文章编号:1671-4598(2026)02-0251-07 DOI:10.16526/j.cnki.11-4762/tp.2026.02.033 中图分类号:TP393 文献标识码:A

干扰条件下基于 MADRL 的多无人机动态信道决策算法

刘正龙，郭肃丽
（中国电子科技集团公司 第五十四研究所，石家庄 050081）

摘要：无人机以其高效和低成本的优势逐渐成为侦查打击、抗震救灾等任务的核心力量，然而在执行任务过程中，无人机网络面临着外部的干扰，其抗干扰能力的强弱直接关系到任务的成败；针对动态干扰环境下的无人机信道接入问题，将问题的求解与集中式训练分布式执行框架融合，提出了一种基于价值分解网络的多智能体无人机动态信道决策算法；将深度强化学习中的固定学习率改进为余弦退火学习率，从而获得更低的损失和更高的模型精度；仿真实验结果表明所提算法能使网络节点智能调整信道选择策略，最大化无线网络的吞吐量，余弦退火算法提升了强化学习的训练效果。

关键词：无人机网络；干扰对抗；动态频谱接入；价值分解网络；余弦退火学习率

A Dynamic Channel Selection Algorithm of UAVs Based on MADRL Under the Interference Environment

LIU Zhenglong, GUO Suli
(The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China)

Abstract: Unmanned aerial vehicles have gradually emerged as a core force in missions such as reconnaissance and strike operations, as well as earthquake relief, thanks to their advantages of high efficiency and low cost. However, UAV networks are exposed to external interference during mission execution, and the strength of their anti-interference capability is directly related to the success or failure of the missions. To address the problem of UAV channel access in a dynamic interference environment, this paper integrates problem solving with a centralized training and distributed execution framework, and proposes a multi-agent UAV dynamic channel decision-making algorithm based on the value decomposition network. In addition, the fixed learning rate in deep reinforcement learning is improved to a cosine annealing learning rate, thereby achieving lower losses and higher model accuracy. Simulation results show that the proposed algorithm enables network nodes to intelligently adjust channel selection strategies and maximize the throughput of wireless networks, while the cosine annealing algorithm enhances the training effect of reinforcement learning.

Keywords: unmanned aerial vehicle network; anti-jamming, dynamic spectrum access; value-decomposition networks; cosine annealing learning rate

0 引言

随着无人技术的快速发展，无人机因具有易于部署、机动性强等优势，成为当前军事和民用领域中的研究热点。在军事行动的侦查打击、民用的荒野搜救以及户外表演等场景中展现出了巨大的应用价值[1]。近年来，随着无线通信网络的快速发展，接入通信网络的无线设备数量急剧增长，对无人机通信性能造成了严重影响；另一方面，由于无线网络的开放特性，无人机网络容易受到恶意用户的干扰或欺骗，这对无人机网络的安

收稿日期:2025-12-08； **修回日期**:2026-01-16。
作者简介:刘正龙(2000-),男,硕士研究生。
引用格式:刘正龙,郭肃丽.干扰条件下基于 MADRL 的多无人机动态信道决策算法[J].计算机测量与控制,2026,34(2):251-257.

全性造成了严重威胁。因此，研究干扰条件下的多无人机动态信道决策技术以有效避免外部复杂多样的干扰，进而实现无人机之间可靠的信息传输对无人机网络通信具有重大意义[2]。

动态频谱接入（DSA，dynamic spectrum access）技术[3]是一种先进的频谱管理技术。它允许认知用户实时探测并使用无线通信网络中未被干扰的信道。传统的动态频谱接入技术（如博弈论和拍卖理论[4]、机器学习等方法）通常是基于模型实现的，在面对实时变化的无线通信网络环境时适用性较差。由于深度强化学习[5]算法可以根据无人机用户与通信网络环境的不断交互来动态调整频谱接入策略，因此将深度强化学习方法用于解决干扰条件下的多无人机信道决策问题，具有灵活性高、通用性强的特点[6]。

基于强化学习的信道决策方案无需提前获得无线通信环境的先验信息，仅根据与当前信道环境交互，即可学习信道决策方案[7]。文献［8］提出一种基于值函数学习的无人机信道决策方法，通过直接评估各状态－动作对所对应的 Q 值来指导无人机智能体做出信道决策。文献［9］提出一种基于深度 Q 网络（DQN，deep Q-network）的信道决策方法，该方法可以在大部分信道先验信息缺失的情况下，与环境互动学习，最终获得一个收敛的信道决策策略。对于无线通信网络中的多无人机信道决策问题，文献［10］开发了一种基于深度 Q 学习的多智能体强化学习框架，仿真结果显示该算法能够在信息开销和系统性能之间取得平衡，提高了多无人机网络的信道决策效率。文献［11］研究了时变无线通信环境下多无人机的联合频谱感知和信道接入问题，提出了一种双重深度 Q 网络（DDQN，double deep Q-network）[12]算法，实验结果表明该算法可以有效提高系统的平均奖励和信道利用率。文献［13］提出了一种基于多智能体近端策略优化算法（MAPPO，multi-agent proximal policy optimization）[14]的多无人机频谱决策方法，仿真结果表明该方法能够有效提升用户传输速率，并保证了用户通信的公平性，但该方法未关注外部恶意干扰的影响。

综上，对于在干扰条件下的多无人机信道决策问题，本文提出了一种基于多智能体强化学习（MADRL，multi-agent deep reinforcement learning）的多无人机动态信道决策算法，该算法利用价值分解网络（VDN，Value-decomposition networks for cooperative multi-agent learning）[15]解决了多智能体的信用分配问题，并利用集中式训练分布式执行（CTDE，centralized training with decentralized execution）框架提升了多智能体的协作能力。在该算法的基础上，提出了余弦退火学习率衰减算法，增强了干扰环境下无人机信道的接入效率。本文的主要贡献如下：

1）提出了一种基于 MADRL 的多无人机动态信道决策算法，首先将多个无人机的动态信道决策过程建模为去中心化的部分可观测马尔科夫过程[16]，其次通过价值分解算法建立单个智能体动作与系统整体回报之间的关联，有效地避免了接入频谱时用户之间的冲突。

2）使用余弦退火学习率衰减算法，避免了固定学习率或步进衰减可能造成的优化过程震荡和不稳定。通过使用余弦函数模拟学习率的衰减过程，增强了训练结果的泛化能力。

1 问题建模

考虑一个干扰机和多个无人机用户共存的多通道无线通信网络系统，如图 1 所示。

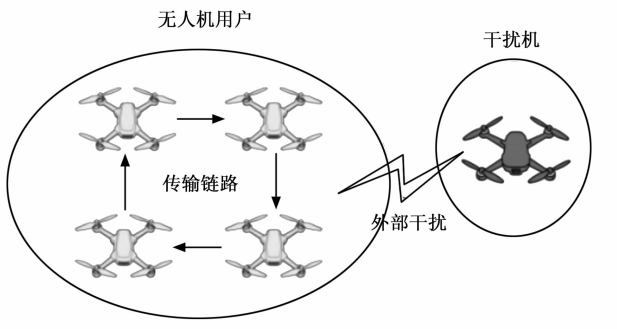

图 1 干扰条件下无人机网络模型示意图

无人机用户在尝试频谱接入操作之前，需要通过感知来获取通信环境的状态信息。本文通过计算信干噪比来判断通信环境中各个信道的占用状态。当信道被其它无人机用户占用时，接收信号可以表示为：

$$y(t)=\sum_{i=1}^{I}\sqrt{P_i(t)}h_i(t)x_i(t)+\omega(t) \tag{1}$$

其中：$P_i(t)$ 表示第 i 个无人机用户的发射功率；$h_i(t)$ 表示当前用户与第 i 个无人机用户之间的信道衰落因子；$x_i(t)$ 为当前用户的发送信号；$\omega(t)$ 为白噪声，其均值为零，方差为 σ^2。

假设在该通信场景中共存在 M 条正交信道，干扰机在每个时隙随机选择 $H(H\leqslant M)$ 条信道进行干扰，可同时干扰最多 M 条正交信道。同时存在 N 个无人机用户，通过随机接入协议共享网络中的 M 条正交信道。在无线通信网络场景中，参考维纳 II 信道模型计算路径损耗：

$$PL=\overline{PL}+A_W\log_{10}(d)+B_W\log_{10}(f_c/5) \tag{2}$$

其中：$\overline{PL}$ 表示参考距离下的路径损失；A_W 表示路径损失因子；B_W 表示路径损失频率依赖系数；d 表示信号传播距离（m）；f_c 表示载波频率（GHz）。用莱斯

信道模型来建模信道[18]：

$$h=\sqrt{\frac{\delta}{\delta+1}}\beta e^{j\theta}+\sqrt{\frac{1}{\delta+1}}CN(0,\beta^2) \tag{3}$$

其中：δ 为视距路径与散射路径之间的接收功率比；θ 是视距路径的信号到达相位；$CN(\cdot)$ 表示高斯随机变量；$\beta^2=10^{-PL/10}$。无人机用户 n 的信干噪比可以表示为：

$$SINR_n=\frac{P_n(t)\left|H_{nn}(t)\right|^2}{P_j(t)\left|H_{jn}(t)\right|^2+\sum_{l=1,l\neq n}^{N}P_l(t)\left|H_{ln}(t)\right|^2+BN_0} \tag{4}$$

其中：$P_n(t)$ 和 $P_j(t)$ 分别表示无人机用户 n 和干扰机 j 的发射功率；B 表示信道带宽；N_0 表示噪声功率谱密度。

设 γ_{th} 为无人机用户在当前通信系统中正常通信所需的 $SINR$ 阈值。在无人机用户 n 尝试接入信道前，随机选择一条信道发射感知信号，并依据公式（4）计算当前的 $SINR_n$，当 $SINR_n>\gamma_{th}$ 时，认为当前信道为空闲状态，可以进行接入；当 $SINR_n\leqslant\gamma_{th}$ 时，认为当前信道存在干扰或者已被其它无人机用户占用，存在接入冲突。

2　基于价值分解网络的 MADRL 算法

多无人机动态信道决策问题是典型的多智能体协作问题[17]，其核心在于多个无人机用户（智能体）需要共享有限的频谱资源，以最大化通信系统的总容量，最小化干扰。多智能体强化学习算法一般采用中心化训练与分布式执行框架，在训练阶段，所有无人机观察通信环境并做出信道接入动作，将得到的奖励和观测动作历史以元组的形式存放在经验回放池中，通过抽取样本更新网络参数的方式进行训练，在训练完成后，每个智能体将学习到的接入策略部署到网络中，仅根据自己的局部观察即可独立选择接入策略。

基于 VDN 的强化学习算法的核心思想是将系统的联合行动值函数分解为单个智能体函数的线性组合，解决了单个智能体在全局视野缺失下的信用分配难题。VDN 算法的理论来源是马尔科夫博弈模型[19]。通常用元组来表示：(S,A,P,R)，其中，S 表示当前智能体所处的环境状态；A 表示智能体的动作选择空间；P 表示在智能体执行动作后，当前环境状态的转移概率；R 表示动作的累计奖励值。在每一个时间步 t，每一个智能体 n 会根据其局部观察 o_n^t 选择一个行动 a_n^t，所有智能体的联合行动共同导致环境的状态转移，并产生该动作的奖励值 $r_n(s,a)$。将多智能体系统总成功次数作为奖励。系统的总回报可以定义为：

$$R^{\mathrm{tot}}=\sum_{t=1}^{T}\gamma^{t-1}\sum_{n=1}^{N}r_t^n(o^n,a) \tag{5}$$

在协作式多智能体强化学习问题中，每个智能体基于自己的局部观测选择动作，得到奖励并以此优化策略。VDN 算法的目标是学习一个最优的联合动作值函数 $Q^{\mathrm{tot}}(\tau,a)$，该函数输入所有智能体的历史行动轨迹 τ 和联合动作，输出执行该联合动作的期望累计回报，将单个智能体的动作－价值函数定义为 Q 值函数。首先，通过每个智能体的 Q 网络输入各自的观测，得到该智能体对应所有可选动作的 Q 值分布；然后，根据每个智能体在当前样本中实际选择的动作，从各自的 Q 值分布中索引提取对应动作的单个 Q 值；最后，对所有智能体提取出的单个 Q 值进行逐样本直接求和，得到最终的联合动作值函数，完成线性组合。即：

$$Q^{\mathrm{tot}}=\sum_{n=1}^{N}Q_n(\tau_n,a_n;\theta_n) \tag{6}$$

其中：Q_n 表示单个智能体 n 的 Q 值函数，由参数为 θ_n 的神经网络进行近似，即由单智能体神经网络近似单智能体 Q 值，混合神经网络近似联合 Q 值。上式表明了联合 Q 值函数被明确归因于每个个体的贡献，每个智能体通过最大化自身的 Q_n，共同最大化全局 Q 值函数。

VDN 算法采用经验回放池将与环境交互过程中产生的单步经验元组以结构化数据元组的形式存储在回放记忆单元中。在神经网络的训练阶段，从回放池的所有已存储样本中，无放回地随机抽取固定批量大小的样本，确保每个样本被抽取到的概率均等，从而有效地提升了算法的稳定性。另一方面，由于该算法在更新网络参数时使用梯度下降法要求训练样本之间不相关，经验回放机制有效降低了训练样本之间的相关性。经验回放机制提高了算法的稳定性和收敛性，提升了样本数据的利用率。

VDN 采用中心式训练、分布式执行模式，该算法网络模型如图 2 所示。

在实际训练过程中，每个智能体依据 ξ－贪婪策略选择动作，得到相应的 Q 值。在集中训练阶段，利用全局信息学习联合策略；在执行阶段，每个智能体依据自己的局部观测和联合策略选择动作，保证了全局最优解的训练策略和执行阶段的高效性。

在训练过程中，VDN 的损失函数与传统的深度 Q 学习网络类似，基于误差来更新 Q 值。对于给定的经验样本，损失函数为：

$$L=E\{[r+\gamma\max_{a'}Q_{\mathrm{tot}}(s',a')-Q_{\mathrm{tot}}(s,a)]^2\} \tag{7}$$

其中：r 是环境给出的全局回报；γ 是折扣因子；s' 是下一个状态；a' 是下一个状态下的最优联合动作。通

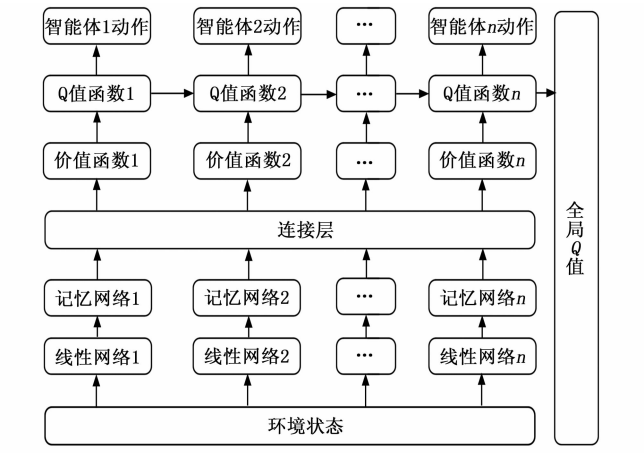

图 2 VDN 强化学习算法网络模型

过随机梯度下降法优化损失函数，梯度会通过加和操作分配给每个智能体网络。这表明如果全局获得一个较高的奖励值，那么所有参与协作的智能体网络的 Q 值都会得到提升；反之，如果全局获得一个较低的奖励值，所有的单个智能体 Q 值都会相应下降。通过足够的环境交互和策略探索，智能体最终会学习到使全局 Q 值收益最高的行动策略。

3 干扰条件下的多无人机动态信道决策算法

针对干扰条件下的多无人机信道决策问题，本文提出了一种基于 MADRL 的动态信道决策算法。由上一节可知，该算法首先需要将问题建模为去中心化的部分可观测马尔科夫过程[20]，并结合问题建模重新定义元组中的部分元素。其中，N 定义为无人机用户；S 表示无人机所处的通信网络环境状态；状态空间、动作空间及奖励函数设计如下。

3.1 状态空间

无人机用户 n 在时隙 t 感知到的信道状态可以表示为：

$$s_n(t)=\left[s_n^1(t),s_n^2(t),\cdots,s_n^m(t),\cdots,s_n^M(t)\right]$$

其中：$s_n^m(t)$ 表示无人机用户 n 在时隙 t 对信道 m 的频谱感知结果。对于通信网络每一条信道，由于受到外部干扰机的影响，在每个时隙开始时，信道的状态以一定概率转换或者不变并在一个时隙内保持该状态不变，其状态转换服从 2 状态马尔科夫链模型。在每一个时隙 t，无人机用户 n 从 M 条正交信道中选取 1 条发送感知信号，当无人机用户 n 计算得到 $SINR_n>\gamma_{th}$ 时，令 $s_n^m=1$，表示在时隙 t 信道 m 可接入；当 $SINR_n>\gamma_{th}$ 时，令 $s_n^m=0$，表示在时隙 t 信道 m 被干扰。

3.2 动作空间

在每一个时隙 t，无人机用户 n 的动作可以表示为：

$$a_n(t)=\left[a_n^1(t),a_n^2(t),\cdots,a_n^m(t),\cdots,a_n^M(t)\right]$$

其中：$a_n^m(t)$，表示无人机用户 n 在时隙 t 时是否接入信道 m。无人机用户的动作空间由其观测得到的信道状态空间决定。当 $a_n^m=1$ 时，表示无人机在时隙 t 选择信道 m 进行接入；当 $a_n^m=0$ 时，表示无人机在时隙 t 感知到信道 m 被干扰，选择不进行接入。

3.3 奖励函数

奖励函数用于评价各个智能体对于当前频谱接入策略性能优劣程度。在信道感知阶段，通过测量无人机用户接收机反馈的信干噪比，判断是否存在接入冲突。用户在时隙末获得的奖励设计为：

$$r_t^n=\begin{cases}+1,\text{所有用户接入不同信道}\\+0.5,\text{部分用户接入信道，另一部分未接入}\\0,\text{所有用户均未接入信道}\\-1,\text{用户接入信道但产生冲突}\end{cases}\tag{8}$$

通过每步的低奖励值引导智能体维持有效状态，加速收敛，解决稀疏奖励下智能体“探索无方向”的问题，同时可避免因吞吐量数值过大导致的梯度爆炸。

在多智能体环境中，由于状态空间庞大且智能体间相互影响，稀疏奖励问题尤为突出。分段奖励通过设置中间奖励信号，为智能体提供了更密集的反馈指导。例如，除了最终的系统性能奖励外，还可以为每个时间步的无冲突接入、公平的信道分配等行为设置即时奖励。这种设计显著改善了信用分配问题，帮助智能体更准确地理解其具体行为与长期回报之间的因果关系，加速学习过程。对于单个无人机用户，其目标是建立在单个回合内累计奖励值最大的接入策略 π^*：

$$\pi^*=\arg\max_{\pi}E\left[\sum_{t=1}^{T}\gamma^{t-1}r_t^n(o_t^n,a_t^n)\mid o_1^n\right]\tag{9}$$

其中：$0\leqslant\gamma\leqslant1$ 是折扣因子，表示动作对回报的长期影响。由于需要智能体维持长期稳定的信道状态，引导智能体选择“长期受益”的动作，故本文中将 γ 值设置为 0.99。

3.4 余弦退火学习率衰减算法

在深度强化学习的训练过程中，固定学习率易使训练结果陷入局部最优值。余弦退火衰减策略使学习率按照周期变化，具体定义为：

$$\eta_t=\eta_{\min}+\frac{1}{2}(\eta_{\max}-\eta_{\min})\left[1+\cos\left(\frac{T_{\text{cur}}}{T_{\max}}\pi\right)\right]\tag{10}$$

其中：η_t 表示在时间步 t 的学习率；$\eta_{\max}$ 为学习率衰减周期的起始值；$\eta_{\min}$ 为学习率衰减的终点值；T_{cur} 为当前退火周期的位置；$T_{\max}$ 为最大周期长度，决定了学习率衰减的速度，当其等于总训练轮数时，则学习率在整个训练过程中从 $\eta_{\max}$ 平滑衰减至 $\eta_{\min}$，当其设置为一个较小的值时，则学习率会周期性重置，有助于模型训练

结果跳出局部最优。

余弦函数导函数的绝对值先增大后减小，对应余弦函数的斜率先较水平，随后迅速降低，最后缓慢回到水平，这样的变化正好满足深度学习对学习率权重衰减的要求。学习率先保持一个较大的值缓慢下降，保证损失函数可以快速接近局部最优解，在接近最优解之后学习率权重迅速减小，开始缓慢接近最优解，最后保持一个较小的值使损失函数收敛于最优解附近。同时，余弦函数本身是一个周期函数，因此可以用单个数学表达式去表现出每一段区间上学习率的变化，降低了算法的复杂性，节约了更多的内存给训练。

3.5　算法流程

多智能体分布式协作频谱接入算法：

初始化通信环境，设置信道数 M 和用户数 N，训练总回合数、周期长度、学习率和学习率等参数

for episode = 1 to I do：

for n= 1 to N do：

初始化回放经验池 D

观测当前时刻状态，并根据贪婪策略选择动作

观测下一时刻状态，并计算奖励值，得到训练样本，储存到回放经验池 D

end

for step = 1 to T do：

从回放经验池中抽取回合训练

计算损失值 L

更新计算网络参数 θ

end

用当前估值网络的参数更新目标网络的参数

end

输出 Q 值并保存模型

4　仿真与分析

实验场景设置为一个拥有 12 条正交信道的带宽相等的无线通信网络，且信道状态转换服从二状态马尔科夫链模型，干扰机在每个时隙随机选择 H（$H\leqslant 12$）条信道进行干扰。同时存在 N 个无人机用户，通过随机接入协议共享网络中的 12 条正交信道，每个无人机用户在尝试频谱接入操作之前，随机选取一条信道进行感知，获取通信环境的状态信息，当感知到当前信道被干扰时，选择不接入；当信道未被干扰时时，选择该信道进行接入。设置 $\gamma_{\text{th}}=10$ dB。探索概率在 400 步内从 0.4 衰减到 0.05；设置余弦退火学习率的衰减周期起始值为 0.005，衰减终点值为 0.000 1，衰减最大周期等于训练总轮数，训练总轮数设置为 200。分别在无人机用户数 N=4，8 的情况下进行对比实验。

在仿真中，选择以下对比实验方案：

1）随机策略：无人机用户随机接入信道。

2）静态分配策略：无人机用户无视干扰，按照固定分配接入信道。

3）固定学习率方案：在其他条件相同的情况下，使用固定学习率进行对比试验，验证余弦退火衰减学习率的有效性。

4.1　随机策略

无人机用户分别采用基于 VDN 的强化学习算法和随机策略的实验结果如图 3 所示。

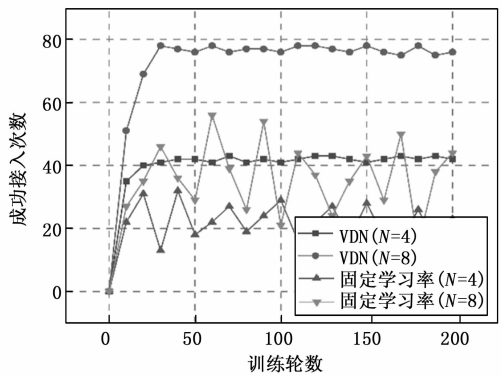

图 3　随机策略对比实验的频谱接入性能

由实验结果可以看出，本文提出的基于 VDN 的强化学习算法在无人机用户数为 4 时已实现了多智能体的隐式协作，避免了多用户同时接入同一空闲信道。而随机策略无协作机制，用户数固定时冲突率恒定，且无法适应干扰机的随机干扰策略。当无人机用户数增加为 8 时，由于全局奖励引导各智能体差异化选择空闲信道，避免了 8 个用户的动作冲突，说明了本算法的可扩展性，并且当用户数量增加时，协作机制的收益放大，性能优势更显著。但训练收敛速度略慢于 用户数为 4 的场景，需要训练更多轮次来优化协作策略。

4.2　静态分配策略

实验结果如图 4 所示。由仿真结果可以得出，静态分配策略的稳定性在一定程度上高于随机接入策略，并在训练轮数较低时，频谱接入表现接近本文提出的算法。但随着总训练轮数的增加，所提算法的接入成功率显著增长。静态分配策略在干扰机未干扰分配信道时，接入成功率稳定，但无法适应干扰机的随机干扰，且在用户数量增加后，固定信道被干扰的概率呈线性上升。而所提算法虽然初期探索阶段协作不足，但可以通过强化学习积累经验，通过协作实现信道资源的最优分配，规避了干扰和冲突。

4.3　固定学习率方案

设置固定学习率为 0.002，实验结果如图 5 所示。

可以从图中直接看出，在不同无人机用户数量情况

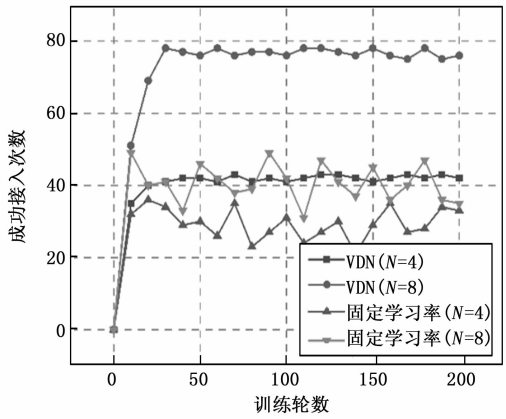

图 4 静态分配策略对比实验的频谱接入性能

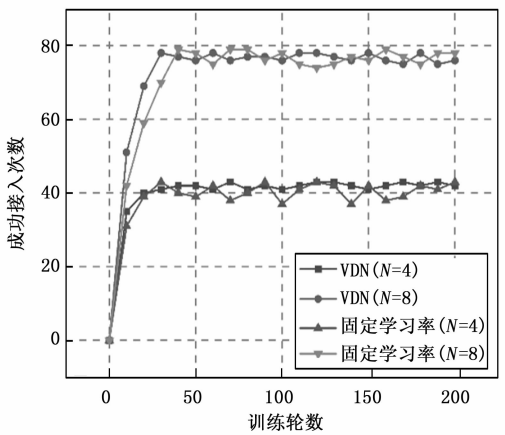

图 5 学习率固定对比实验的频谱接入性能

下，余弦退火学习率与固定学习率的接入成功率都比较接近。在无人机用户数量为 4，训练轮数较低时，余弦退火学习率与固定学习率的成功接入次数曲线几乎一致，这是由于余弦退火学习率的起始值与固定学习率接近，均能快速更新网络参数，降低初始损失。随着训练轮数的增加，训练后期参数接近最优解时，固定学习率曲线在最优解附近震荡，导致 Q 值估计不稳定；而余弦退火学习率的成功接入曲线相较于固定学习率更加稳定。在无人机用户数量为 8，训练轮数较低时，相较于固定学习率，余弦退火学习率初期较大，能快速探索复杂状态空间，积累有效经验，加速初期收敛。的频谱接入表现优于固定学习率，并且在训练轮数增加时更加稳定，同时后期小步长更新保证参数稳定，提升全局最优解的探索概率。

4.4 实验结果分析

在实验中，将本文提出的算法与随机接入策略、静态分配策略和固定学习率算法进行了比较。在多个无人机用户协作频谱接入场景中，由 N 个智能体代表 N 位无人机用户，每个智能体独立训练频谱接入算法。在每个回合开始时，N 个智能体各自选择 1 条信道进行感知，并依据感知信道的状态自主判断是否进行接入，通过与环境不断地交互训练，最终获得一个协作接入策略。

实验误差主要来源于环境随机性和算法训练随机性。实验中的训练轮数 200 轮，在训练轮数大于 50 时，每轮可视为一个独立样本，样本量为 200，满足统计检验的样本量要求，并且在无人机用户数 $N=4$ 和 $N=8$ 两种场景下，所提算法均优于对比方法，且性能差距随训练轮数扩大，趋势一致，排除偶然因素。实验结果表明，基于 MADRL 的多无人机动态信道决策算法性能明显优于随机接入策略和静态分配策略，并且在不同无人机数量的情况下，该算法的接入成功率较为稳定。

对于本文提出的余弦退火学习率衰减算法，由以上实验结果可以得出：在共有 12 条正交信道的通信环境中，在不同无人机用户数量的情况下，衰减学习率的收敛速度和稳定性都要优于固定学习率，这是因为在训练初期，衰减学习率的起始值高于固定学习率，较大的学习率可以快速降低损失。随着训练进行，参数逐渐接近最优解，损失曲面变得平坦，此时，如果学习率不变，参数更新可能会在最小值附近来回震荡，无法进一步收敛。衰减学习率可以在后期减小步长，使参数稳定在最优解附近，从而获得更低的损失和更高的模型精度。固定学习率过大时，虽然在训练初期的收敛速度会加快，但训练后期会导致震荡，固定学习率较小时，会使训练初期的收敛速度明显下降。

4.5 算法复杂度分析

在 VDN 中，每个智能体拥有一个独立的 Q 网络。前向传播时，需要并行或串行地对 N 个智能体的网络进行一次前向计算，以得到各自的 Q 值函数。因此，单次前向传播的时间复杂度为 $O(NC_f)$，其中 C_f 表示单个智能体网络前向传播的复杂度。VDN 的损失函数基于联合 Q 值 Q^{tot} 与联合目标 Q 值之间的时序差分误差。由于 Q^{tot} 是各个 Q 值的简单求和，其梯度可以简洁地分配给每个智能体网络。因此，反向传播的过程相当于对 N 个独立的网络分别进行反向传播。其时间复杂度同样为 $O(NC_b)$，其中，C_b 是单个网络的反向传播复杂度。算法整体空间复杂度为 $O(N+B)$，其中 N 为智能体数量，B 为回放池大小，呈线性增长特性且开销较低，因而能较好的保证实时性。

5 结束语

本文针对干扰条件下的多无人机信道决策问题，提出了一种基于 MADRL 的多无人机动态信道决策算法。

首先将多个无人机用户的频谱接入问题建模为去中心化的部分可观测马尔科夫过程，每个无人机用户视作一个智能体通过与无线通信环境的自主交互获取信道状态信息，并根据交互经验预测信道状态变化。其次通过价值分解算法建立单个智能体动作与系统整体回报之间的关联，促进智能体间的隐式协作，更有效地避免了接入频谱时用户之间的冲突。最后，使用余弦退火学习率衰减算法，避免了固定学习率或步进衰减可能造成的优化过程震荡和不稳定。

针对实际使用场景，由于 VDN 价值分解机制的天然适配性，可以将联合 Q 值分解为单个智能体 Q 值的线性组合，无需随无人机用户数量修改网络结构。并且分布式执行模式意味着在执行阶段每个无人机用户仅依赖局部观测决策，无需与其他智能体实时通信，避免了无人机用户数量增加导致的通信开销爆炸，在大规模无人机网络中具备较强的实用价值。

参考文献：

[1] GUPTA L，JAIN R，VASZKUN G. Survey of important issues in UAV communication networks [J]. IEEE Communications Surveys & Tutorials，2015，18 (2)：1123－1152.

[2] ZOU Y，ZHU J，WANG X，et al. A survey on wireless security：technical challenges，recent advances，and future trends [J]. Proceedings of the IEEE，2016，104 (9)：1727－1765.

[3] 宋　波，叶　伟，孟祥辉. 基于多智能体强化学习的动态频谱分配方法综述 [J]. Systems Engineering & Electronics，2021，43 (11)：1－9.

[4] RAWAT D B，SHETTY S，RAZA K. Game theoretic dynamic spectrum access in cloud-based cognitive radio networks [C] // 2014 IEEE International Conference on Cloud Engineering，IEEE，2014：586－591.

[5] SUTTON R S，BARTO A G. Reinforcement learning：an introduction [M]. Cambridge：MIT Press，1998.

[6] TAN X，ZHOU L，WANG H，et al. Cooperative multi-agent reinforcement-learning-based distributed dynamic spectrum access in cognitive radio networks [J]. IEEE Internet of Things Journal，2022，9 (19)：19477－19488.

[7] LEI W，YE Y，XIAO M. Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks [J]. IEEE Transactions on Cognitive Communications and Networking，2020，6 (3)：970－979.

[8] CHE Y，LAI Y，LUO S，et al. UAV-aided information and energy transmissions for cognitive and sustainable 5G networks [J]. IEEE Transactions on Wireless Communications，2020，20 (3)：1668－1683.

[9] 何一汕，王永华，万频. 面向多用户动态频谱接入的改进双深度 Q 网络方法研究 [J]. 广东工业大学学报，2023，40 (4)：85－93.

[10] CUI J，LIU Y，NALLANATHAN A. Multi-agent reinforcement learning-based resource allocation for UAV networks [J]. IEEE Transactions on Wireless Communications，2019，19 (2)：729－743.

[11] JIANG W，YU W，WANG W，et al. Multi-agent reinforcement learning for joint cooperative spectrum sensing and channel access in cognitive UAV networks [J]. Sensors，2022，22 (4)：1651.

[12] VAN H H，GUEZ A，SILVER D. Deep reinforcement learning with double q-learning [C] //Proceedings of the AAAI Conference on Artificial Intelligence，2016，30 (1).

[13] 胡昊南，韩　铭，李文鹏，等. 面向无线传感器网络信息年龄的多无人机轨迹优化算法 [J]. 电子与信息学报，2024，46 (4)：1222－1230.

[14] YU C，VELU A，VINITSKY E，et al. The surprising effectiveness of PPO in cooperative multi-agent games [J]. Advances in Neural Information Processing Systems，2022，35：24611－24624.

[15] SU J，ADAMS S，BELING P. Value-decomposition multi-agent actor-critics [C] //Proceedings of the AAAI Conference on Artificial Intelligence，2021，35 (13)：11352－11360.

[16] PUTERMAN M L. Markov decision processes [J]. Handbooks in Operations Research and Management Science，1990，2：331－434.

[17] OROOJLOOY A，HAJINEZHAD D. A review of cooperative multi-agent deep reinforcement learning [J]. Applied Intelligence，2023，53 (11)：13677－13722.

[18] GIRI M K，MAJUMDER S. Distributed spectrum sensing and access through deep recurrent Q-networks [C] // 2022 IEEE 6th Conference on Information and Communication Technology (CICT)，IEEE，2022：1－5.

[19] LITTMAN M L. Markov games as a framework for multi-agent reinforcement learning [C] //Machine Learning Proceedings 1994，Morgan Kaufmann，1994：157－163.

[20] OLIEHOEK F A，AMATO C. A concise introduction to decentralized POMDPs [M]. Cham，Switzerland：Springer International Publishing，2016.