本章目录

多智能体强化学习与规模扩展

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

10.5.1 马尔可夫博弈与学习架构

教材对应：第10.1.3.5节 · 主流框架与问题描述

10.5.2 MADDPG数据流与更新

教材对应：第10.1.3.5节 · 网络更新与训练过程

10.5.3 大规模集群扩展

教材对应：第10.1.3.6节 · 挑战与优化方案

10.5.4 方案选择与验证

教材对应：第10.1.3.5–10.1.3.6节 · 方案选择

10.5.1 马尔可夫博弈与学习架构：任务与奖励

多个学习者共同改变环境

多智能体强化学习（MARL）：多个智能体在共享环境中交互，同时改进各自策略。

贯穿任务：双机沿巡检航线保持间距。各机输出速度，邻机动作会改变相对距离。

合作与竞争：共享任务奖励表示合作；奖励冲突表示竞争，也可同时存在两种关系。

前置知识：课件10.3的回报与价值函数、课件10.4的Actor策略网络和Critic价值网络。

10.5.1 马尔可夫博弈与学习架构：问题建模

马尔可夫博弈如何描述多机决策

$$\mathcal{G} = \langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^{N}, \{\mathcal{O}_i\}_{i=1}^{N}, \mathcal{P}, \{r_i\}_{i=1}^{N}, \gamma \rangle \tag{10.39}$$

符号与含义	巡检任务中的对象
$s \in \mathcal{S}, \quad o_i \in \mathcal{O}_i$	全体位置与任务状态；本机及邻域观测
$\mathbf{a} = (a_1, \ldots, a_N)$	所有机器本次同时执行的速度指令
$\mathcal{P}(s' \mid s, \mathbf{a}),\ r_i$	联合动作决定转移；各机获得即时奖励
$\gamma \in (0,1), \quad o_i = O_i(s)$	折扣因子；另需指定观测规则或观测概率

联合观测可以用于训练，但不必然包含完整的真实状态。

10.5.1 马尔可夫博弈与学习架构：可复算转移

一步联合动作决定下一次队形误差

教学假设：一维质点，步长1 s，期望间距4 m，速度即时执行。

$$p_i^+ = p_i + v_i \Delta t, \quad e^+ = p_2^+ - p_1^+ - 4, \quad r = -(e^+)^2$$

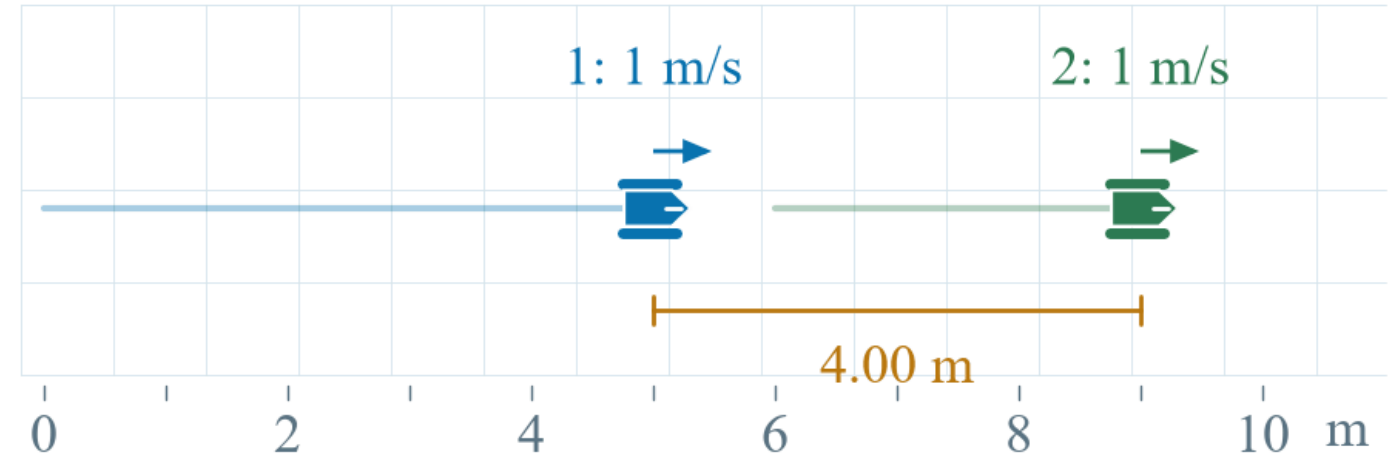
步	初始位置 / m	联合速度 / (m/s)	下一位置 / m	奖励
1	(0, 6)	(2, 1)	(2, 7)	-1
2	(2, 7)	(2, 1)	(4, 8)	0
3	(4, 8)	(1, 1)	(5, 9)	0

第一步误差由2 m降为1 m；保持原速度差会越过目标间距，因此到位后改为同速。

10.5.1 马尔可夫博弈与学习架构：联合动作

两机都在运动，速度差决定间距怎样变化

两机同时移动



(0, 6) → (2, 7) → (4, 8) → (5, 9) m

t = 3.00 s

期望间距 4 m

p = (5.00, 9.00)

r[3] = 0

同速保持间距

图：一维双机运动；轨迹、位置、间距与各步终点奖励来自同一次计算

间距从6 m减到4 m后，两机改为同速；固定地面网格显示它们仍在前进。

10.5.1 马尔可夫博弈与学习架构：主流框架

独立学习、通信与CTDE的区别

方法	训练时的信息与更新	执行时的动作输入
独立学习	各机依自己的数据和奖励更新	本机观测；相互影响未被显式建模
基于通信的学习	学习如何利用或形成消息	本机观测与允许获得的邻居消息
CTDE	联合信息辅助Critic或联合损失	各机策略仅用部署时可获得的信息

CTDE全称为集中式训练与分布式执行。通信机制可以与独立学习或CTDE组合。

10.5.1 马尔可夫博弈与学习架构：非平稳性

邻机策略变化造成非平稳学习目标

教学假设：双机共用单通道。双方前进奖励-10，只有本机前进奖励+2。

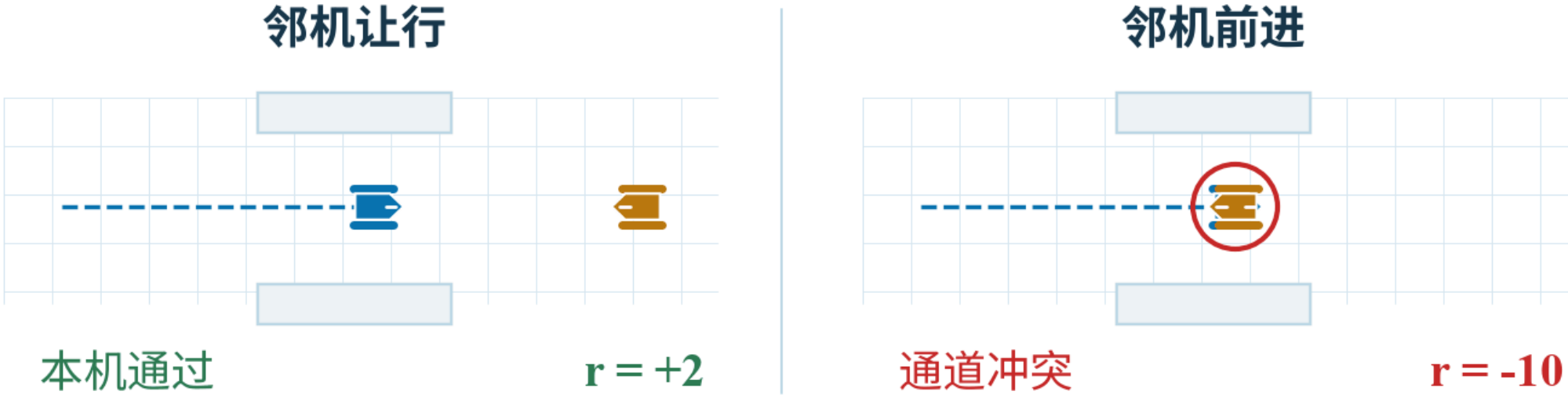
$$\mathbb{E}[r_1 \mid a_1 = \text{前进}] = -10p + 2(1-p) = 2 - 12p$$

邻机前进概率	本机同一动作的期望奖励	解释
0.2	$2 - 12 \cdot 0.2 = -0.4$	邻机多数让行
0.8	$2 - 12 \cdot 0.8 = -7.6$	邻机策略改变，旧评价失准

联合信息帮助解释交互，但不保证学习收敛或消除全部非平稳性。

10.5.1 马尔可夫博弈与学习架构：策略变化

自己做同一个动作，邻机改变了结果



图：蓝色为本机、赭黄色为邻机；左右保持本机初态与动作相同，只改变邻机动作

邻机越倾向抢先进入通道，本机“前进”的期望奖励就越低。

本章目录

多智能体强化学习与规模扩展

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

10.5.1 马尔可夫博弈与学习架构

教材对应：第10.1.3.5节 · 主流框架与问题描述

10.5.2 MADDPG数据流与更新

教材对应：第10.1.3.5节 · 网络更新与训练过程

10.5.3 大规模集群扩展

教材对应：第10.1.3.6节 · 挑战与优化方案

10.5.4 方案选择与验证

教材对应：第10.1.3.5–10.1.3.6节 · 方案选择

10.5.2 MADDPG数据流与更新：教材框架图

从DDPG扩展到多个Actor与Critic

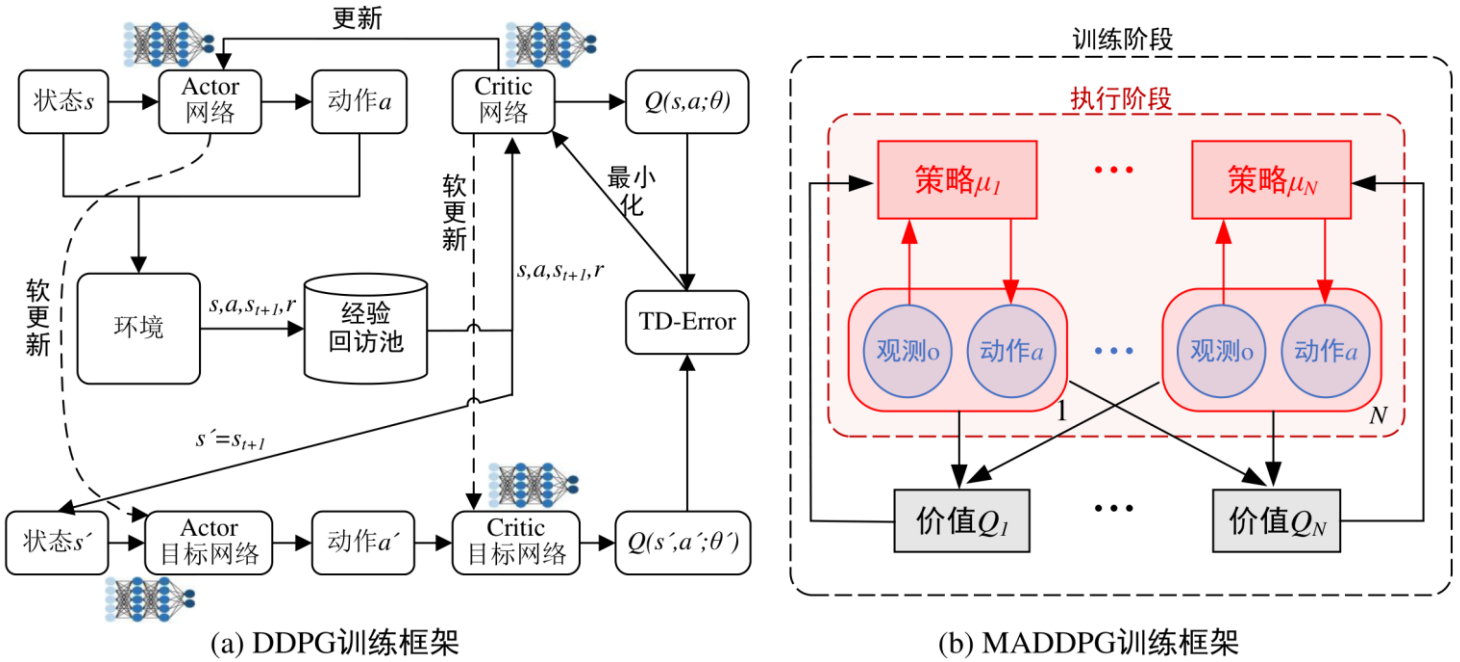
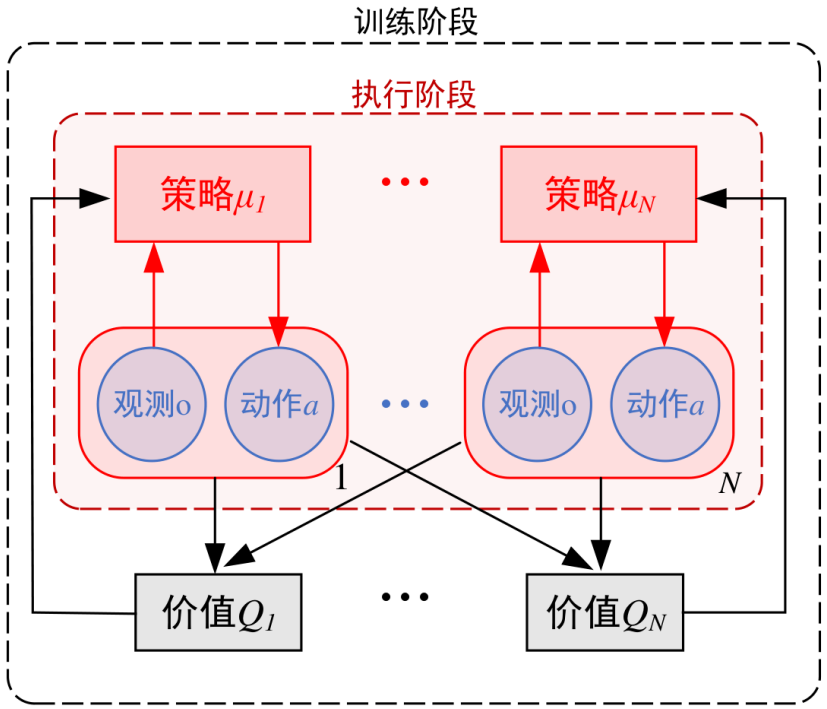


图10.7 DDPG与MADDPG框架

在线网络学习、目标网络软更新；多智能体训练联合评估，执行时各机使用自己的策略。

10.5.2 MADDPG数据流与更新：教材图局部放大

MADDPG子图中的两条信息路径



(b) MADDPG训练框架

图10.7(b) MADDPG训练框架

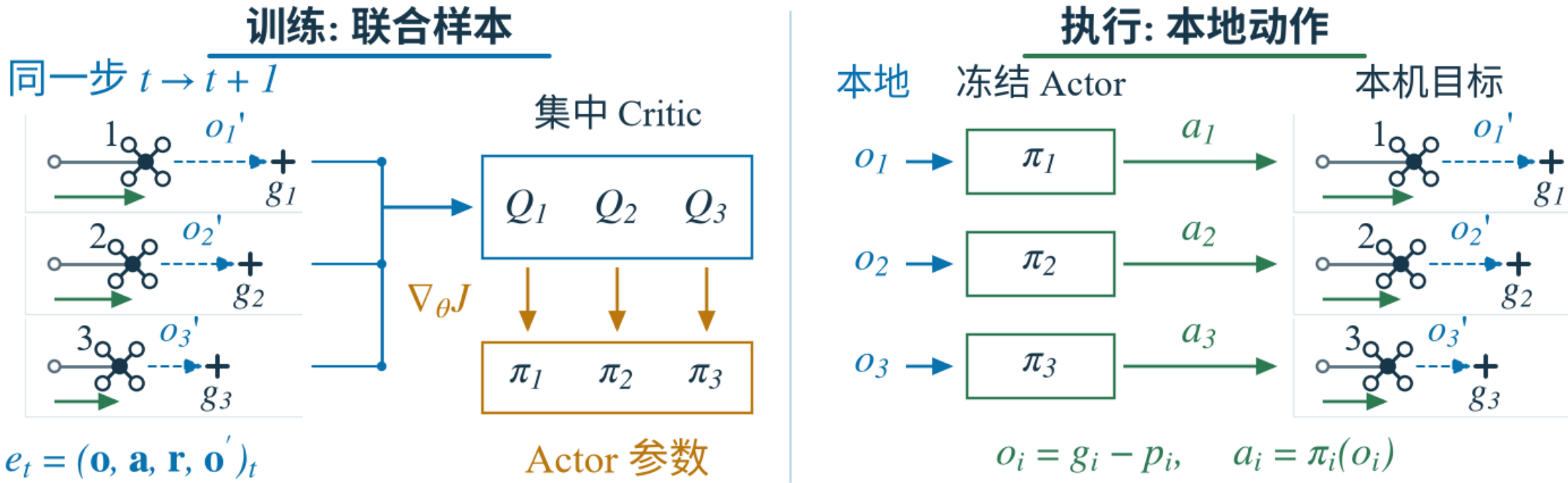
红色边界内：本地观测进入本机策略，输出本机动作。

价值网络：接收联合观测与动作，评价对应智能体收益。

外圈反馈：训练中更新策略；部署时不需要集中式Critic。

10.5.2 MADDPG数据流与更新：CTDE信息边界

训练联合评价，执行本地决策



图：CTDE集中训练与分布执行示意

训练Critic使用联合信息；执行Actor仅用本地观测，图中未加入机间通信。

10.5.2 MADDPG数据流与更新：经验回放

回放中的一条经验来自同一次交互

$$(\mathbf{x}, \mathbf{a}, \mathbf{r}, \mathbf{x}') \tag{10.45}$$

$$\mathbf{x} = (o_1, \ldots, o_N), \quad \mathbf{a} = (a_1, \ldots, a_N), \quad \mathbf{r} = (r_1, \ldots, r_N)$$

对齐：同一环境、同一步、同一组节点ID；下一观测与执行过的动作相对应。

采样：整条联合经验一起抽取；本机Actor只取其中属于自己的观测。

实现补充：记录终止标志与有效节点，另存训练可用全局状态；观测不完整时可用历史或循环网络。

10.5.2 MADDPG数据流与更新：目标值

目标Critic给出下一时刻的价值

$$y_i = r_i + \gamma Q_i'(s', a_1', \ldots, a_N') \tag{10.40}$$

$$a_j' = \mu_j'(o_j')$$

符号：此处Critic使用全局状态；目标Actor生成下一动作，目标Critic估计收益；目标计算不反传梯度。

教学数值：即时奖励1，折扣0.9，目标价值2，则目标值为 $1 + 0.9 \times 2 = 2.8$。

终止条件：式10.40用于非终止样本。真正终止时目标为即时奖励；时间截断须按任务定义处理。

10.5.2 MADDPG数据流与更新：价值损失

Critic用TD误差修正价值估计

$$\mathcal{L}_i(\theta_i^Q) = \mathbb{E}_{(s,\mathbf{a},r_i,s')}\left[(y_i - Q_i(s,\mathbf{a};\theta_i^Q))^2\right] \tag{10.41}$$

$$\delta = 2.8 - 1.5 = 1.3, \quad \ell = 1.3^2 = 1.69$$

$$Q = \theta, \quad \theta^+ = \theta - 0.1 \cdot 2(\theta - y) = 1.76$$

教学简化：把Critic输出视为单一参数，目标值固定。真实训练对小批量取平均，并对网络参数做梯度下降。

10.5.2 MADDPG数据流与更新：策略梯度

Actor沿着提高本机价值的方向更新

$$J_i(\theta_i^{\mu}) = \mathbb{E}_{(s,o_i)\sim\mathcal{D}}[Q_i(s, a_1, \ldots, a_i, \ldots, a_N)] \tag{10.42}$$

$$\nabla_{\theta_i^{\mu}} J_i = \mathbb{E}[\nabla_{a_i} Q_i(s, \mathbf{a}) \nabla_{\theta_i^{\mu}} \mu_i(o_i; \theta_i^{\mu})] \tag{10.43}$$

$a_i = \mu_i(o_i; \theta_i^{\mu})$。更新本机Actor时，其他动作视为常量；梯度在本机策略输出处求值。

训练仍使用联合Critic的梯度；执行仅保留本地Actor。Critic估计不准时，策略改进方向也可能失准。

10.5.2 MADDPG数据流与更新：可复算梯度实验

固定解析Critic下的四次策略更新

$$Q_1 = -(a_1 - a_2)^2, \quad a_2 = 2, \quad a_1 = \theta o_1, \quad o_1 = 1$$

$$\theta_{k+1} = \theta_k + 0.1[-2(\theta_k - 2)]$$

更新次数	本机动作	解析价值	距最优动作的误差
0	0.0000	-4.0000	2.0000
1	0.4000	-2.5600	1.6000
2	0.7200	-1.6384	1.2800
3	0.9760	-1.0486	1.0240
4	1.1808	-0.6711	0.8192

假设：固定Critic与邻机动作、无动作饱和；数值由递推式直接计算。

10.5.2 MADDPG数据流与更新：软更新机制

目标网络缓慢跟随主网络

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \tag{10.44}$$

$$\tau = 0.01, \quad \theta = 2, \quad \theta' = 1 \quad \Rightarrow \quad \theta'^{+} = 1.01$$

初始化：主Actor、主Critic各有对应目标副本，目标副本通常先复制主网络参数。

取舍：小更新系数使目标平滑但滞后；系数及更新频率共同决定跟随速度。

10.5.2 MADDPG数据流与更新：关键训练过程

MADDPG的训练流程与评估

环节	输入、计算与产物
1 初始化与交互	首次复制目标网络；本地Actor加探索，形成联合动作
2 存储与采样	同步经验写入回放池，随机抽取整条联合转移
3 Critic与Actor	先算目标并拟合价值，再用动作梯度更新本机策略
4 软更新与评估	目标副本慢速跟随；冻结Actor后评估独立任务回合

执行输入在训练时就必须受到约束，不能靠部署时临时删掉信息。

本章目录

多智能体强化学习与规模扩展

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

10.5.1 马尔可夫博弈与学习架构

教材对应：第10.1.3.5节 · 主流框架与问题描述

10.5.2 MADDPG数据流与更新

教材对应：第10.1.3.5节 · 网络更新与训练过程

10.5.3 大规模集群扩展

教材对应：第10.1.3.6节 · 挑战与优化方案

10.5.4 方案选择与验证

教材对应：第10.1.3.5–10.1.3.6节 · 方案选择

10.5.3 大规模集群扩展：扩展瓶颈

节点增加分别扩大哪些开销

$$d_Q = N(d_o + d_a) = 15N, \quad d_o = 12, \quad d_a = 3$$

瓶颈	增长或机制	设计方向
联合Critic输入	10机150维；50机750维	共享编码、邻域聚合
多个独立Critic	每机一套网络放大参数与更新量	按角色共享，压缩联合表示
策略漂移与相关性	节点同时更新，群体经验强相关	评估旧策略影响，增加场景多样性

枚举联合离散动作时，$|\mathcal{A}| = m^N$体现组合爆炸。

10.5.3 大规模集群扩展：参数共享

参数共享保留个体输入差异

$$a_i = \mu(o_i, z_i; \theta), \quad i = 1, \ldots, N$$

共享对象：同构节点共用Actor参数；Critic是否共享取决于奖励、角色与输入结构。

个体差异：各机观测不同，角色编码不同，得到的动作仍可不同。

教学算例：每套Actor含10万参数，50套为500万；共用权重为10万，推理仍覆盖50个输入。

适用条件：观测与动作语义兼容；异构平台可先统一接口，或按角色使用不同网络。

10.5.3 大规模集群扩展：注意力与图结构

注意力与GNN选择和聚合邻居信息



注意力：按任务给邻居特征加权；权重表示模型中的相关性，不能直接解释因果。

图神经网络（GNN）：沿边聚合节点特征，允许邻居数量变化；效果仍取决于边、特征与训练分布。

部署区别：训练Critic可用联合图；执行Actor若使用图消息，必须满足实际通信与时效约束。

10.5.3 大规模集群扩展：图聚合算例

一次邻域聚合可以手算

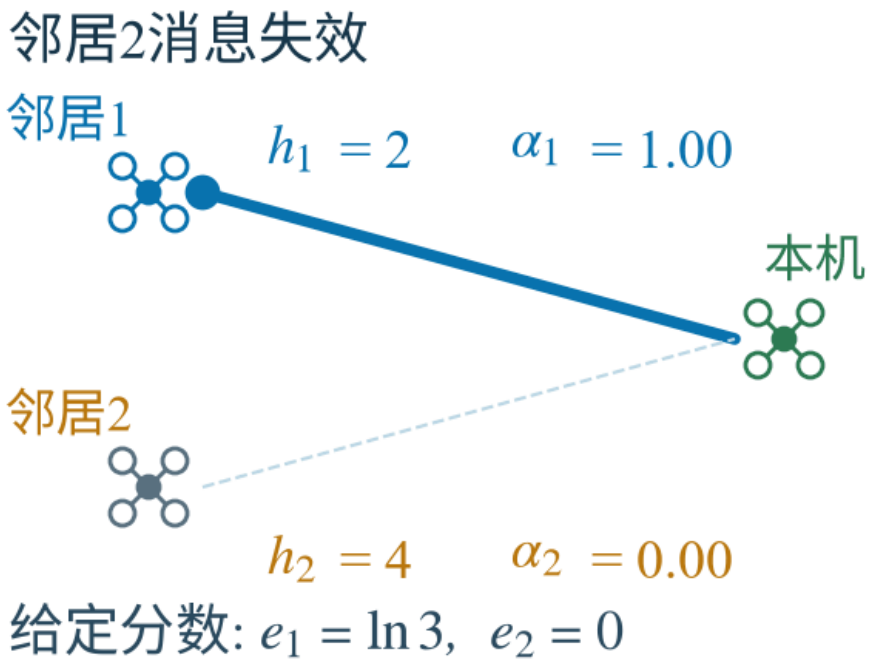
$$h_1 = 2, \quad h_2 = 4, \quad m_{\text{mean}} = (2+4)/2 = 3$$

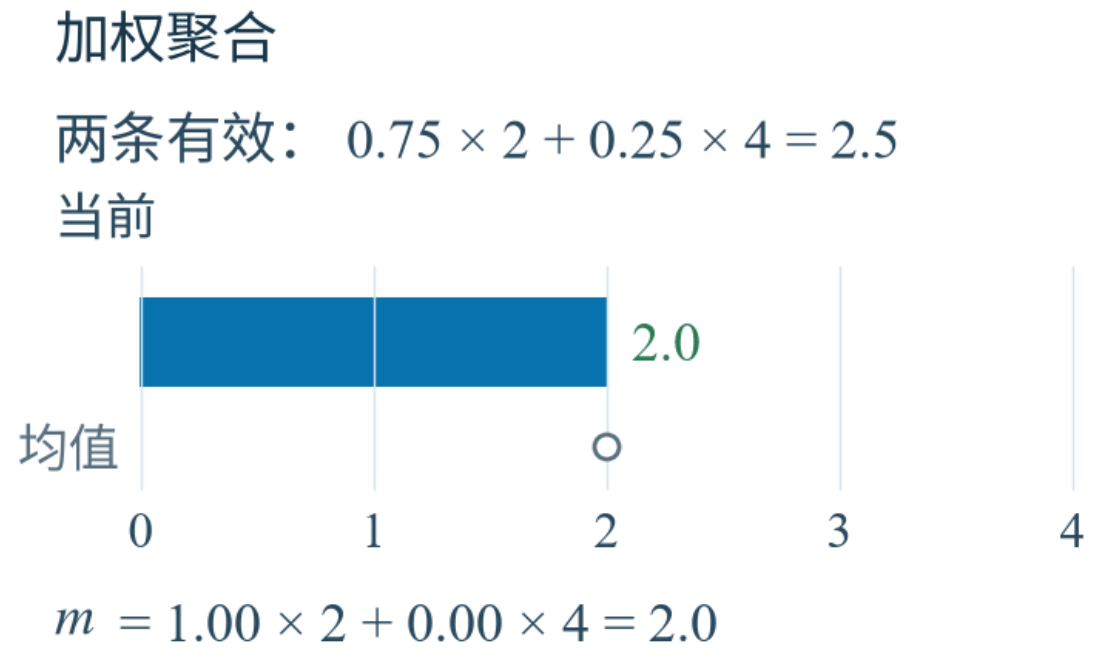
$$\alpha_1 = 0.75, \quad \alpha_2 = 0.25, \quad m = 0.75 \cdot 2 + 0.25 \cdot 4 = 2.5$$

消息状态	有效权重	聚合输出
两位邻居都有效	0.75，0.25	2.5
第二位邻居丢失	重新归一化为1，0	2.0
无有效邻居	使用本地状态及缺失标记	不能用旧消息假装新观测

10.5.3 大规模集群扩展：邻域注意力算例

消息失效后，按有效邻居重新归一化





图：给定分数经指数归一化为权重；连线中的圆点表示有效消息传递

两条消息有效时聚合为2.5，仅第一条有效时为2；权重来自给定分数。

10.5.3 大规模集群扩展：值函数分解

值分解把团队评价连接到局部选择

$$Q_{\text{tot}} = \sum_{i=1}^{N} Q_i(o_i, a_i) \quad \text{(VDN)}$$

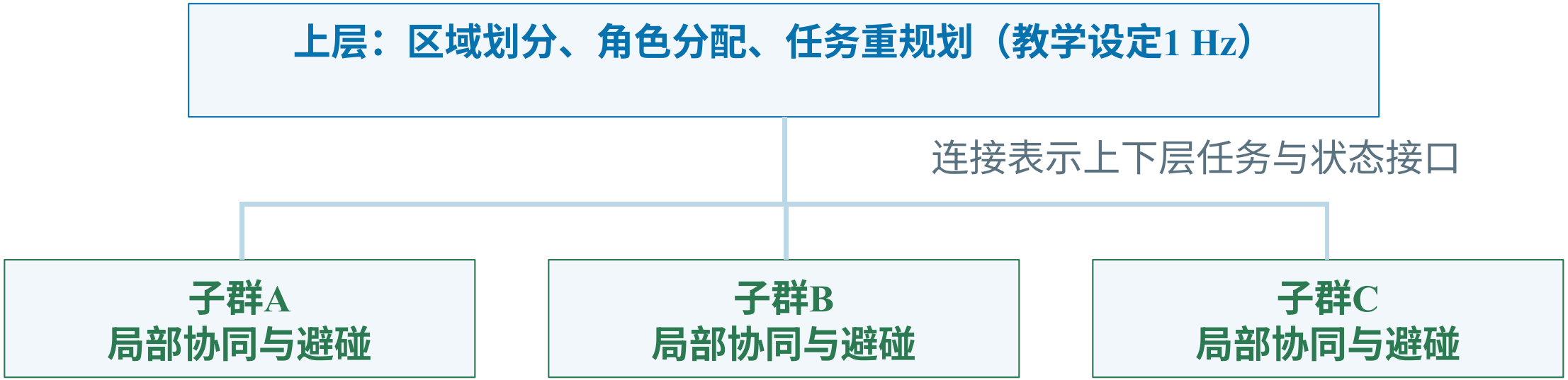
$$Q_{\text{tot}} = f_s(Q_1, \ldots, Q_N), \quad \frac{\partial Q_{\text{tot}}}{\partial Q_i} \geq 0 \quad \text{(QMIX)}$$

教学数值	二号选A：4	二号选B：2
一号选A：1	5	3
一号选B：3	7（各自贪心）	5

适用共同奖励与离散动作；非单调团队价值可能超出表示能力。

10.5.3 大规模集群扩展：层结构设计

分层把任务分配与快速控制解耦



下层教学设定20 Hz：依据本地观测生成动作。接口还需给出目标有效期、资源约束与失败反馈；上层任务变更不能破坏下层安全约束。

本章目录

多智能体强化学习与规模扩展

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

10.5.1 马尔可夫博弈与学习架构

教材对应：第10.1.3.5节 · 主流框架与问题描述

10.5.2 MADDPG数据流与更新

教材对应：第10.1.3.5节 · 网络更新与训练过程

10.5.3 大规模集群扩展

教材对应：第10.1.3.6节 · 挑战与优化方案

10.5.4 方案选择与验证

教材对应：第10.1.3.5–10.1.3.6节 · 方案选择

10.5.4 方案选择与验证：方案选择

按任务与资源选择组合方案

任务条件	候选结构	需要核验
连续动作、交互紧密	MADDPG与联合Critic	样本效率、价值估计与策略漂移
同构、局部交互	共享Actor + 稀疏GNN	角色表达、节点变化与消息缺失
共同奖励、离散动作	VDN/QMIX与参数共享	价值分解假设与联合行为表达
多子任务、不同时间尺度	分层分组，可配图聚合	跨层接口、资源冲突与失联降级

10机与50机是教材中的规模示例，选择依据是信息维数与资源约束。

10.5.4 方案选择与验证：验证与拓展

训练后的评估与辅助模型边界

验证指标：固定策略后测试任务成功率、全时段最小间距、能耗与推理截止时间。

泛化测试：改变节点数、初值、消息缺失与定位误差；记录性能下降与失败原因。

大语言模型（LLM）辅助：用于任务描述、数据扩充和上层分析；输出受资源、安全与时效约束。

平台实践：MARLExps（完整版及以上）为单机PPO与并行环境；协同训练需另建共享环境与多机学习器。

练习与讨论

课堂讨论：哪些信息能进入机载策略

讨论问题	可见参考答案
Critic用全体真值，Actor只用本地观测，是否符合CTDE?	符合。联合信息用于训练评价，部署Actor输入保持一致。
Actor训练依赖全体真值，部署删除真值输入，能否直接运行?	不能直接保证有效。需重设计输入并重新训练、评估。
从8机扩到60机，优先改什么?	先测输入维数、内存、通信和时延，再选共享、稀疏图或分层。

本节小结

本节小结：联合学习与本地执行

模型：联合动作影响转移，奖励表达合作或竞争；局部观测不必然构成完整状态。

MADDPG：Critic读联合信息，Actor输出本机动作；回放、目标网络与软更新闭合训练循环。

扩展：共享减少参数，GNN组织邻域，值分解用于合作，分层组织不同时间尺度。

易错点：CTDE不意味着执行集中控制，也不保证安全、收敛或跨规模泛化。