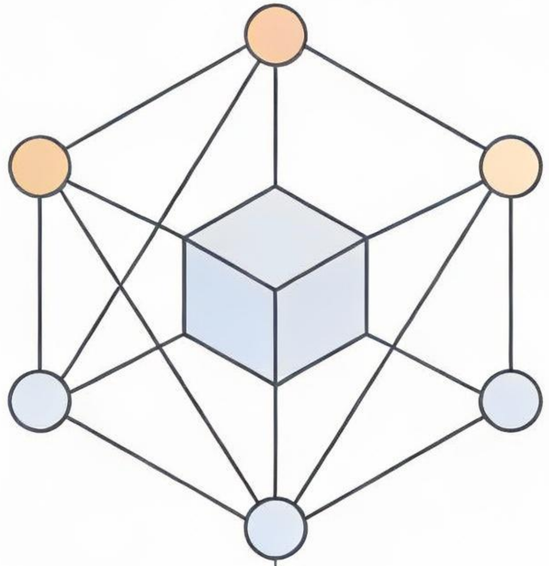
MORANDI

面向目标的传输调度：基于结构引导的统一双策略深度强化学习方法

- 核心创新：
  应用驱动目标转向
- 理论突破：
  单调性/渐近凸性证明
- 算法设计：
  SUDO-DRL统一双策略框架
- 实验验证：
  40设备20信道大规模场景收敛性



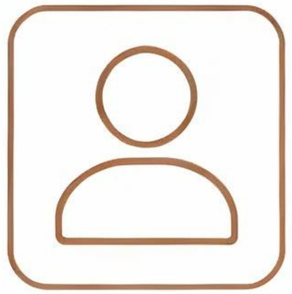
传统DRL：
40设备场景不收敛

SUDO-DRL：性能提升30%-40%

作者：Jiazheng Chen, Wanchun Liu*（通讯作者）

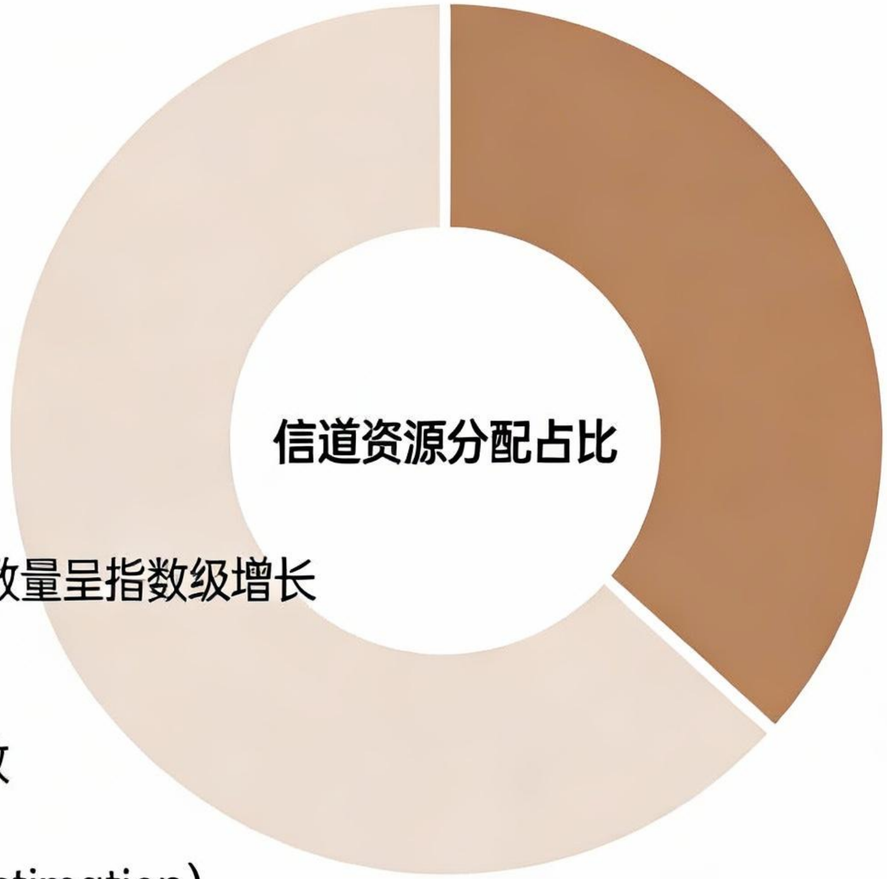
研究背景与问题定义：面向目标的传输调度

MORANDI



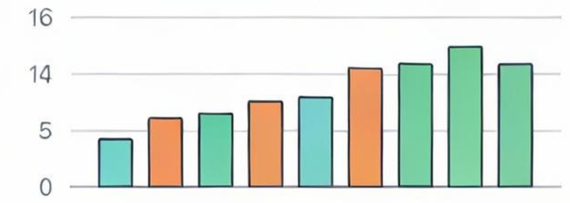
多设备多信道系统

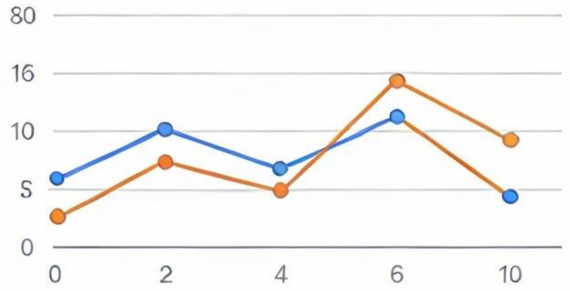
在多设备多信道无线通信系统中，N个边缘设备共享M个有限衰落信道（M < N）。研究的核心问题是寻找最优调度策略π，以最小化所有设备的长期期望总成本。在远程状态估计场景下，这等同于最小化与信息年龄（AoI）相关的估计状态均方误差（MSE）

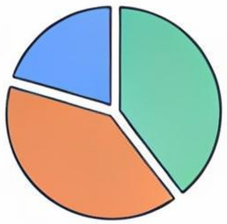


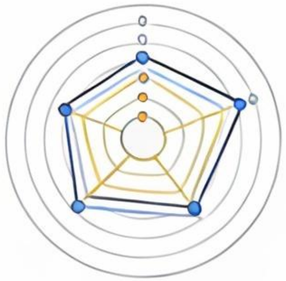
- **高维状态空间**：必须实时跟踪每个设备的AoI和信道状态，复杂度随设备数量呈指数级增长
- **庞大的动作空间**：对于N设备M信道，动作数为N! / (N-M)!。
  在10设备5信道场景下，动作数已达30,240个，传统MDP求解器完全失效
- **目标转向**：从单纯的传输比特转向满足特定应用需求（Remote State Estimation）

现有方法的局限性与研究空白









1. 传统方法：

启发式方法（如Whittle”s Index）虽快但非最优；动态规划（值迭代/策略迭代）面临“维度灾难”

2. 标准DRL算法：

- 离策（Off-Policy）如DQN/DDPG：虽样本效率高，但在大型动动态系统中训练极不稳定，40设备场景下无法收敛-
- 同策（On-Policy）

如PPO/TRPO：稳定性好但样本效率极低，数据利用率差，易陷入局部最优

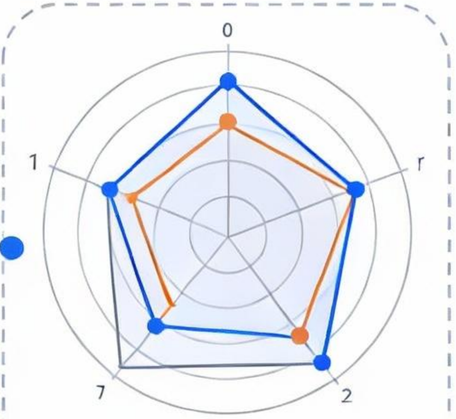
3. 理论结合缺失：

现有研究多采用“暴力优化”，未深入挖掘Aol状态最优值函数的数学性质（如单调性、凸性），导致在大规模系统（20-40个传感器）中性能严重下降或不收敛

4/12 SUDO-DRL框架概述：理论与算法的统一

核心思想：结合PPO在线策略稳定性与SAC离线策略效率，用理论结构属性指导训练

1. **理论属性推导**：证明V函数单调性与渐近凸性
2. **结构评分机制**：转化为CM（单调性）、CC（凸性）、AM（策略单调性）评分
3. **统一双损失函数**：通过α, β_1, β_2桥接在线与离线更新
4. **结构引导回放池**：基于评分选择性存储与优先级采样



结构评分分布

理论基础：最优策略的结构属性证明

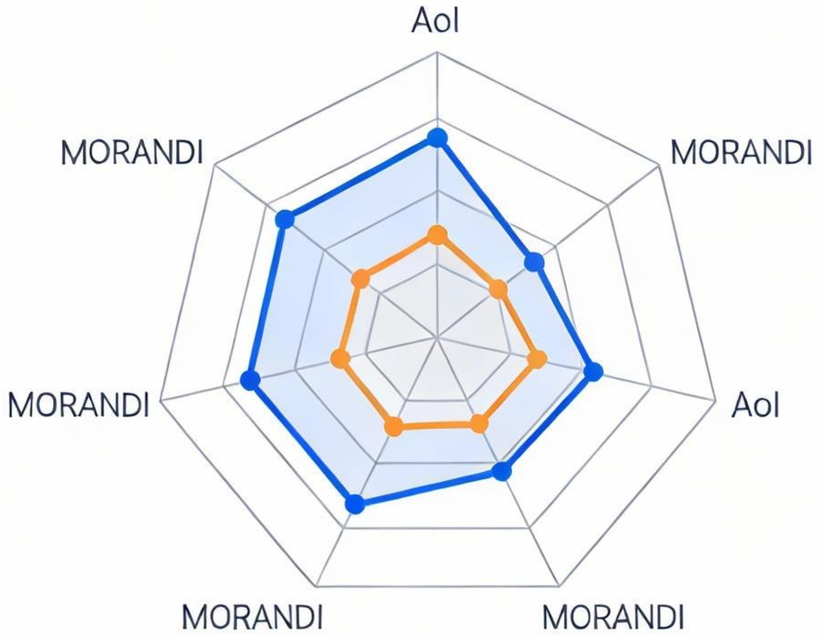
1.V函数单调性：价值函数随AoI和信道状态增加单调递增

2.渐近凸性：传输调度领域首次证明V函数随AoI状态渐近凸性

3.策略单调性：最优策略关于信道状态具有单调性

4.渐近贪婪结构：共址设备AoI极大时最优策略倾向强制调度（预训练理论依据）

结构属性量化评分



目标函数：

$\min \pi \lim T \to \infty (E\pi[\perp Tt=1 \sum n=1N \gamma tcn(\delta n,t)]$

（$\delta n,t$为设备n在时间t的AoI）

结构属性评估框架与量化指标

MORANDI

数学证明

CM Score (Critic-Monotonicity)

评估Critic网络预测的V值是否随AoI增加，若违反则产生惩罚项：

$$\check{y}_{AoI} = \max(0, v(s; v) - v(\hat{s}_{(n)}; v))$$

CC Score (Critic-Convexity)

评估V函数的二阶差分是否符合凸性：

$$\tilde{c}_{AoI} = \max(0, 2v(s; v) - (v(\hat{s}_{(n)}; v) + v(\hat{s}_{(n)}; v)))$$

AM Score (Actor-Monotonicity)

评估Actor网络的动作选择是否符合信道单调性逻辑，评分决定数据是否进入回放池及采样优先级

统一双损失函数

1. Critic统一损失:

$$L_{SUDO}(v) = L_{On}(v) + \beta_1 L_{Off}(v) + \text{Structural Penalties}$$

其中L_On处理当前轨迹TD误差，L_Off处理回放池历史数据TD误差

2. Actor统一损失:

$$L_{SUDO}(\phi) = L_{On}(\phi) + \beta_2 L_{Off}(\phi)$$

结合PPO裁剪损失（稳定性）与SAC风格离线策略梯度（效率），解决传统离策方法大规模调度发散问题

公式推导

回放池管理与结构引导预训练

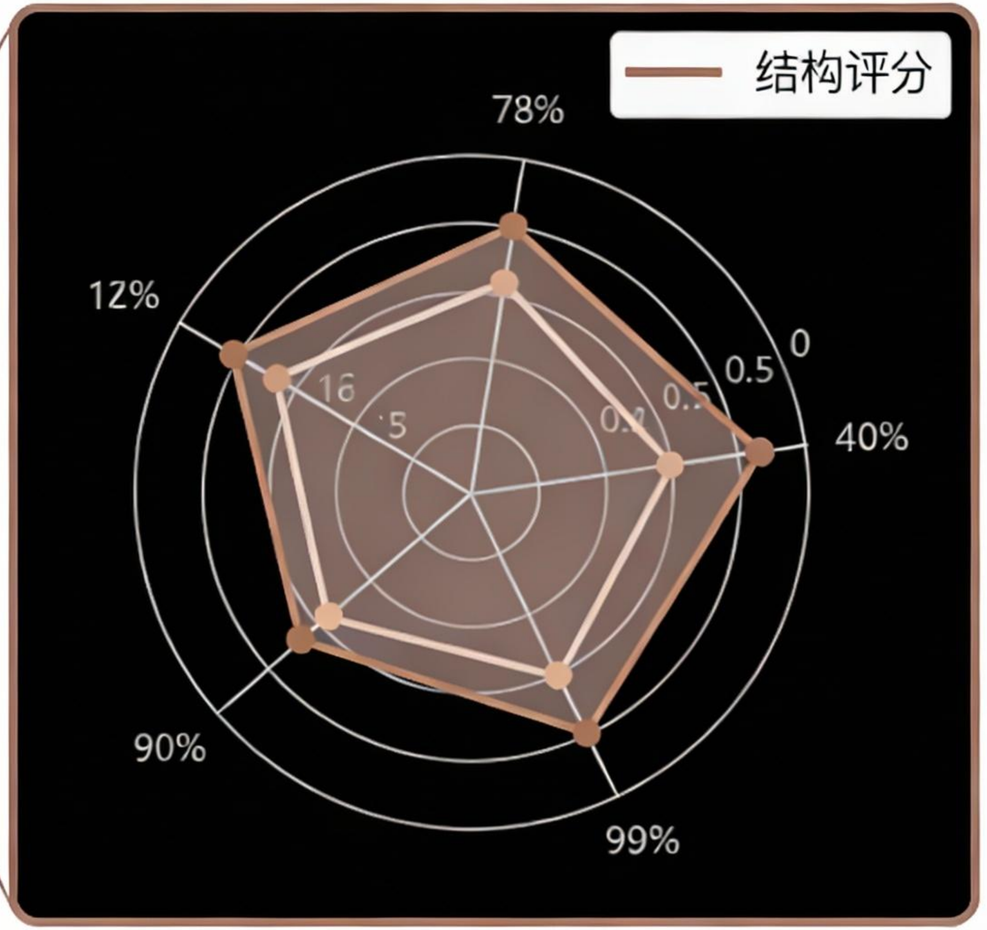
MORANDI

1. 选择性存储：
仅存储结构评分（CM, CC, AM）超历史平均的轨迹，确保回放池为高质量样本

2. 优先级采样：
采样概率Pb由结构评分和新近度决定，公式为Pb

$$Pb \approx (pb \geq varrho^{\wedge}b) / \sum (pb \geq varrho^{\wedge}b)$$

3. 渐近贪婪预训练：
前10 × N 回合用定理 5贪婪结构指导动作，收敛速度提升约40%



实验设置：远程状态估计仿真

MORANDI

规模覆盖：
从10设备5信道扩展至40设备20信道（超大规模）

信道模型：
i.i.d.块衰落信道，量化为5级状态，丢包率范围0.01至0.2

神经网络：
Actor和Critic均为3层隐藏层的全连接网络

训练参数：
10,000训练回合，每回合128步；
γ=0.99，学习率Critic=0.001, Actor=0.0001

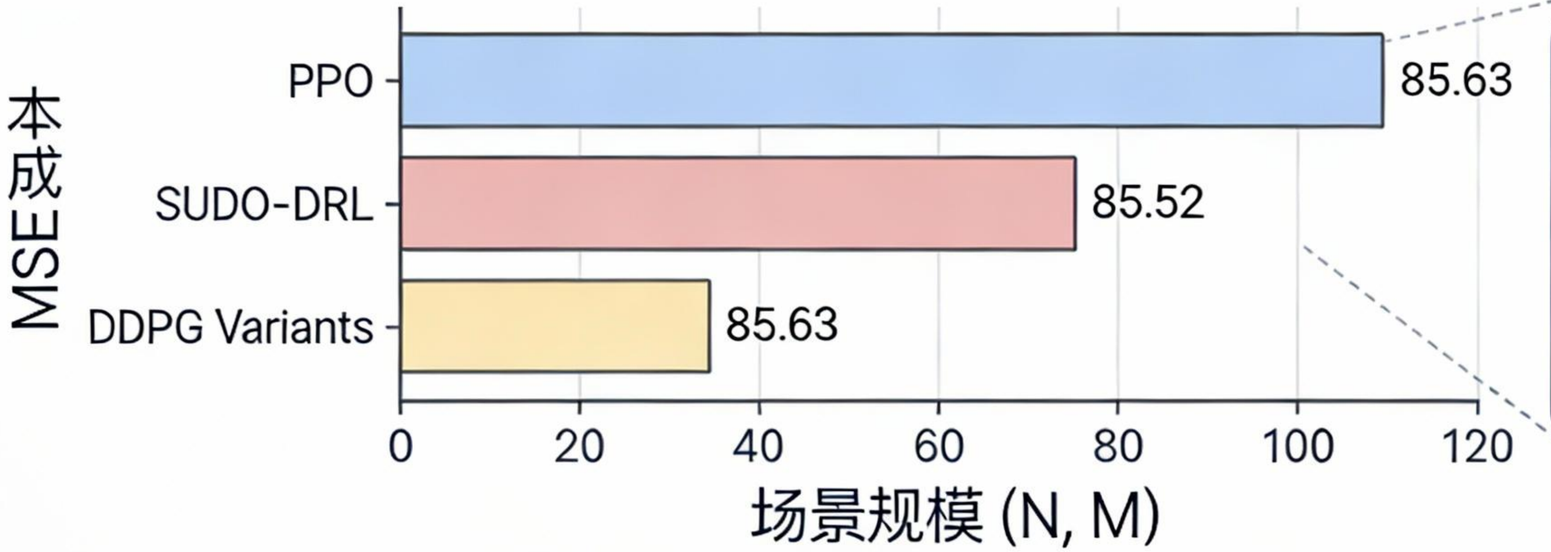
硬件支持： NVIDIA RTX 3060Ti GPU。
评估指标侧重于平均总MSE成本和收敛速度

Hyperparameters	Value
Discount factor, γ	0.99
Clipping parameter, ε	0.2
Clipping parameter, ε	0.9
Unified loss weight, β1, β2	0.9
Replay buffer, R	200
Pre-training episodes	10 × N

性能对比：SUDO-DRL的卓越表现

1. 小规模场景 (10, 5)：SUDO-DRL成本为85.52，接近专门优化的SE-DDPG（77.14），远优于PPO（119.63）。
2. 大规模场景 (30, 15) 及以上：所有离策算法（DDPG变体）均宣告失败（无法收敛）。

   SUDO-DRL不仅成功收敛，且性能比PPO提升了30%-40%。
3. 鲁棒性：在不同参数设置下（Para. 1-16），SUDO-DRL始终保持最低的MSE成本，，证明了其在复杂、动态无线环境下的稳健性。



Doc Table 2

Scale (N, M)	PPO	SUDO-DRL	DDPG Variants
(10, 5)	119.63	85.52	77.14-89.26
(20, 10)	569.96	370.63	357.14 (Partial Fail)
(30, 15)	900.71	518.03	Failed (—)
(40, 20)	1291.54	994.80	Failed (—)

消融研究与收敛效率分析

1 结构引导的作用：
SUDO-DRL在200回合内达到100分凸性评分（CC Score），PPO徘徊80分以下，结构引导强制模型学习正确物理规律

2 预训练的价值：
全功能版本比无预训练版本收敛速度快40%，预训练使初始MSE成本从800+降至700以下

3 可扩展性突破：SUDO-DRL是首个在40设备20信道规模下保持高性能且稳定收敛的DRL调度算法，填补大规模目标导向通信技术空白

200
100

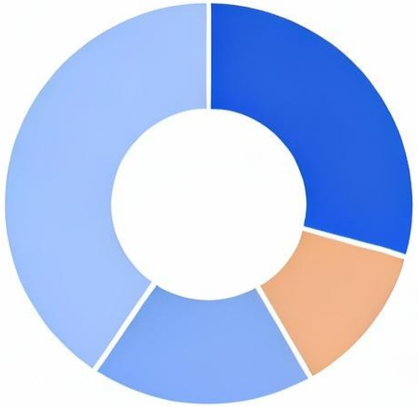
80
80

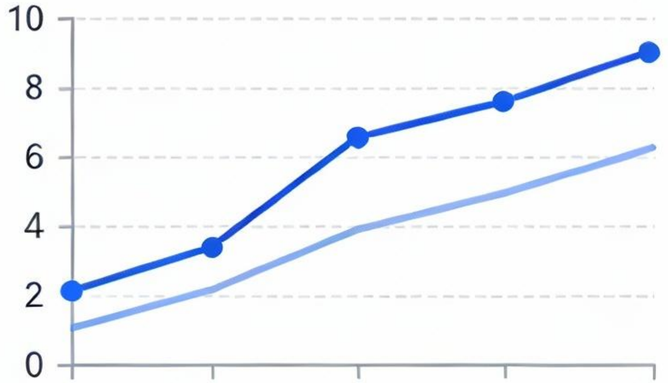
40%

800+

700
40设备20信道

SUDO-DRL与PPO的CC Score收敛情况





有无预训练的MSE成本对比

结论与贡献总结

MORANDI

1. **理论突破**：首次证明AoI调度中价值函数的渐近凸性，利用单调性构建约束学习框架

2. **算法创新**：提出统一双策架构，兼顾PPO稳定性与SAC效率

3. **性能飞跃**：相比PPO，系统性能提升25%-45%，收敛速度提升40%

4. **大规模适用性**：解决40台设备维度灾难，实现理论最优与计算可行平衡

5. **未来研究**：探索多跳网络或异构业务流应用