ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 3,May 2026,pp. 687－700
DOI:10. 16337/j. 1004-9037. 2026. 03. 005

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于波束-信道-功率联合优化的多干扰机协同决策方法

戴　进，冯智斌，余　帅，童晓兵，徐逸凡，龚玉萍，李欣然

（中国人民解放军陆军工程大学通信工程学院，南京 210007）

摘　要：随着认知电子战的快速发展，多干扰机协同已成为提升复杂电磁环境下干扰效能的重要手段，然而现有方法普遍面临能量分散、决策耦合及动作空间维度爆炸等难题。提出一种基于深度强化学习的波束方向-信道-功率联合决策方法，构建“分布式执行、集中优化”的多智能体架构，各干扰机基于局部观测独立决策并共享全局奖励以实现策略协同；设计融合双目标网络与玻尔兹曼探索策略的改进深度Q网络（Deep Q-network，DQN）算法，解决Q值过估计问题并自适应平衡探索与利用，实现波束指向、信道选择与功率分配的三维联合优化。仿真结果表明，与独立Q学习及独立深度强化学习方法相比，所提方法干扰成功率提升至约90%，有效解决了多干扰机协同决策难题，为智能化电子对抗提供了新的技术途径。

关键词：智能干扰；深度强化学习；多智能体协同；波束成形；资源联合优化

中图分类号：TN972　　**文献标志码：**A

引用格式：戴进，冯智斌，余帅，等．基于波束-信道-功率联合优化的多干扰机协同决策方法[J]．数据采集与处理，2026，41(3)：687-700. DAI Jin，FENG Zhibin，YU Shuai，et al. Multi-jammer cooperative decision-making via joint beam-channel-power optimization[J]. Journal of Data Acquisition and Processing，2026，41(3)：687-700.

引　　言

认知电子战的兴起推动无线干扰技术从传统的“盲扰”模式向“感知-学习-决策-行动”（Observe orient decide act，OODA）的智能闭环演进[1-2]。其核心在于赋予干扰系统环境感知能力与自主决策智能[3-5]，通过实时侦测、分析和理解敌方通信信号的物理层特征与时空行为模式，干扰方可构建动态电磁态势图，并在此基础上实施精准、高效的干扰策略[4]。这种智能化转型为应对现代通信网络的复杂对抗场景提供了新的技术路径。

从实际对抗需求分析，随着信息化作战体系的持续演进，敌方对无线通信网络的依赖程度显著增强[6]，涵盖单兵战术通信、无人机集群协同及指挥控制系统信息回传等关键领域，均高度依赖稳定、低时延、高可靠的数据链路。这种结构性依赖为无线干扰技术提供了重要的应用价值与战略意义[7]。关键通信链路的有效压制将显著削弱敌方的态势感知、指挥控制与协同打击能力，从而为我方夺取战场主动权创造有利条件[8]。

现有的无线干扰技术研究大多集中在单干扰机对抗单目标场景，并已取得显著进展。文献[9]针对目标和环境动态未知，提出基于深度强化学习的干扰功率智能分配方法，提升了干扰效率与决策准确性。文献[10]针对转发式干扰中因侦察截取导致的干扰假目标滞后问题，提出一种基于生成对抗网络的干扰波形生成方法，有效提升了干扰波形的欺骗性。文献[11]针对跳频通信场景，提出了基于深度强化学习和元学习的干扰资源分配方法，能够有效应对高维动态环境挑战。然而，上述研究主要集中于单目标场景，干扰能力有限，鉴于实际通信系统中通常以多目标用户为主，相关方法无法直接推广

基金项目：国家自然科学基金（62401625，62571548）。
收稿日期：2026-04-12；**修订日期：**2026-05-10

应用。

针对多机协同场景,多智能体深度强化学习成为研究热点。文献[12]研究多干扰机协同压制分布式雷达问题,构建虚假目标拒绝概率模型,并提出改进粒子群算法求解非凸优化问题。文献[13]针对单一干扰源效率低下的问题,提出基于强化学习的多智能体协同干扰算法,有效降低用户吞吐量。文献[14]针对跳频通信场景下传统跟随式干扰面临的硬件性能与信号同步挑战,提出基于异步深度强化学习的协同干扰决策算法,显著加快学习收敛速度并实现多干扰节点高效协同。上述研究为设计多干扰机协同机制与算法提供了有益参考,但主要聚焦于单一维域的干扰决策,在复杂电磁环境下的干扰效能仍有提升空间。

进一步地,将干扰资源优化从单一维域拓展至多维域联合优化,是提升干扰效能的重要方向。传统干扰通常仅针对功率、频率或时间等单一维度进行优化,虽实现简单,但难以充分利用干扰资源的自由度,在应对采用抗干扰技术的现代通信系统时效果受限。引入多维域联合优化,如将波束成形技术与功率控制相结合,通过定向发射将能量集中于特定方向的目标接收机,可显著提升干扰功率利用率并降低被侦测概率[15-16]。然而,多维域联合优化在提升干扰效能的同时,也带来了新的挑战:不同维域的决策变量相互耦合,如波束方向的选择直接影响功率分配的有效性,信道选择又与波束覆盖范围密切相关,这种决策耦合使得传统的分步贪婪策略难以保证全局最优[15-16]。因此,如何设计适用于高维混合动作空间、能够处理不同维域决策耦合关系的多智能体协同算法,仍是亟待解决的关键问题。

基于此,本文构建多智能体协同干扰系统模型,针对多用户通信网络的动态特性,提出一种基于深度强化学习的波束方向、信道与功率联合干扰决策算法。相较于传统基于博弈论或凸优化的方法,非合作对抗场景下的深度强化学习决策方法,无需知道环境和目标用户的先验信息,通过与环境交互,仅需共享低维全局奖励即可实现多干扰机协同,自主学习训练实现波束-信道-功率动态决策[17-18]。在此基础上,该算法采用“分布式执行、集中优化”的多智能体架构,各干扰机基于局部观测独立决策以降低通信开销,同时引入全局干扰效能作为共享奖励信号引导策略收敛;设计融合双目标网络与玻尔兹曼探索策略的优化机制,缓解Q值过估计问题并实现探索与利用的自适应平衡;在总功率受限条件下实现波束方向、干扰信道与发射功率的三维联合优化。仿真结果表明,所提算法相较于基准方案在干扰成功率和能量利用效率方面均有显著提升,为复杂电磁对抗场景下的多目标协同干扰提供了有效的技术支撑。

主要贡献包括:

(1) 针对多干扰机协同中的决策耦合问题,提出“分布式执行、集中优化”的多智能体架构。与现有方法仅针对功率或频率等单一维度进行优化不同,将动作空间构造为波束方向、信道选择与功率等级三维联合空间,各干扰机基于局部观测独立决策,采用独立深度Q网络实现去中心化控制;同时设计全局干扰效能作为共享奖励信号,引导个体策略向系统最优收敛,有效缓解多机资源冲突与干扰重叠问题。

(2) 为提升多智能体深度强化学习的训练稳定性,设计融合双目标网络与自适应玻尔兹曼探索策略的改进DQN机制。与标准ε-greedy策略不同,采用的玻尔兹曼策略根据动作价值的相对概率动态调整探索行为,并结合指数衰减温度实现探索与利用的自适应平衡;双目标网络通过软更新机制有效缓解Q值过估计问题,提升训练收敛稳定性。

(3) 面向多用户干扰场景,实现波束方向、干扰信道与发射功率的三维联合优化。与现有分步贪婪策略或单维域优化方法相比,所提算法能够同时处理不同维域决策变量之间的耦合关系,在总功率受限条件下实现多维资源的协同配置。仿真实验结果表明,相较于单维域干扰策略和独立决策方法,所提算法在多目标干扰成功率、干扰能量效率等性能上有明显提升,并分析了不同用户数量下的干扰成功率。

1　系统模型与问题建模

考虑一个面向多用户通信网络的协同干扰场景,如图1所示。该场景包含M对通信用户(发送端-接收端对),用户对的集合表示为$\mathcal{M}=\{u_1,u_2,\cdots,u_M\}$,存在$W$个可用信道,信道集表示为$\mathcal{W}=\{f_1,f_2,\cdots,f_W\}$,用户发送端将向接收端传输信息,且所有用户的传输功率相同,记为P。系统中随机分布N个智能干扰机,干扰机集合表示为$\mathcal{N}=\{j_1,j_2,\cdots,j_N\}$,所有干扰机的总发射功率均为$P_j$。干扰机采用天线阵列波束成形技术,通过动态调整干扰波束方向及各波束功率分配策略实施协同干扰,以最大化对敌方网络通信的干扰效能。

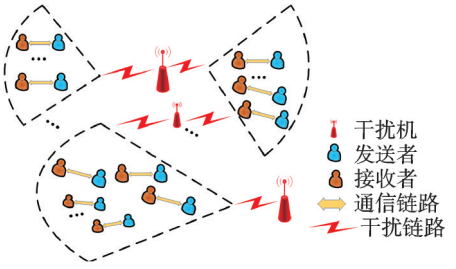

图1　多智能体协同干扰系统模型

Fig.1　Multi-agent cooperative jamming system model

图2展示了用户信道的动态变化过程。绿色表示用户对1当前所占用的通信信道,紫色表示用户对2的当前信道,假设通信用户之间按照一定的跳频规律进行通信,第n对用户的跳频序列为$\mathcal{F}=\{f_n^1,f_n^2,\cdots,f_n^{W-1},f_n^W\}$。

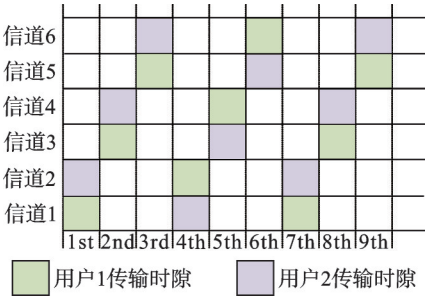

图2　用户信道变化模型

Fig.2　User channel variation model

系统时间被离散化为等长的时隙。在每个时隙内,各干扰机可选择一个或多个波束方向进行干扰;与此同时,每对通信用户持续在其占用信道上进行数据传输。时隙结构如图3所示,定义干扰机和用户的时隙集合分别为$\mathcal{T}_j=\{t_{j1},t_{j2},\cdots,t_{jT}\}$和$\mathcal{T}_u=\{t_{u1},t_{u2},\cdots,t_{uT}\}$。

智能干扰机具备频谱感知与行为学习能力,能够实时监测通信方的频率使用情况,并通过历史交互数据学习其用频行为模式。每个干扰时隙T_j由两个阶段组成,频谱感知子时隙T_{wss}和干扰实施子时隙T_s。

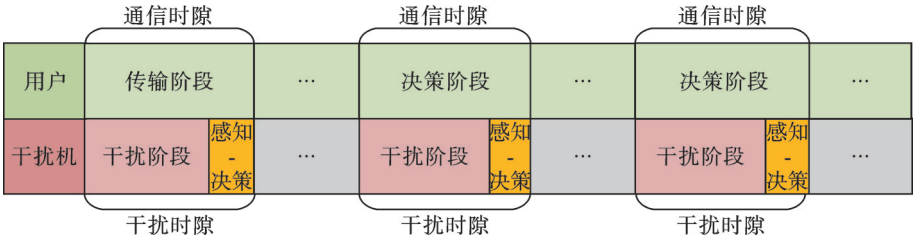

图3　干扰-用户时隙结构

Fig.3　Jammer-user timeslot structure

在非合作对抗环境下,干扰方必须依赖在线感知与学习机制来推断网络状态。因此,在不调整干扰机部署位置及总发射功率的前提下,考虑多个干扰机协同工作,利用天线阵列实现波束成形,最大化对敌方通信网络的干扰效果。在每个干扰时隙内,干扰机的总发射功率有限且保持不变,若其生成多个波束以同时干扰多个方向,则总功率需在各波束之间进行合理分配,以达到最优的干扰效果,每个干扰机可选波束数量为K,波束的最大宽度为定值θ,每个波束分配的功率占总功率可选集合为$P=\{w_1,w_2,\cdots,w_Q\}$且同时所有波束分配的功率比例之和等于1。

在实施干扰的过程中,考虑干扰机有V个波束方向,其波束方向策略集为$\{\phi_1,\phi_2,\cdots,\phi_V\}$,频率策略集为$\{f_j^1,f_j^2,\cdots,f_j^W\}$,对于第$n$个干扰机,可选择的波束数量为$K_n\in\{1,2,\cdots,K\}$,波束方向策略$\phi_n\in\{\phi_n^1,\phi_n^2,\cdots,\phi_n^{K_n}\}$,可选功率等级为$P_n\in\{w_n^1,w_n^2,\cdots,w_n^{K_n}\}$,频率策略为$F_n\in\{f_j^1,f_j^2,\cdots,f_j^W\}$,在某一

干扰时隙内,其干扰策略可表示为一个四元组$s_n=(K_n,\phi_n,P_n,F_n)$。

对于第n个干扰机的第k个波束的发射功率为$P_{\mathrm{j},n,k}=w_n^k*P_{\mathrm{j}}$,定义$g_{\mathrm{u}}^m$表示第$m$对用户链路的信道增益。假设信道服从瑞利衰落模型[19-20],则第m对用户发送端与接收端之间的信道增益g_{u}^m可表示为

$$g_{\mathrm{u}}^m=\rho_0\cdot\left(\frac{d_{\mathrm{u}}^m}{d_0}\right)^{-\sigma}|\varepsilon|^2 \tag{1}$$

式中:σ表示路径损耗因子;$\varepsilon\in\mathcal{CN}(0,1)$表示服从瑞利分布的小尺度衰落分量;$d_0$为发收两端的参考距离;$\rho_0$为参考距离处的路径损耗基准值。

第n个干扰机的第k个波束所覆盖的干扰区域S_n^k是以ϕ_n^k为中心、宽度为θ_n^k的扇形区域,其中θ_n^k表示第n个干扰机的第k个波束的宽度,u_m表示用户对m所在的位置。此外,当干扰波束的宽度为360°时,即可视为全向干扰模式。需注意,通信用户是否受到干扰仅取决于其接收端是否位于干扰波束的覆盖区域内。

对于用户u_m,定义指示函数δ_1表征其接收端是否位于第n个干扰机的第k个波束的扇形干扰区域,其数学表达式为

$$\delta_1(u_m,\mathcal{S}_n^k)=\begin{cases}1 & \text{用户}u_m\text{的接收端位于干扰机}n\text{的干扰扇形区域}\mathcal{S}_n^k\text{内部}\\0 & \text{用户}u_m\text{的接收端位于干扰机}n\text{的干扰扇形区域}\mathcal{S}_n^k\text{外部}\end{cases} \tag{2}$$

其次,给出一个指示函数δ_2来表示在t时刻干扰机n的信号是否精准覆盖第m对用户通信的信道,即

$$\delta_2(f_{\mathrm{j}}^n(t),f_{\mathrm{u}}^m(t))=\begin{cases}1 & f_{\mathrm{j}}^n(t)=f_{\mathrm{u}}^m(t)\\0 & f_{\mathrm{j}}^n(t)\neq f_{\mathrm{u}}^m(t)\end{cases} \tag{3}$$

因此,第m对用户在第t个通信时隙的信干噪比[7]可以表示为

$$\mathrm{SINR}_{\mathrm{u}}^m(t)=\sum_{n=1}^{N}\sum_{k=1}^{K_n}\frac{P(t)g_{\mathrm{u}}^m(t)}{P_{\mathrm{j},n,k}^m(t)g_{\mathrm{j},n}^m(t)+\sigma^2} \tag{4}$$

式中σ^2为噪声功率。

定义在第t个通信时隙干扰机n对M个用户对进行干扰的奖励值,即

$$V_n=\sum_{m=1}^{M}\delta_1(u_m,S_n)\cdot\delta_2(f_{\mathrm{j}}^n(t),f_{\mathrm{u}}^m(t))\cdot r_m(t) \tag{5}$$

式中$r_m(t)$为第m个通信对回报给干扰机的即时奖励,表达式如下

$$r_m(t)=\begin{cases}1 & \mathrm{SINR}_{\mathrm{u}}^m(t)<\xi_1\\0 & \text{其他}\end{cases} \tag{6}$$

式中ξ_1为用户的最低解调门限。

然而,由于环境信息受限,当一对用户被成功干扰时,难以确定通信性能下降是否由特定干扰机引起,故$\delta_1(u_m,S_n)$和$\delta_2(f_{\mathrm{j}}^n(t),f_{\mathrm{u}}^m(t))$的判定存在困难,因此引入全局奖励值如下

$$V(t)=\sum_{m=1}^{M}r_m(t) \tag{7}$$

干扰机的奖励函数V是对所有M对通信用户的干扰效能之和。

针对未知的多用户通信网络,多个智能干扰机利用其配备的天线阵列实现波束成形,对特定用户或区域实施定向干扰,以提升干扰的空间选择性与能量集中度。以最大化干扰效能为优化目标,将干扰机j_n的决策策略定义为π_n,则所有干扰机的联合策略为$\pi=\{\pi_1,\pi_2,\cdots,\pi_N\}$。多智能体的共同目标是通过优化波束方向、功率等级与信道选择,获得最优联合干扰策略$\pi^*=\{\pi_1^*,\pi_2^*,\cdots,\pi_N^*\}$,其数学表述为

$$\begin{cases} G:\max\limits_{\pi}\sum\limits_{t=1}^{\infty}V(t)=\max\limits_{\pi}\sum\limits_{t=1}^{\infty}\sum\limits_{m=1}^{M}r_m(t) \\ \text{s.t.}\ \ \phi_n\in\{\phi_n^1,\phi_n^2,\cdots,\phi_n^V\} \\ F_n\in\{f_{\mathrm{j}}^1,f_{\mathrm{j}}^2,\cdots,f_{\mathrm{j}}^W\} \\ P_n\in\{w_n^1,w_n^2,\cdots,w_n^Q\} \end{cases} \tag{8}$$

2　基于深度强化学习的干扰波束方向、信道与功率联合干扰决策算法

传统的强化学习方法[21](如Q学习)依赖表格形式存储状态-动作值函数。然而,在多智能体通信与对抗场景中,环境通常具有高维且部分可观测的状态特征,多个并发信号共同作用导致状态空间呈指数级增长。在此类复杂环境下,传统Q学习方法面临“维度灾难”,难以实现有效收敛且可扩展性受限。为克服上述局限,深度强化学习(Deep reinforcement learning, DRL)[22]通过引入深度神经网络对Q值函数进行逼近,显著提升对高维状态的建模能力。具体而言,深度Q网络(Deep Q-network, DQN)采用卷积神经网络或全连接网络对动作值函数进行参数化表示,实现高维状态空间下的有效策略学习。动作值函数可表示为$Q(s_t,a_t;\theta)$,其中θ为Q网络的可训练参数,用于评估在状态s_t下执行动作的预期累积回报。为平衡探索与利用,智能体可采用基于估计Q值的探索策略生成动作,例如玻尔兹曼(Boltzmann)策略,根据动作价值的相对大小分配选择概率,实现对未知策略空间的有效探索。针对干扰对抗场景的特殊需求进行三方面针对性改进:设计全局干扰效能作为共享奖励信号以解决非合作环境下多机干扰效果难以精确归因的信用分配问题,采用自适应玻尔兹曼探索策略替代传统ε-greedy策略以适配高维混合动作空间的分布式决策,并将“分布式执行、集中优化”(Centralized training with decentralized execution, CTDE)架构与波束方向、信道选择、功率等级三维联合动作空间深度融合以实现去中心化执行与集中式协同训练的有机结合。

所设计的CTDE多智能体干扰决策架构如图4所示。该架构的核心优势在于执行阶段通信需求极低:各干扰机仅依赖本地观测进行决策,无需交换高带宽的原始感知数据或复杂策略参数。协同所需的唯一信息为周期性全局干扰效能奖励标量,通过低速率、高可靠的反向广播链路分发,其通信开销相对于协同增益可忽略不计,有效避免通信瓶颈。在该架构中,各干扰机配备独立的DQN策略网络,进行局部观测与动作决策,但共享同一全局奖励信号进行网络参数更新。协同优化所需的关键信息通过专用反向通信链路广播,采用简单广播协议,由逻辑控制节点在每个决策周期结束后统一计算并分发全局奖励。这种极简通信设计在保证协同效能的同时,最大限度降低通信开销与暴露风险。

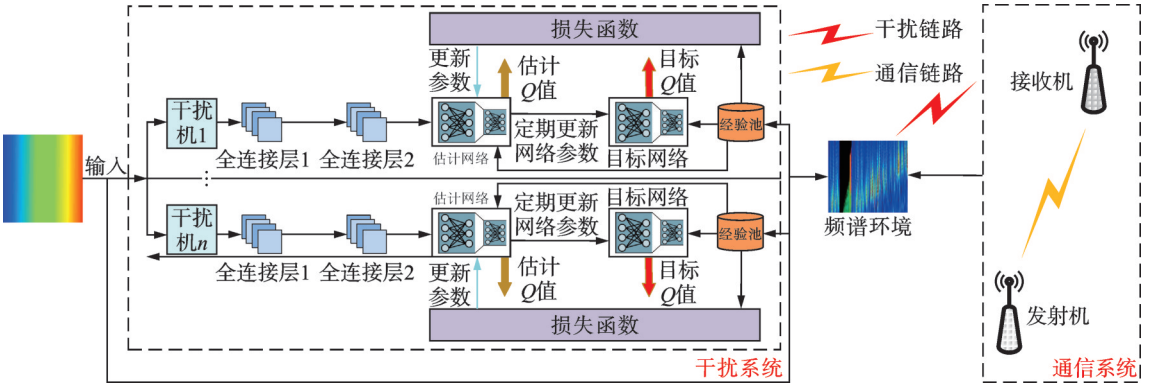

图4　“分布式执行、集中优化”多智能体干扰框架

Fig.4　CTDE multi-agent jamming decision architecture

传统单智能体方法需要同时处理波束方向选择、信道匹配和功率分配的三维联合决策，动作空间复杂度高，探索难度极大。本架构通过分布式决策将复杂度分解，每个智能体只需学习$K\times V\times Q$种动作组合的价值函数，显著降低了单个网络的探索学习负担，同时通过多智能体协作覆盖了更完整的干扰策略空间。

为提升训练稳定性，采用目标网络机制。每个智能体配置一对结构相同的Q网络：一个用于当前值估计，另一个用于目标值计算。目标网络参数定期从估计网络通过软更新或硬复制方式更新，以降低目标值波动，缓解训练不稳定性。

在动态、不确定的环境中，多智能体系统的序列决策问题通常采用多智能体马尔可夫决策过程(Multi-agent Markov decision process, MA-MDP)进行建模，N个智能体所构成的多智能体马尔可夫决策过程可形式化定义为一个元组

$$\langle N,\mathcal{S},\mathcal{A}_1,\mathcal{A}_2,\cdots,\mathcal{A}_N,\Pr,R_1,\cdots,R_N\rangle \tag{9}$$

式中：$N=\{j_1,j_2,\cdots,j_N\}$表示智能干扰机构成的智能体集合；$\mathcal{S}$为全局环境的状态空间，描述系统在任一时隙的完整观测信息(如通信链路状态、位置、信道占用等)；$\mathcal{A}_n$表示智能干扰机j_n的动作空间集合，表示其可执行的干扰行为集合；$\Pr:\mathcal{S}\times\mathcal{A}_1\times\cdots\times\mathcal{A}_N\to[0,1]$为环境状态转移概率函数；$R_n:\mathcal{S}\times\mathcal{A}_1\times\cdots\times\mathcal{A}_N\to[0,1]$为智能干扰机$j_n$的奖励函数。各元素的具体含义如下：

环境状态S。为捕捉通信用户的时-空-频动态特性并提供必要的历史信息，将t时刻的状态定义为

$$S_t=[o_t,o_{t-1},\cdots,o_{t-\Phi+1}]^{\mathrm{T}} \tag{10}$$

式中Φ表示历史时长，即用于决策的历史信息。

动作空间$\mathcal{A}_n$。智能干扰机j_n的动作空间可以表示为

$$\mathcal{A}_n\triangleq\{a_n=(f_{\mathrm{j}}^n,p_{\mathrm{j},k},s_n^k):f_{\mathrm{j}}^n\in\mathcal{W},p_{\mathrm{j},k}\in\mathcal{P},s_n^k\in\phi\} \tag{11}$$

因此，智能干扰系统的联合干扰工作空间表示为

$$\mathcal{A}\triangleq\mathcal{A}_1\otimes\mathcal{A}_2\otimes\cdots\otimes\mathcal{A}_N \tag{12}$$

式中“$\otimes$”表示笛卡尔乘积。

干扰系统决策网络中，动作价值函数通过贝尔曼方程迭代更新，以逼近最优策略。具体而言，在每次交互后，智能体依据当前环境反馈对Q值进行时序差分(Temporal difference, TD)学习。其更新过程遵循如下贝尔曼最优性方程

$$Q(\boldsymbol{s}_t,a_t)=Q(\boldsymbol{s}_t,a_t)+\alpha[r(\boldsymbol{s}_t,a_t)+\gamma\max_{a^*}Q(\boldsymbol{s}_{t+1},a^*)-Q(\boldsymbol{s}_t,a_t)] \tag{13}$$

式中：α表示学习率，控制Q值更新过程中新获取信息对历史估计的权重；γ表示折扣因子，用于调节智能体对未来奖励的重视程度。为提升训练的稳定性与样本利用率，算法采用经验回放机制，计算估计Q网络的损失函数，在每一轮训练中，从经验回放库中随机抽取一个小批量样本，并基于这些样本构建Q网络的损失函数。具体地，损失函数定义为当前Q网络输出与目标Q值之间均方误差(MSE)，即

$$L(\theta_t)=[y_t-Q(\boldsymbol{s}_t,a_t;\theta_t)]^2 \tag{14}$$

式中：θ_t表示干扰机所维护的估计Q网络(也称本地网络或在线网络)的可训练参数，y_t为对应的目标Q值，具体公式为

$$y_t=r_t+\gamma\max_{a}Q(\boldsymbol{s}_{t+1},a;\theta_t^-) \tag{15}$$

估计Q网络的参数θ通过随机梯度下降(Stochastic gradient descent, SGD)进行迭代优化，其更新过程依据损失函数的负梯度方向调整网络权重，具体更新规则如下

$$\theta_{i+1}=\theta_i+\alpha\cdot\nabla_{\theta_i}L(\theta_i) \tag{16}$$

式中：α表示学习率，控制参数更新的步长；θ_i表示估计网络中的可训练参数。

鉴于传统ε-greedy策略在分布式网络中可能导致环境不稳定,采用玻尔兹曼探索策略,动作选择策略$a(t)$的更新公式为

$$a(t)=\frac{\tau^{Q(\boldsymbol{s}_t,a_t)/\beta}}{\sum\tau^{Q(\boldsymbol{s}_t,a_t)/\beta}} \tag{17}$$

式中β为玻尔兹曼模型的相关参数。

$$\beta=\begin{cases}\beta_0 v & \beta\geqslant\beta_{\text{final}}\\ \beta_{\text{final}} & \beta<\beta_{\text{final}}\end{cases} \tag{18}$$

式中:β_0与探索时间正相关,β_{final}表示探索阶段的结束条件,v表示温度衰减率。

基于上述分析,提出基于多智能体深度强化学习的信道、功率与波束方向联合决策算法,具体步骤如表1所示。

表1　基于深度强化学习的波束方向、信道与功率联合干扰决策算法

Table 1　DRL-based joint beam-channel-power jamming algorithm

算法设计
初始化:
1. 设置学习率α,折扣因子γ等参数;
2. 生成神经网络、网络权重,初始状态。
循环开始
当$k<N$时:训练回合数达到预设最大值$T_{\max}=2\,000$或连续100回合内干扰成功率变化不超过1%时终止
1. 根据当前观测环境并构建状态向量$\boldsymbol{s}_t=[o_t,o_{t-1},\cdots,o_{t-\Phi+1}]^{\mathrm{T}}$。
2. 采用玻尔兹曼策略选择联合动作$a_t=(a_1,a_2,\cdots,a_M)$,并获得奖励值。
3. 干扰机观测下一状态$\boldsymbol{s}_{t+1}=[o_{t+1},o_t,\cdots,o_{t-M+2}]^{\mathrm{T}}$,并将经验值$e_t=(\boldsymbol{s}_t,r_t,\boldsymbol{s}_{t+1})$存储在经验池$D$中。
4. 干扰机从经验池中随机抽取Batch个样本数据。
5. 根据式(15)更新目标值$y_i=r_i+\gamma\max\limits_{a}Q(\boldsymbol{s}_{t+1},a;\bar{\theta})$;
6. 根据式(14)计算损失函数$L(\theta_t)$并更新网络参数。
循环结束

3　仿真分析与讨论

3.1　仿真参数设计

以双干扰机协同场景为例开展仿真验证,所提多智能体决策框架在结构上支持任意数量干扰节点的扩展。仿真参数设置参照文献[23],如表2所示。系统中随机分布4对用户和2个干扰机。仿真场景设定在400 m×400 m的矩形区域内,4对通信用户均匀分布于东、南、西、北4个方向,每对用户的接收端距中心约160 m,发送端位于其附近12 m处,构成外围通信链路。两个干扰机部署于中心区域([160,240]×[160,240]),形成对多方向用户的协同压制。

3.2　收敛性分析

为验证所提算法的有效性,将其性能与文献[22]中的独立Q学习方法进行对比。在该方法中,各干扰机独立运行Q学习算法,仅依据自身观测的状态进行决策,彼此之间不进行信息交互或协同,忽略群体间的合作效应。

为全面评估所提协同干扰策略的性能,仿真从干扰成功率、全局奖励值、通信方信号质量及算法可扩展性4个维度展开分析。首先对各评价指标进行明确定义。

表 2 仿真参数设置

Table 2 Simulation parameter settings

参数名称	参数设置
用户对数量M	4
干扰机数量N	2
干扰机的可选功率等级集合	$\{0.2,0.5,0.7\}$
干扰波束方向策略$\{\phi_1,\phi_2,\cdots,\phi_V\}$	$V=6,\{0,\pi/3,2\pi/3,\pi,4\pi/3,5\pi/3\}$
用户传输功率P_u/dBm	12
干扰总功率P_j/dBm	20
路径衰落系数σ	3
背景噪声功率N_0/dBm	−90
用户可用信道数	5
初始温度β_0	8
最小温度β_{final}	2.5
温度衰减率v	0.999
学习率α	0.000 1
折扣因子γ	0.95
目标网络软更新系数τ	0.005
经验回放值容量	150 000
batchSize	256

干扰成功率衡量系统对多目标通信网络的覆盖压制能力，定义为被成功干扰的用户数与通信用户总数之比。为抑制随机波动并清晰反映策略收敛趋势，采用指数移动平均(Exponential moving average, EMA)进行平滑处理，其表达式为

$$\text{JSR}_t=\alpha\cdot\frac{N_{\text{succ},t}}{N_{\text{total}}}+(1-\alpha)\cdot\text{JSR}_{t-1} \tag{19}$$

式中：$N_{\text{succ},t}$为第t个训练回合中被成功干扰的用户数，判定条件为通信方信干比SINR低于0.25；N_{total}为通信用户总数；α为平滑系数；JSR_{t-1}为第$t-1$回合的平滑后成功率。

全局奖励值对应系统总收益，定义为各干扰机在单轮决策中所获奖励的累积值，表达式为

$$V(t)=\sum_{m=1}^{M}r_m(t) \tag{20}$$

该指标综合反映了单位决策周期内系统取得的干扰成效与资源利用效率，数值越大表明整体协同效果越优。

通信方平均SINR定义为所有通信用户接收端信干噪比的算术平均值，通过计算各用户接收信号功率与干扰加噪声功率之比的对数平均获得。该指标直接反映干扰对通信链路质量的破坏程度，数值越低表明压制效果越显著。

图5展示了3种算法下干扰成功率的收敛曲线。训练初期，由于尚未掌握通信方的用频规律与空间分布特征，3种方法的干扰成功率均从10%左右开始缓慢上升。随着训练推进，基于独立Q学习的干扰策略因采用离散化的状态-动作空间表示，难以处理高维连续的波束与功率联合决策空间，学习效率受限，最终于约500个训练回合后收敛至50%附近。独立深度强化学习方法虽具备更强的环境感知能力，但因各干扰机独立决策，存在资源竞争与干扰重叠问题，多个干扰机易同时覆盖同一目标，造成能

量浪费,最终收敛至60%左右。所提协同深度强化学习算法通过全局奖励信号引导多机协同,有效避免资源冗余配置,于约800个训练回合后趋于稳定,最终干扰成功率达到90%左右,较独立深度强化学习方法提升约50%,较独立Q学习方法提升约80%。上述性能优势主要源于3个层面:全局奖励信号引导多机决策与系统目标对齐以避免资源冗余,三维联合优化使各干扰机根据信道占用与功率约束自适应调整而非盲目探索,以及玻尔兹曼温度衰减机制实现从广泛探索到精细利用的平滑过渡。

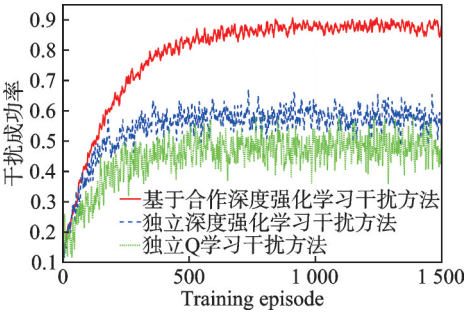

图5　干扰成功率对比曲线

Fig.5　Convergence curves of jamming success rate for three algorithms

图6给出了3种算法下全局奖励值的对比结果。初始阶段,3种方法的奖励值均处于较低水平。独立Q学习方法因表征能力受限,策略优化能力有限,奖励值增长缓慢,最终收敛于1.19附近。独立深度强化学习方法通过独立优化各干扰机的动作策略,奖励值逐步提升至2.22左右,但由于存在干扰重叠导致的能量分散,未能充分发挥多机协同优势。所提算法通过波束方向、信道与功率的三维联合优化,结合全局奖励引导的多机协同,有效避免了能量浪费。在约1 000个训练回合后趋于稳定,最终全局奖励值达到3.93,较独立深度强化学习方法提升约77%,较独立Q学习方法提升约230%。这充分验证了协同机制在提升系统整体干扰收益方面的关键作用,表明所提方法能够以更优的资源配置获得更高的综合效能。

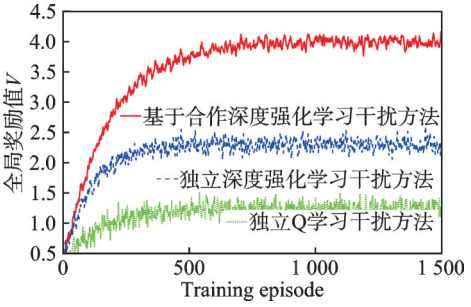

图6　全局奖励值对比曲线

Fig.6　Comparison results of global reward value for three algorithms

图7展示了通信方平均SINR的收敛特性。独立Q学习方法由于策略优化能力有限,对通信方信号质量的劣化程度不足,最终平均SINR维持于27 dB左右,通信方仍保持较高链路质量。独立深度强化学习方法通过优化波束指向与功率分配,将通信方平均SINR降至22 dB附近,干扰效果有所改善但有限。协同深度强化学习算法通过三维联合优化实现空间、频谱与功率资源的精准配置,能够更有效地降低通信方的信干噪比。随着训练推进,SINR值持续下降并最终稳定于15 dB以下,较对比方法降低约10 dB以上,表明该算法对通信质量的压制效果更为显著,能够有效破坏通信链路的可靠性。

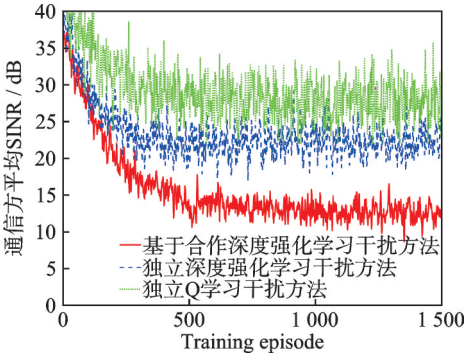

图7　用户平均SINR值对比曲线

Fig.7　Convergence characteristics of average SINR for communication

为排除几何对称性对算法性能的潜在偏置影响,进一步验证所提方法在非理想配置下的泛化能力,本文额外设计了用户随机分布场景进行对比验证。在该场景中,4对通信用户的接收端在距中心100～200 m范围内随机均匀分布,发送端随机位于接收端附近10～15 m范围内,干扰机位置保持与原对称场景一致,其余仿真参数与表2相同。如图8所示,随机分布场景下的干扰成功率最终收敛至约85%,与对称场景(88%)相比性能差距仅约3%,且收敛趋势与稳定水平基本一致。这一结果表明,所提方法对用户几何分布不具有强依赖性,能够有效适应非对称、非规则的实战电磁环境,排除了对称配置带来的性能增益假象。

为进一步验证方法在实际部署中的可行性，在收敛性分析的基础上，对算法的计算复杂度与实时运行开销进行系统评估。如表3所示，所提方法与独立DQN均采用“分布式执行”架构，各干扰机基于局部观测独立决策并维护一套DQN网络，单次决策时间复杂度均为$O(|A|)$($|A|$为波束方向、信道选择与发射功率联合动作空间的笛卡尔积规模)，空间复杂度均为$O(|\Theta|)$，网络参数量同为1.2×10^6。三者的核心差异体现在训练效率与决策稳定性：所提方法通过全局奖励协同机制引导策略向系统最优收敛，将平均训练时间由独立DQN的62±11 min缩短至45±7 min；Q-learning采用表格型存储，时间复杂度为$O(|S|\cdot|A|)$，在高维连续状态空间下随状态维度线性增长，不具备可扩展性，约15%的独立实验未能收敛。

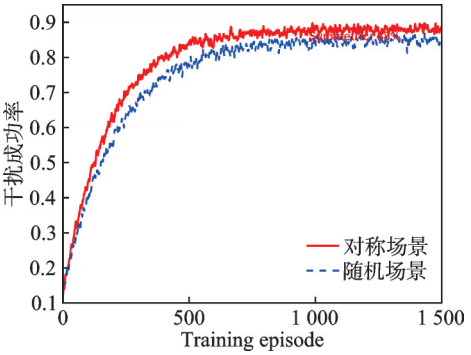

图8 随机用户分布场景下的干扰成功率曲线

Fig.8 Convergence curves of jamming success rate under random user distribution scenario

表3 计算复杂度与资源占用对比

Table 3 Comparison of computational complexity and resource occupancy

方法	时间复杂度	空间复杂度	参数量	FLOPs	Decision time/ms	Training time/min
基于合作深度强化学习干扰方法	$O(\|A\|)$	$O(\|\Theta\|)$	1.2×10^6	2.8×10^6	12.3±2.1	45±7
独立深度强化学习干扰方法	$O(\|A\|)$	$O(\|\Theta\|)$	$1.2M\times10^6$	2.8×10^6	28.5±4.3	62±11
独立Q学习干扰方法	$O(\|S\|\cdot\|A\|)$	$O(\|S\|\cdot\|A\|)$			5.1±0.8	>150

图9展示了所提算法在不同通信网络规模下的干扰成功率收敛曲线，用于验证方法的可扩展性。当用户数为3时，由于干扰资源相对充裕，算法收敛速度最快，于约600个训练回合后即达到稳定状态，最终干扰成功率维持于92%。用户数增加至4时，需干扰的目标增多而总功率保持不变，资源分配难度加大，收敛速度略有放缓，最终成功率稳定于86%附近。当用户数进一步增加至5时，干扰资源竞争加剧，算法需要更多训练回合探索最优分配策略，最终成功率收敛至78%。尽管网络规模扩大导致绝对成功率有所下降，但算法在各场景下均保持75%以上的成功率，且相较于独立学习方法仍具显著优势。上述结果表明所提算法具备较强的规模适应性，能够有效应对不同复杂度的多目标干扰场景，在动态变化的网络环境中保持鲁棒性。

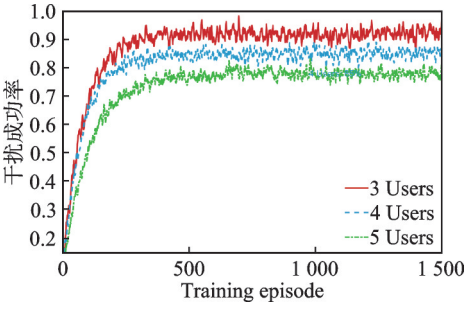

图9 不同用户干扰成功率对比曲线

Fig.9 Comparison curves of jamming success rate for different numbers of users

3.3 消融实验

为验证所提方法各关键组件的独立贡献，设计消融实验。保持其他条件不变，依次移除全局奖励机制、双目标网络及玻尔兹曼探索策略，构建3种消融配置：(1)移除全局奖励，各干扰机仅使用独立局部奖励训练；(2)采用单目标网络替代双目标网络；(3)将玻尔兹曼探索替换为固定ϵ-greedy策略。实验结果如表4及图10所示。

表4　消融实验结果对比

Table 4　Comparison of ablation experiment results

Configuration	JSR/%	全局奖励	收敛性	奖励标准差
所提方法	88	3.80	~1 800	0.15
去除全局奖励后算法	65	2.45	~2 200	0.28
替换为单目标网络后算法	74	2.95	~2 000	0.20
替换为ε-greedy后算法	80	3.35	~2 400	0.18

消融实验表明,完整方法在干扰成功率与全局奖励值上均达到最优。由图10可知,移除全局奖励后,训练曲线出现显著振荡,干扰成功率最终仅收敛至65%,全局奖励值降至2.45,且由表4可知,奖励标准差由0.15增至0.28,验证了协同奖励信号对抑制多机资源冲突的关键作用;采用单目标网络时,中期出现明显策略振荡,成功率降至74%,标准差增加约33%,证明双目标网络对缓解Q值过估计、保障收敛稳定性的有效性;采用ε-greedy策略时,前期上升斜率明显放缓,达到同等性能需额外约1 000回合,最终成功率降至80%,证明自适应玻尔兹曼策略在高维混合动作空间中具有更优的探索效率。

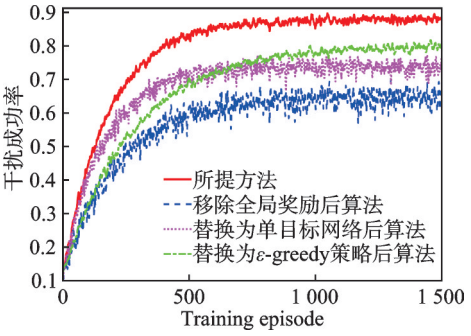

图10　不同配置下的干扰成功率收敛曲线

Fig.10　Comparison results of ablation experiments

3.4　参数敏感性分析

进一步分析关键超参数对算法性能的影响。图11展示了不同折扣因子$\gamma\in\{0.90,0.95,0.99\}$下的干扰成功率收敛曲线。$\gamma=0.99$时智能体更关注长期回报,收敛后性能最优,约为89%,但因价值传播慢,收敛速度明显放缓;$\gamma=0.90$时折扣弱,智能体过于关注即时奖励,收敛最快但最终成功率降低约5%;$\gamma=0.95$在收敛速度与最终性能之间取得平衡。图12展示了温度衰减率$v\in\{0.995\,0,0.999\,0,0.999\,5\}$对探索行为的影响。$v=0.999\,0$在探索深度与收敛速度之间取得较好平衡;$v=0.995\,0$衰减过快导致前期探索不足,陷入局部最优,最终成功率降至约85%且曲线存在明显震荡;$v=0.999\,5$衰减慢,探索过度导致训练全程波动大,收敛速度减缓。结果表明,所提方法在超参数合理范围内具有较好的鲁棒性。

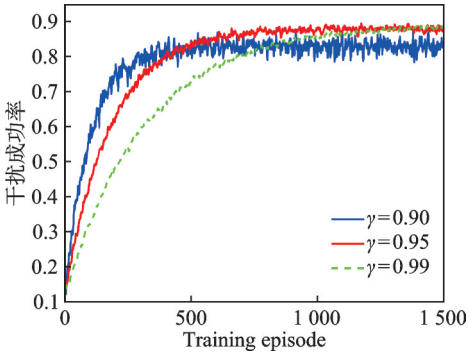

图11　不同折扣因子下的干扰成功率收敛曲线

Fig.11　Convergence curves of jamming success rate under different discount factors

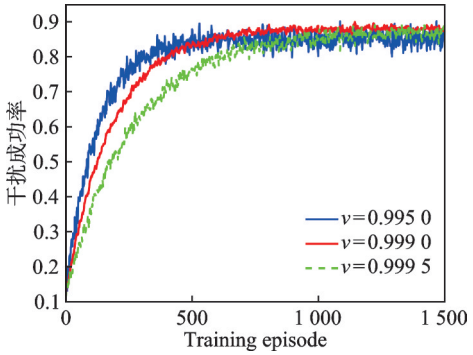

图12　不同温度衰减率下的干扰成功率收敛曲线

Fig.12　Impact of different temperature decay rates on exploration behavior

4 结束语

针对多用户通信网络干扰中协同性差、资源利用率低的问题，提出一种基于深度强化学习的多智能体协同干扰方法，实现波束方向、信道与功率的联合优化。采用“分布式执行、集中优化”框架，各干扰机基于局部观测独立决策，共享全局奖励，兼顾自主性与协同性，并通过双目标网络与玻尔兹曼策略提升算法稳定性。仿真结果表明，该方法显著优于独立学习策略，可将通信方平均SINR有效压制至较低水平，干扰成功率提升至约90%，有效增强了动态环境下的自适应协同干扰能力。

需要进一步指出的是，全局奖励机制虽然能有效引导多机协同，但存在信用分配模糊的问题，即单个干扰机难以直接判断自身动作对系统整体收益的具体贡献。所提玻尔兹曼探索策略通过动作价值的相对概率进行决策，在一定程度上弱化了个体对精确信用估计的依赖。后续研究可考虑引入差分奖励或值分解网络等显式信用分配机制，以进一步增强个体策略的可解释性。此外，将所提方法推广至Rician、Nakagami-m等更复杂信道条件以及考虑旁瓣泄漏等非理想波束模型，亦是值得深入的研究方向。

参考文献：

[1] HAYKIN S. Cognitive radio: Brain-empowered wireless communications[J]. IEEE Journal on Selected Areas in Communications, 2005, 23(2): 201-220.

[2] OSINGA F P B. ‘Getting’ a discourse on winning and losing: A primer on Boyd’s ‘Theory of Intellectual Evolution’[J]. Contemporary Security Policy, 2013, 34(3): 603-624.

[3] GUERCI J R. Cognitive radar: The knowledge-aided fully adaptive approach[M]. [S.l.]: Artech House, 2010.

[4] 黄知涛, 王翔, 赵雨睿. 认知电子战综述[J]. 国防科技大学学报, 2023, 45(5): 1-11.
HUANG Zhitao, WANG Xiang, ZHAO Yurui. Cognitive electronic warfare: A review[J]. Journal of National University of Defense Technology, 2023, 45(5): 1-11.

[5] GRIFFITHS H, COHEN L, WATTS S, et al. Radar spectrum engineering and management: Technical and regulatory issues [J]. Proceedings of the IEEE, 2015, 103(1): 85-102.

[6] ADAMY D. EW 104: Electronic warfare against a new generation of threats[M]. [S.l.]: Artech House, 2015.

[7] NERI F. Introduction to electronic defense systems[M]. 3rd ed. [S.l.]: Artech House, 2018.

[8] SCHLEHER D C. Electronic warfare in the information age[M]. [S.l.]: Artech House, 1999.

[9] JIANG H, LI G, XIE J, et al. Action candidate driven clipped double Q-learning for discrete and continuous action tasks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5269-5279.

[10] WU J, SHI C, ZHANG W, et al. Joint beamforming design and power control game for a MIMO radar system in the presence of multiple jammers[J]. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60(1): 759-773.

[11] RAO N, XU H, QI Z, et al. Fast adaptive jamming resource allocation against frequency-hopping spread spectrum in wireless sensor networks via meta-deep-reinforcement-learning[J]. IEEE Transactions on Aerospace and Electronic Systems, 2024, 60 (6): 7676-7693.

[12] SUN J, YUAN Y, GRECO M S, et al. Coordinated deception jamming power scheduling for multijammer systems against distributed radar systems[J]. IEEE Transactions on Radar Systems, 2024, 2: 1076-1088.

[13] WANG L, SONG F, FANG G, et al. A multi-agent reinforcement learning-based collaborative jamming system: Algorithm design and software-defined radio implementation[J]. China Communications, 2022, 19(10): 38-54.

[14] RAO N, XU H, DAN W, et al. Efficient jamming resource allocation against frequency-hopping spread spectrum in WSNs with asynchronous deep reinforcement learning[J]. IEEE Sensors Journal, 2024, 24(8): 13560-13577.

[15] HE R, CHEN J, LI G. Channel-aware jammer selection and power control in covert communication[J]. IEEE Transactions on Vehicular Technology, 2024, 73(2): 2266-2279.

[16] FENG Z, LI G, XU Y. Spectrum compensation-based joint communication and jamming channel allocation: A hierarchical stochastic potential game approach[J]. IEEE Transactions on Vehicular Technology, 2023, 72(3): 4046-4051.

[17] 代海波, 吴天奇, 梁轶群, 等. 基于区块链的无人机辅助铁路通信系统低能耗传输方法[J]. 数据采集与处理, 2025, 40(1):

72-85.
DAI Haibo, WU Tianqi, LIANG Yiqun, et al. Low-energy transmission method for UAV-assisted railway communication system based on blockchain[J]. Journal of Data Acquisition and Processing, 2025, 40(1): 72-85.

[18] 李国鑫, 甘麒, 陈瑾, 等. 基于指针网络深度强化学习的NOMA用户配对和功率分配方案[J]. 数据采集与处理, 2025, 40(6): 1477-1489.
LI Guoxin, GAN Qi, CHEN Jin, et al. NOMA user pairing and power allocation scheme based on pointer network deep reinforcement learning[J]. Journal of Data Acquisition and Processing, 2025, 40(6): 1477-1489.

[19] LI W, QIN Y, FENG Z, et al. "Advancing secretly by an unknown path": A reinforcement learning-based hidden strategy for combating intelligent reactive jammer[J]. IEEE Wireless Communications Letters, 2022, 11(7): 1320-1324.

[20] FENG Z, LI G, XU Y, et al. Fight against smart communication rival: An intelligent jamming approach with trend-oriented efficacy evaluation[J]. IEEE Wireless Communications Letters, 2022, 11 (11): 2290-2294..

[21] NADEEM A, ULLAH A, CHOI W. Social-aware peer selection for energy efficient D2D communications in UAV-assisted networks: A Q-learning approach[J]. IEEE Wireless Communications Letters, 2024, 13(5): 1468-1472.

[22] NIKPOUR B, SINODINOS D, ARMANFARD N. Deep reinforcement learning in human activity recognition: A survey and outlook[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4267-4278.

[23] ZHANG S, TIAN H, CHEN X, et al. Design and implementation of reinforcement learning-based intelligent jamming system [J]. IET Communications, 2020, 14(18): 3231-3238.

作者简介:

戴进(2000-),男,硕士研究生,研究方向:智能通信干扰,E-mail: 347674221@qq.com。

冯智斌(1995-),男,讲师,研究方向:无线通信对抗、智能频谱博弈。

余帅(2003-),男,硕士研究生,研究方向:无线通信对抗、智能频谱博弈。

童晓兵(1978-),**通信作者**,男,教授,研究方向:无线通信,E-mail: txb_w@sina.com。

徐逸凡(1995-),男,副教授,研究方向:无线通信、智能通信抗干扰。

龚玉萍(1978-),女,教授,研究方向:短波通信。

李欣然(1998-),女,讲师,研究方向:无线通信。

(编辑:夏道家)

Multi-jammer Cooperative Decision-Making via Joint Beam-Channel-Power Optimization

DAI Jin, FENG Zhibin, YU Shuai, TONG Xiaobing*, XU Yifan, GONG Yuping, LI Xinran

(College of Communications Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: This study aims to address the critical challenges of energy diffusion, resource conflicts, and high-dimensional action spaces inherent in multi-jammer cooperative jamming within complex electromagnetic environments. Conventional omnidirectional jamming suffers from severe energy inefficiency, while independent decision-making among multiple jammers frequently results in interference overlap. Furthermore, the joint optimization of beam direction, jamming channel, and transmit power creates an exponentially growing action space that traditional reinforcement learning methods struggle to handle. To overcome these limitations, we propose a collaborative decision-making framework based on deep reinforcement learning to achieve three-dimensional joint resource optimization with minimal communication overhead. The proposed method constructs a multi-agent architecture featuring "centralized training with decentralized execution" (CTDE), where each jammer utilizes an independent deep Q-network to approximate action-value functions based on local observations. Centralized training is achieved through a shared global reward signal defined as the total number of successfully jammed users, aligning individual policies with system-wide objectives without high-bandwidth data exchange. To mitigate Q-value overestimation, double target networks with soft parameter updating are integrated. An adaptive Boltzmann exploration strategy with exponentially decaying temperature is employed to dynamically balance the exploration and the exploitation. The action space is formulated as a three-dimensional joint space integrating beam direction, frequency channel, and power level assignment. Comprehensive simulations conducted in a 400 m×400 m scenario with four communication user pairs and two intelligent jammers demonstrate the effectiveness of the proposed approach. Quantitative results indicate that the jamming success rate reaches approximately 90%, representing a 50% improvement over independent deep reinforcement learning and an 80% improvement over independent Q-learning. This approach effectively resolves resource conflicts in multi-jammer systems through global reward sharing while ensuring low communication overhead. The integration of double target networks and adaptive Boltzmann exploration successfully addresses training instability in high-dimensional spaces. By achieving joint optimization of spatial, spectral, and power resources, the method significantly enhances energy utilization efficiency, providing a robust technical foundation for intelligent electronic countermeasures.

Highlights:

1. A novel "distributed execution with centralized optimization" multi-agent architecture is proposed to achieve collaborative jamming with minimal communication overhead and exposure to risk.
2. An improved deep Q-network algorithm integrating double target networks and adaptive Boltzmann exploration is designed to address Q-value overestimation and balance exploration-exploitation trade-offs.
3. A three-dimensional joint optimization framework for beam direction, jamming channel, and transmit power is proposed, and simulation results validate that the proposed method achieves approximately 90% jamming success rate, outperforming independent learning.

Key words: intelligent jamming; deep reinforcement learning; multi-agent cooperation; beamforming; joint resource optimization

Foundation items: National Natural Science Foundation of China (Nos.62401625, 62571548).

Received: 2026-04-12; **Revised**: 2026-05-10

***Corresponding author**, **E-mail**: txb_w@sina.com.