动作优势分支合成Q值
优先经验回放	平均抽样可能低效	按绝对TD误差等确定优先级并修正偏差

优势表示某动作相对该状态通常水平的好坏；高TD误差也可能来自噪声。



10.4 单智能体强化学习经典算法



本章目录

单智能体强化学习经典算法

第十章 集群协同与博弈对抗

- 10.1 集群协同基础与控制架构
- 10.2 一致性与编队控制
- 10.3 智能博弈与强化学习基础
- 10.4 单智能体强化学习经典算法**
- 10.5 多智能体强化学习与规模扩展
- 10.6 集群关键技术与挑战
- 10.7 RflySim集群框架与接口
- 10.8 本章实验案例与综合验证
- 10.9 习题与课堂讨论

本节 4 个教学单元

10.4.1 Q-Learning

教材对应：第10.1.3.4节第1项

10.4.2 深度Q网络DQN

教材对应：第10.1.3.4节第2项

10.4.3 Actor-Critic与DDPG

教材对应：第10.1.3.4节第3–4项

10.4.4 改进算法与选型

教材对应：第10.1.3.4节第5项



10.4.3 Actor-Critic与DDPG: Actor-Critic框架

Actor选动作, Critic提供评价



- **Actor**: 改变动作选择, 使长期目标改善。
- **Critic**: 用奖励和后续估计学习价值, 提供策略优化的方向。

10.4.3 Actor-Critic与DDPG: A2C与A3C

优势函数衡量动作相对水平

$$A_{\mu}(s, a) = Q_{\mu}(s, a) - V_{\mu}(s)$$

策略等概率选直接或绕行: $V_{\mu}(S) = 0.5 \times 2 + 0.5 \times 2.15 = 2.075$

$$A_{\mu}(S, \text{via}) = 2.15 - 2.075 = 0.075$$

方法	采样与更新组织	共同内容
A2C	多个环境采样后同步更新	Actor依据优势改进策略
A3C	多个工作线程异步更新	Critic估计价值, 减少梯度方差

10.4.3 Actor-Critic与DDPG：图10.7(a)总览 DDPG训练包含四个网络

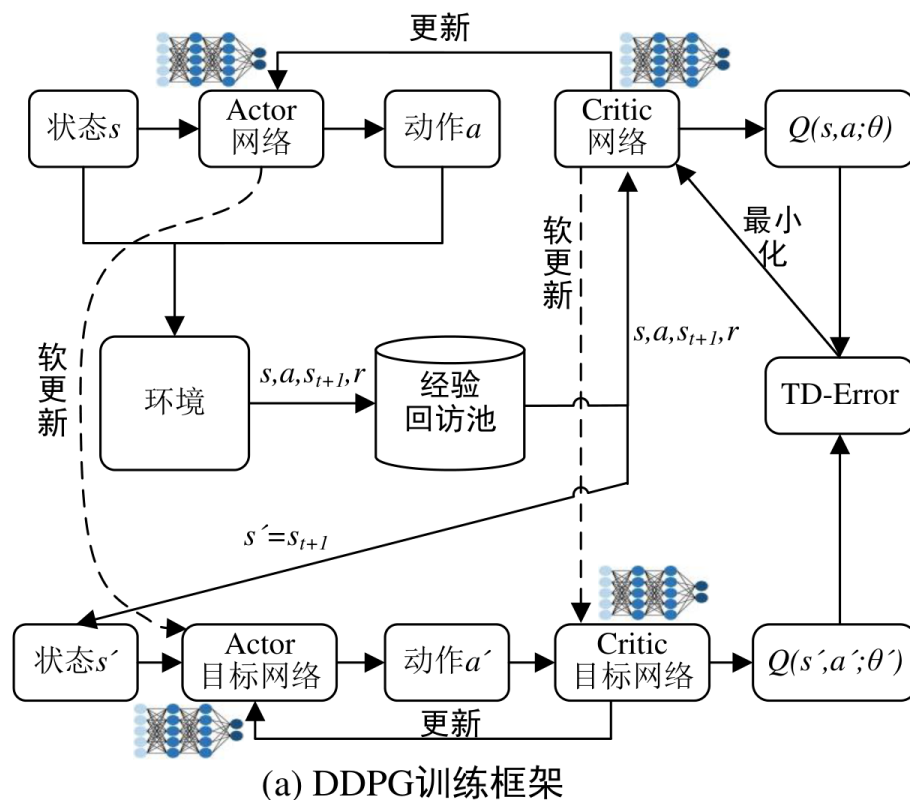


图10.7(a) DDPG训练框架

- **在线Actor:** 从状态生成连续动作。
- **在线Critic:** 估计当前状态动作价值。
- **两个目标网络:** 共同构造学习目标。

底部“更新”箭头不适用。
目标网络仅由各自在线网络
按式10.37软更新。

10.4.3 Actor-Critic与DDPG：图10.7(a)在线分支 在线网络采样并学习动作价值



Actor更新时沿Critic对动作的梯度改进；Critic本身用经验的TD目标训练。

10.4.3 Actor-Critic与DDPG：图10.7(a)目标分支

目标Actor和目标Critic构造后续值



$$s' \rightarrow \mu'(s') \rightarrow Q'(s', \mu'(s'))$$

- **输入关系：** 目标Critic同时接收下一状态和目标Actor生成的动作。
- **更新关系：** 两目标网络仅按式10.37由各自在线网络软更新，不互相梯度更新。

10.4.3 Actor-Critic与DDPG: 式10.35

Actor最大化Critic给出的价值

$$J(\theta^\mu) = \mathbb{E}_{s \sim D}[Q(s, \mu(s; \theta^\mu); \theta^Q)] \quad (10.35)$$

$$\nabla_{\theta^\mu} J \approx \mathbb{E}[\nabla_a Q(s, a)|_{a=\mu(s)} \nabla_{\theta^\mu} \mu(s)]$$

- **更新方向:** Actor做梯度上升以增大Q; 优化器可通过最小化负的平均Q实现。
- **固定对象:** 这一步固定Critic参数, 但保留它对动作的梯度, 才能传回Actor。

10.4.3 Actor-Critic与DDPG：式10.36

Critic拟合带目标动作的TD目标

$$\mathcal{L}(\theta^Q) = \mathbb{E}_{(s,a,r,s') \sim D} [(y - Q(s, a; \theta^Q))^2] \quad (10.36)$$

$$y = r + \gamma Q'(s', \mu'(s'))$$

$$y = r + \gamma(1 - d)Q'(s', \mu'(s'))$$

第二行是教材非终止目标，第三行补入终止标志；目标Actor和Critic均不参与此步梯度更新。

10.4.3 Actor-Critic与DDPG：式10.37

软更新让目标网络缓慢跟随

$$\theta^{Q'} \leftarrow \tau \theta^Q + (1 - \tau) \theta^{Q'} \quad (10.37)$$

$$\theta^{\mu'} \leftarrow \tau \theta^{\mu} + (1 - \tau) \theta^{\mu'}$$

$$\theta_{\text{target,new}} = 0.01 \times 2 + 0.99 \times 1 = 1.01$$

软更新比例越小，目标跟随越慢；两个目标网络分别跟随自己的在线网络。

10.4.3 Actor-Critic与DDPG：式10.38

训练探索与执行动作不同

$$a_t = \mu(s_t; \theta^\mu) + \mathcal{N}_t \quad (10.38)$$

$$a_t = \text{clip}(\mu(s_t) + \mathcal{N}_t, a_{\min}, a_{\max})$$

阶段	动作生成	用途
训练	策略输出+噪声，再限幅	探索并生成覆盖更广的经验
固定策略评估	策略输出，再执行相同限幅	比较策略本身的任务表现

10.4.3 Actor-Critic与DDPG：标量Actor-Critic算例

连续动作的评价与改进可以手算

$$Q(a) = 4 - (a - 0.5)^2, \quad a \in [-1, 1], \quad a_0 = 0$$

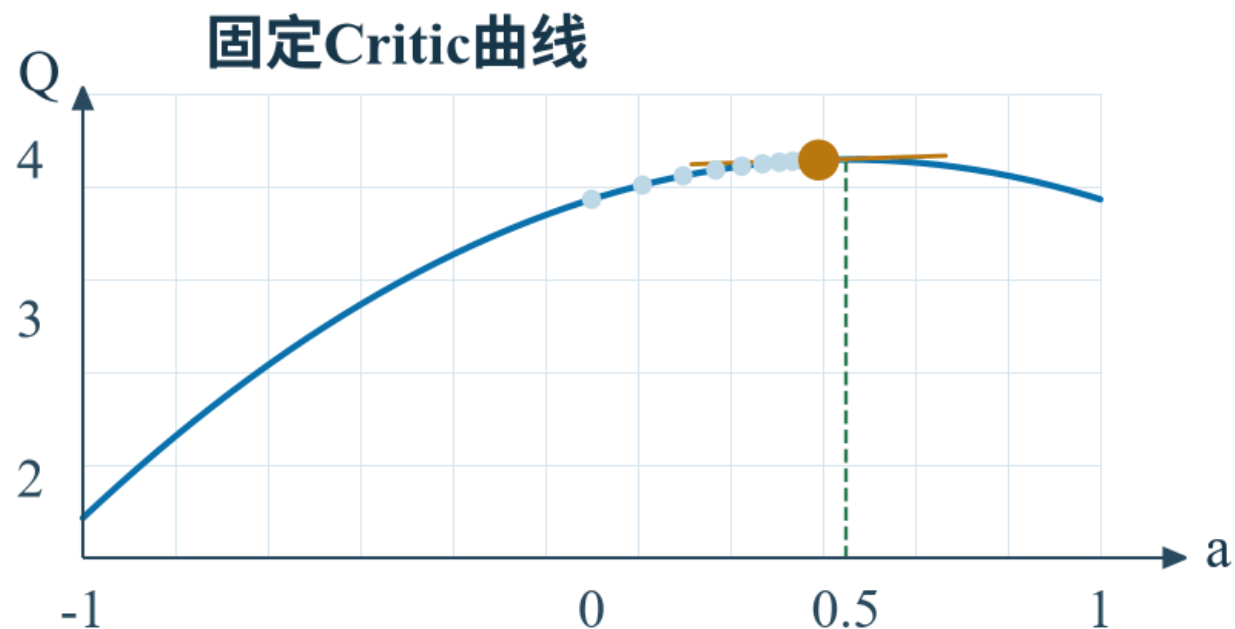
本例固定状态与Critic；Actor参数就是动作，学习率为0.1。

$$a_{k+1} = a_k + 0.1[-2(a_k - 0.5)] = 0.8a_k + 0.1$$

更新次数	动作	固定Critic给出的Q值
0	0	3.7500
1	0.1000	3.8400
5	0.3362	3.9732
10	0.4463	3.9971

10.4.3 Actor-Critic与DDPG：动作梯度

Actor沿着Critic曲线提高动作评价



参数更新次数 10

Actor输出 0.4463

$Q(a) = 3.9971$

$dQ/da = 0.1074$

$a^* = 0.5$

图：固定Critic的十次梯度上升；金色切线表示当前位置的斜率

靠近曲线峰值时斜率变小，动作更新也随之缩小；本例只更新Actor。

10.4.3 Actor-Critic与DDPG：采样、学习与评估

DDPG的完整训练循环



循环环节	参数怎样变化	检查对象
Critic更新	只更新在线Critic	TD误差、终止mask、目标冻结
Actor更新	只更新在线Actor	动作梯度、限幅与Q估计偏差
独立评估	全部策略参数固定	回报、成功率与约束违反次数



10.4 单智能体强化学习经典算法



本章目录

单智能体强化学习经典算法

第十章 集群协同与博弈对抗

- 10.1 集群协同基础与控制架构
- 10.2 一致性与编队控制
- 10.3 智能博弈与强化学习基础
- 10.4 单智能体强化学习经典算法**
- 10.5 多智能体强化学习与规模扩展
- 10.6 集群关键技术与挑战
- 10.7 RflySim集群框架与接口
- 10.8 本章实验案例与综合验证
- 10.9 习题与课堂讨论

本节 4 个教学单元

10.4.1 Q-Learning

教材对应：第10.1.3.4节第1项

10.4.2 深度Q网络DQN

教材对应：第10.1.3.4节第2项

10.4.3 Actor-Critic与DDPG

教材对应：第10.1.3.4节第3–4项

10.4.4 改进算法与选型

教材对应：第10.1.3.4节第5项



10.4.4 改进算法与选型：双Critic与延迟更新

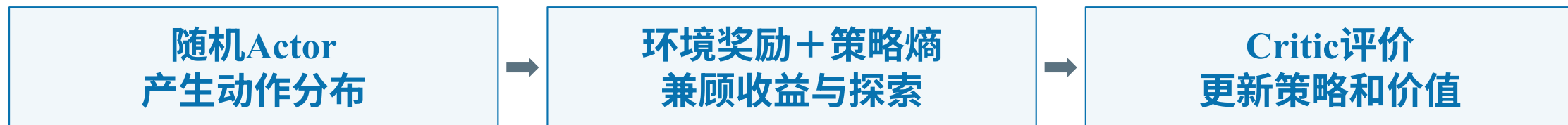
TD3针对价值高估与策略振荡

- **双Critic**：对同一目标动作估值并取较小值，减少过估计被Actor放大的机会。
- **延迟更新**：Critic更新多次后再更新Actor，使策略使用相对稳定的评价。
- **目标平滑**：目标动作加入限幅噪声，再对动作限幅，减弱尖锐虚高Q峰的影响。

策略会利用评价误差，改进评价也是改进决策。

10.4.4 改进算法与选型：随机策略与探索

SAC同时优化收益与策略熵



- **熵**：衡量动作分布的随机程度；分布越集中，通常熵越低。
- **温度**：决定熵项权重，可按目标熵调整探索强度。
- **梯度**：将采样写成参数与随机噪声的可微组合，使策略参数可以得到梯度。

10.4.4 改进算法与选型：概率比剪裁

PPO用剪裁目标抑制过大更新

$$\rho_t = \frac{\pi_{\theta}(a_t | s_t)}{\pi_{\text{old}}(a_t | s_t)}$$

$$L^{\text{clip}} = \mathbb{E}[\min(\rho_t A_t, \text{clip}(\rho_t, 1 - \epsilon, 1 + \epsilon) A_t)]$$

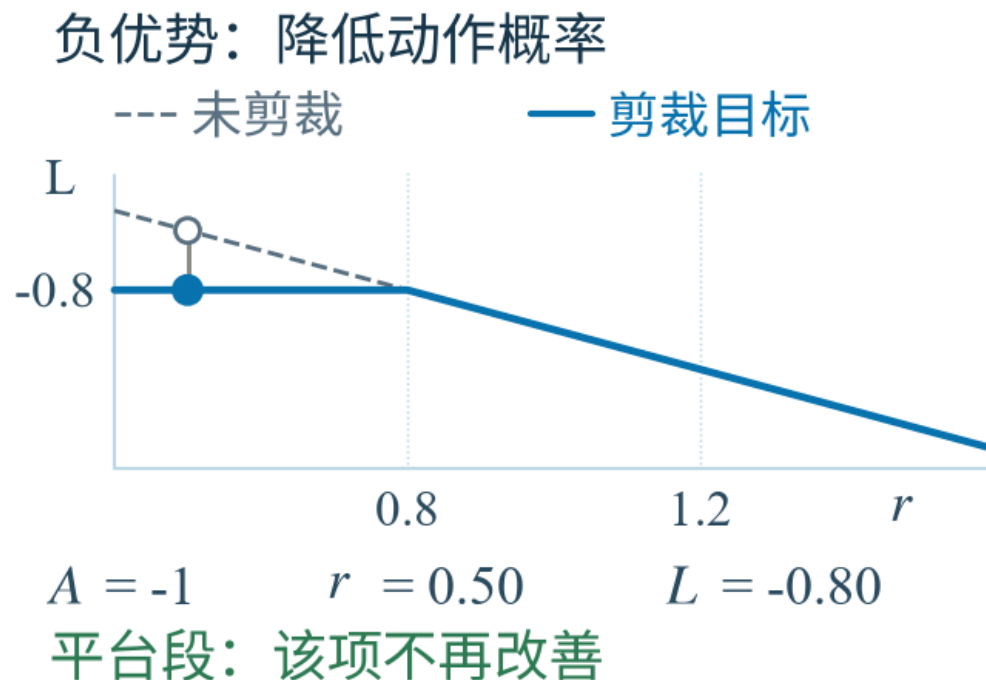
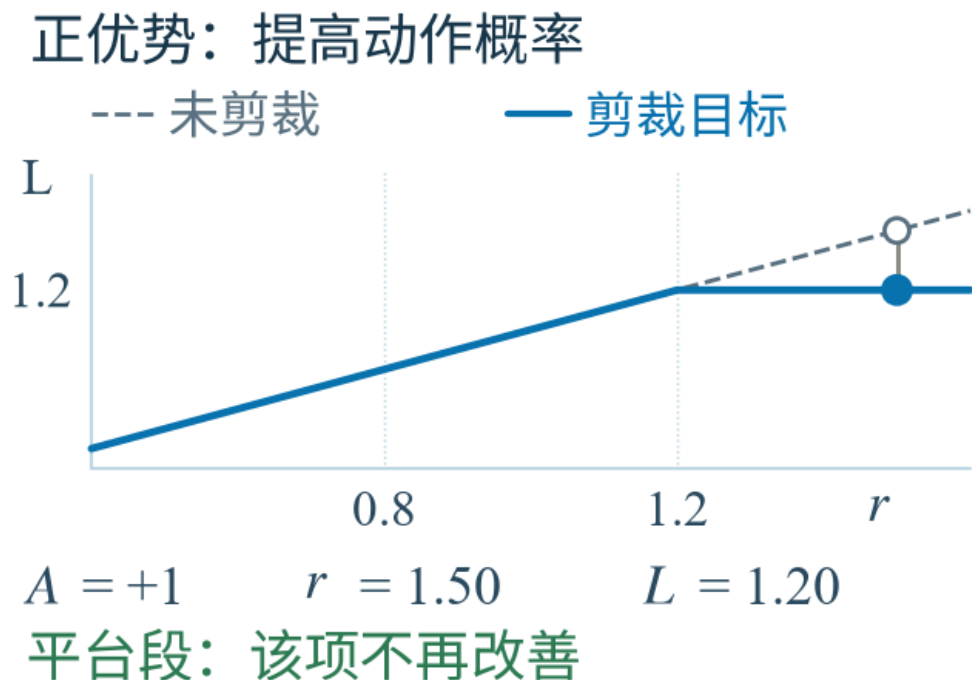
- **方向：**最大化剪裁目标； ρ_t 为新旧动作概率（密度）比。
- **参数：** A_t 为旧策略优势估计， ϵ 为剪裁宽度。
- **流程：**采样后以GAE（广义优势估计）加权多步TD信息，折中偏差与方差；同批有限次更新再采样。

$$A_t = 1, \quad \epsilon = 0.2, \quad \rho_t = 1.5: \quad \min(1.5, 1.2) = 1.2$$

该正优势样本的目标项在概率比超过1.2后不再增加。

10.4.4 改进算法与选型：PPO单样本目标

正负优势的剪裁平台位于不同一侧



图： r 为前页概率比，剪裁宽度0.2；实心点为取较小值后的目标
正优势在右侧变平，负优势在左侧变平；实际概率比不被硬限幅。

10.4.4 改进算法与选型：TD-MPC与TD-MPC2

TD-MPC在学习模型中规划短期轨迹



- **TD-MPC**: 学习任务相关的动态模型、奖励与终端价值，在潜在空间做滚动规划。
- **TD-MPC2**: 引入可扩展世界模型、多任务嵌入、Q值集成和价值尺度处理。
- **计算取舍**: 执行时需要模型内规划；模型误差与规划时延会直接影响决策。

10.4.4 改进算法与选型：算法对照

按动作空间、数据与执行预算选型

方法	常见动作空间	数据与执行特点
Q表/DQN	有限离散动作	异策略；执行时比较各动作Q值
DDPG/TD3	连续动作	异策略回放；确定性Actor前向计算
SAC	常见为连续动作	异策略回放；随机策略与熵目标
PPO	离散或连续动作	近期策略数据；一批数据有限次更新
TD-MPC(2)	连续动作	学习潜在模型；执行时还要滚动规划

10.4.4 改进算法与选型：材料与任务边界

RflySim的单机PPO资料对应什么

资料目录（完整版及以上，按发布包核对）

10.RflySimSwarm/3.CustExps/e0.CustApiExps/3.MARLExps

文件	读取时关注什么
Readme.pdf	Genesis + RflySim四旋翼加速度悬停控制的任务与依赖
train/env_acc.py	观测、动作、奖励与终止如何定义
train/train_acc.py	PPO训练配置与并行采样
sil/rfAcc.py	固定策略如何转换为SITL控制输入

材料索引经只读核查，运行效果未实测；多环境采样与多智能体协作是两类问题。



10.4 单智能体强化学习经典算法



10.4.4 改进算法与选型：继续学习

算法原理与实现阅读入口

DDPG目标、终止处理与完整伪代码

<https://spinningup.openai.com/en/latest/algorithms/ddpg.html>

PPO剪裁目标与策略更新机制

<https://spinningup.openai.com/en/latest/algorithms/ppo.html>

TD-MPC2论文、世界模型与实现入口

<https://www.tdmcp2.com/>





10.4 单智能体强化学习经典算法



练习与讨论

课堂练习：目标值与网络职责

- **问题1：** 终止奖励5，旧Q为2，学习率0.5。下一状态网络输出100，新Q是多少？
- **问题2：** 在线参数2，目标参数1，软更新比例0.1。更新后的目标参数是多少？
- **问题3：** DDPG固定策略执行时，是否必须运行Critic、目标网络和回放池？
- **问题4：** 同一PPO策略在100个独立单机环境采样，是否就是100智能体协同？





10.4 单智能体强化学习经典算法



参考要点

参考要点：终止后没有未来回报

$$y = 5, \quad Q_{\text{new}} = 2 + 0.5(5 - 2) = 3.5$$

$$\theta'_{\text{new}} = 0.1 \times 2 + 0.9 \times 1 = 1.1$$

- **问题3参考：**固定Actor、观测预处理、动作限幅和接口参与执行；Critic等通常只供训练。
- **问题4参考：**这是并行单机采样；多智能体在共同环境中形成相互影响的决策。





10.4 单智能体强化学习经典算法



本节小结

价值学习与连续动作优化

- **Q-Learning与DQN**: 按样本修正动作价值; 网络近似使样本互相影响, 需管理数据与目标。
- **Actor-Critic与DDPG**: 策略生成动作, Critic评价; 连续动作可沿价值梯度优化。
- **算法选择**: 比较动作空间、经验利用、探索方式与执行计算预算, 再设计评估。
- **易错点**: 终止目标无后续项; 目标网络不随损失反传; 评估参数与探索规则固定。

