复杂电磁环境下基于HRDQN的智能干扰决策算法

刘天一[1,2]，吴宣利[1]，许涛[1]，王吉彬[2]，李广华[2]
（1.哈尔滨工业大学电子与信息工程学院，黑龙江 哈尔滨 150001；2.中国人民解放军63861部队，吉林 白城 137001）

摘　要：针对通信对抗中现有智能干扰决策面对复杂电磁环境收敛速度慢以及干扰能效低等问题，提出了一种基于分层Rainbow DQN（HRDQN）的智能干扰决策算法。首先，构建了存在非合作智能干扰的通信系统模型，将干扰决策过程建模为马尔可夫决策过程（MDP），并推导了压制系数门限作为干扰效果的判断依据；其次，基于分层结构设计了智能体的动作空间和决策方法，从而提升了决策效率；最后，结合压制系数门限及所估计干信比（JSR）设计了算法的奖励函数，确保算法稳定收敛。仿真结果表明，所提算法能够在快速生成理想干扰决策的同时降低干扰功耗，相较于传统智能干扰决策算法，具有更快的收敛速度，验证了所提算法的有效性。

关键词：通信对抗；智能干扰决策；马尔可夫决策过程；分层Rainbow-DQN算法；能量效率
中图分类号：TN975
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2026001

Intelligent jamming decision-making algorithm based on HRDQN in complex electromagnetic environments

Liu Tianyi[1,2], Wu Xuanli[1], Xu Tao[1], Wang Jibin[2], Li Guanghua[2]
1. School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China
2. Unit 63861 of PLA, Baicheng 137001, China

Abstract: To address the issues of slow convergence speed and poor energy efficiency of existing intelligent jamming decision-making in communication countermeasures scenarios, a hierarchical Rainbow deep Q-network (HRDQN) algorithm was proposed. Firstly, a communication system model subject to non-cooperative intelligent jamming was formulated, and the jamming decision-making process was modeled as a Markov decision process (MDP), deriving the suppression coefficient threshold to quantify jamming effectiveness. Secondly, the action space and decision-making method of the agent were designed to improve decision-making efficiency based on a hierarchical structure. Finally, the reward function was designed to combine the suppression coefficient threshold with the estimated jamming-to-signal ratio (JSR) to guarantee stable convergence of the algorithm. Simulation results demonstrate that the proposed algorithm rapidly generates ideal jamming decisions while reducing power consumption, and outperforms traditional algorithms in convergence speed, thereby corroborating the merits of the proposed algorithm.

Keywords: communication countermeasure, intelligent jamming decision-making, Markov decision process, hierarchical Rainbow-DQN algorithm, energy efficiency

0　引言

在信息化战争日益激烈的今天，通信对抗已成为决定战争胜负的关键因素之一。通信干扰作为通信对抗的重要组成部分，通过干扰敌方信息传输以

收稿日期：2025–08–30；修回日期：2025–11–10
通信作者：吴宣利，xlwu2002@hit.edu.cn
基金项目：国家自然科学基金资助项目（No.U23A20278）
Foundation Item：The National Natural Science Foundation of China (No.U23A20278)

削弱其作战能力，具有重要的战略意义。随着通信技术的快速发展，通信系统日益复杂多变，并且随着无人机、机器狗等小型智能设备在通信对抗场景中的大规模应用，通信系统日益向着智能化、无人化和小型化方向发展，此类微型化平台受物理尺寸与能源供给限制，且需降低暴露风险，故必须兼顾干扰效果与能耗约束。因此，如何以最低的功耗实现精准和高效的干扰，探索基于降低干扰功率的最佳通信干扰理论与实现方法，也成为当前通信对抗研究中亟待解决的问题[1-2]。

传统的干扰手段无法适应日益智能化和小型化的通信对抗场景。通过引入人工智能技术，可以构建具备认知能力的通信对抗系统，实现对敌方通信系统的智能感知、学习和干扰。这种智能干扰系统能够动态适应复杂电磁环境变化，学习敌方通信系统的状态变化规律，精准预测敌方动作，自主调整干扰策略，从而实现最优干扰。智能干扰区别于传统固定样式的干扰手段，其本质在于对基础干扰方式进行动态组合与自适应应用，体现出高度的智能化特征[3]。在人工智能技术推动下，采用基于强化学习的智能干扰方法与环境持续交互自主学习最优策略，适用于需要动态调整的复杂电磁环境，因此在通信对抗场景中具有重要应用价值，有望成为未来通信对抗领域的重要发展方向[4]。

针对智能干扰决策的设计方面，国内外学者纷纷展开相关研究，力求在复杂电磁环境中找到更高效与自适应的智能干扰策略。文献[5]设计了一种名为“贪婪强盗”的新算法，该算法依据功率和时序奖励规则，借助收发双方持续交互提取功率及信号样式等关键参数，进而达成有效干扰效果。在此基础上，文献[6]设计了一种智能干扰建模方法，该方法具备对用户通信行为规律的快速自适应学习能力，并且可基于学习结果实时调控，表现出显著的干扰效能。文献[7]针对跳频干扰资源分配难题，设计了分层强化学习模型分级决策频段与带宽，使干扰资源分配的问题得到了优化。文献[8]提出了一种基于高斯扰动机制的智能干扰策略生成算法，通过在干扰参数空间中进行定向探索，并结合干扰效果反馈进行迭代优化，显著提高了算法收敛速度及在未知环境中的策略生成能力。文献[9]通过深度挖掘信号的空时特性，设计了新型智能噪声干扰波形，同时具备压制性干扰效果与欺骗性干扰特性。但是以上方法策略生成速度较慢，难以适应日益复杂多变的通信环境。随着强化学习在电子战干扰策略领域的探索日益深入，为了更加简便快速地生成智能干扰策略，文献[10]采用深度神经网络生成无线信号的方式，提出了一种高效的智能欺骗干扰方法。文献[11-12]分别提出了两种创新性干扰决策算法，分别是融合动作剔除机制的深度竞争双Q网络算法和结合有效方差置信上界策略的Q学习算法，这两种算法的核心思想在于筛除低效干扰行为，从而提高最优干扰策略的搜索效率。文献[13]针对多用户场景中的多智能体干扰问题，使用多智能体马尔可夫决策过程（Markov decision process, MDP）框架对其进行建模和分析，并提出了一种基于强化学习的协同Q-learning算法，干扰效果优于独立Q-learning算法。随着深度强化学习的发展，进一步地，文献[14]针对跳频通信信号因高频切换导致干扰效率低的问题，提出了基于深度Q网络（deep Q-network, DQN）的跳频干扰决策方法。文献[15]提出了全并行DQN算法，提升了干扰机在多频道动态切换场景下的收敛速度与干扰成功率。相比于Q-learning算法，DQN算法在探索效率上有显著优势，但是由于其在设计与训练机制上的局限性，很容易收敛到次优解，且策略存在振荡风险，因此干扰决策效果不佳。近期，针对传统DQN算法在认知干扰中收敛速度慢、训练稳定性差的问题，文献[16]提出了改进的A2C算法，引入基准网络降低Critic网络方差，并通过势能函数奖励塑造引导策略优化，使收敛速度得到提升。文献[17]提出了基于先验知识嵌入的长短期记忆（long short term memory, LSTM）网络–近端策略优化（proximal policy optimization, PPO）模型，结合PPO算法限制策略更新幅度，从而提升训练稳定性。然而，以上算法未将干扰功率列为决策目标，难以满足微型化平台的能耗需求。文献[18]提出了改进的策略爬山算法，通过混合策略决策和动态奖励机制优化干扰信道与功率分配，实现了收敛速度的提升与自适应策略调整。然而，将干扰功率纳入决策目标后，决策维度增加，决策空间呈几何级数增长。上述方法在决策参数众多的高维场景中仍面临探索效率不足的问题。

尽管相关研究在智能干扰决策算法方面已取得显著成果，但在应对复杂电磁环境时仍面临一些挑

战。首先，现有算法多采用扁平的决策模型，在干扰参数维度增加时，动作空间急剧膨胀，导致在动态环境中探索效率低、策略生成慢，适应能力不足。其次，部分算法过于依赖先验知识或者初始参数的设定，收敛速度较慢且存在策略振荡风险，限制了其在未知或实时变化通信环境下的应用能力。此外，现有研究多聚焦于策略性能优化，对能耗约束等实际需求关注较少，难以兼顾微型化平台的功耗控制与隐蔽性需求。

本文主要的创新性工作如下。

1) 通过分析通信与干扰链路的特性，建立系统模型。将干扰决策的过程构建为MDP模型，分析了在不同调制方式下接收端干信比（jamming-to-signal ratio, JSR）与误码率（bit error rate, BER）的关系，并提出了压制系数门限α作为干扰效果的判断依据。

2) 提出了基于分层Rainbow DQN（hierarchical rainbow deep Q-network, HRDQN）的智能干扰决策算法。首先，定义了状态空间和动作空间，将MDP模型具体化；随后，基于分层结构设计了智能体的动作空间与决策方法，将高维动作空间解耦为两个低维子空间，从结构上降低了决策复杂度，提高了算法的探索效率与收敛速度。

3) 基于压制系数门限α及所估计的JSR设计了智能决策算法的奖励函数，确保算法稳定收敛。仿真结果表明，本文算法具有更快的收敛速度、更高的探索效率和更低的能耗。通过将本文算法与RDQN和PPO算法进行对比分析，验证了本文算法在探索效率、收敛速度等性能上具有明显优势。

1　模型构建

1.1　系统模型

智能干扰系统模型如图1所示，主要包括一对通信收发机和一个智能干扰机。发送机到接收机之间存在通信链路，发送机到智能干扰机之间存在感知链路，智能干扰机到接收机之间存在干扰链路。通信收发机负责发送和接收信息，包括信源、编码器、调制器、信道、解调器和解码器等多个部分。智能干扰机具备感知电磁环境的能力，并部署有智能体，通过对通信方行为和规律的学习，发送干扰信号以破坏通信收发机的正常通信。

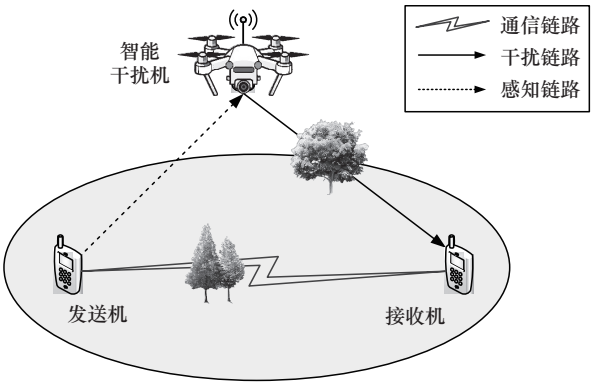


图1　智能干扰系统模型

通信干扰只能干扰接收机，干扰方法是向目标接收机注入不希望其接收到的通信信号，伴随着希望接收到的通信信号一同进入接收机，而且干扰信号必须足够强，才能使接收机无法从接收到的通信信号中恢复出所需信息。在该系统模型中，通信接收端的JSR是决定干扰效果的核心指标。考虑通信链路与干扰链路的信道衰落，通信接收端的干扰功率与通信信号功率之比可表示为

$$\mathrm{JSR}=\frac{P_j\left|H_j\right|^2}{P_c\left|H_c\right|^2} \tag{1}$$

其中，P_j与P_c分别为智能干扰机与发送机的发射功率，H_j与H_c分别为干扰链路与通信链路的复信道增益。

干扰决策的核心目标可归结为一个具有明确物理意义的数学优化问题：在动态未知的复杂电磁环境中，确保对敌方通信实施有效压制的同时，最大限度地节约干扰功率，以提升平台隐蔽性与续航能力。该目标可形式化为如式(2)所示优化问题。

$$\begin{aligned}&\min_{P_j,m_j,f_j}\quad P_j\\&\text{s.t. C1: } f_j=f_c\\&\qquad \text{C2: } \frac{P_j\left|H_j\right|^2\eta(m_j,m_c)}{P_c\left|H_c\right|^2}\geqslant\alpha(m_c)\\&\qquad \text{C3: } P_j\leqslant P_{j,\max}\\&\qquad \text{C4: } m_j\in M, f_j\in F\end{aligned} \tag{2}$$

其中，干扰功率P_j、干扰调制方式m_j和干扰频点f_j共同作为决策变量，目标函数设定为最小化P_j。C1频点匹配约束$f_j=f_c$是干扰生效的先决条件，确保干扰能量能够有效对准通信信号。C2干扰有效性约束则通过干信比条件JSR $\geqslant\alpha$保证了通信链路的

有效压制，该约束中引入的干扰效率函数 $\eta(m_j,m_c)\in(0,1]$，表示当干扰与通信调制方式完全匹配时，$\eta\approx1$，此时能量利用效率最高；一旦失配则 $0<\eta<1$，意味着达成相同干扰效果需付出更高功率代价。此外，干扰调制方式 m_j 与干扰频点 f_j 作为关键决策变量，其作用在于通过约束条件间接制约干扰功率 P_j 的可达下界：系统唯有在满足 $f_j=f_c$ 且通过 $\eta(m_j,m_c)$ 维持较高水平时，才可能以最低的 P_j 满足干扰门限的要求。约束C3和C4为智能干扰机的物理特性。

常规优化方法难以求解该优化问题。首先，环境参数具有高度的不确定性与对抗性：信道 h_j 和 h_c 是随机过程；通信方具备抗干扰能力，其核心参数 f_c 会主动跳变，这使优化问题中的约束条件本身就是一个快速移动的目标。其次，干扰效率函数 $\eta(m_j,m_c)$ 的精确形式难以获取，这些因素共同导致常规优化方法失效。

1.2　智能干扰系统

智能干扰机可实现自主决策，其核心是一个由感知、决策、执行和评估模块构成的闭环系统，结构如图2所示。感知模块负责实时监测和捕获目标通信系统的信号特征并进行分析，获得频率、调制方式及信号强度等关键参数。决策模块基于感知数据，运用机器学习和算法分析，智能选择最优干扰策略，实现以最小的代价达到理想的干扰效果。执行模块根据决策指令进行干扰信号的生成与发射。评估模块通过对干扰前后通信系统性能的对比分析，得出干扰效果参数。一般通过估计接收端JSR，以及观察通信方是否改变通信参数等手段获得。随后反馈干扰效果至决策模块，形成闭环优化机制。智能干扰系统的决策目标是智能地选择干扰参数，动态地控制JSR达到理想值。鉴于在通信感知和干扰评估方面已经有许多现有方案[19-21]，因此在设计智能干扰决策时，对感知和评估模块进行理想化处理，即感知模块可以直接获得通信方发射端的频率和调制方式，评估模块可以直接获得通信方接收端的JSR。

将智能干扰决策过程构建为一个MDP模型，通过定义状态、动作、状态转移概率以及奖励函数，能够很好地捕捉并描述通信对抗中干扰决策的动态特性和决策目标。具体地，用（S,A,P,R）四元组表示，其中 S 表示智能干扰机的状态空间，即所面临的通信环境及位置信息；A 表示智能干扰机的动作空间，即可以采取的干扰策略；P 表示状态转移的概率；R 表示采取动作后获得的奖励[22]。

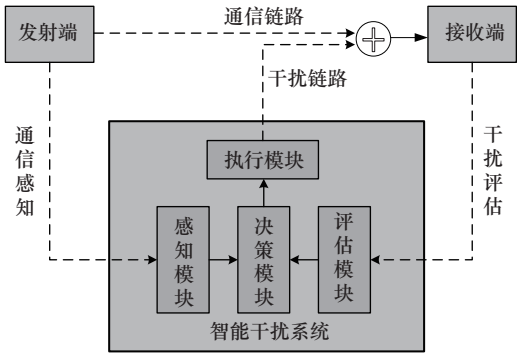


图2　智能干扰机结构

基于上述模型，智能干扰决策的本质可归结为一个序列优化问题：智能体需要学习一个策略 π，通过与环境交互，最大化其所能获得的长期累积折扣奖励。该优化问题在强化学习框架下可形式化表示为

$$\max_{\pi} J(\pi)=E_{s_0,a_t\sim\pi(\cdot|s_t),s_{t+1}\sim P(\cdot|s_t,a_t)}\left[\sum_{t=0}^{\infty}\gamma^t R\left(s_t,a_t,s_{t+1}\right)\right] \tag{3}$$

其中，在离散时间步 t，智能体观察状态 $s_t\in S$，依据策略 π 采取动作 $a_t\in A$，环境根据 P 转移到新状态 s_{t+1}，并产生一个标量奖励 $r_{t+1}=R\left(s_t,a_t,s_{t+1}\right)$，通过设计奖励函数，将式(2)的目标与约束内化到智能体的优化目标中；$\gamma\in[0,1)$ 为折扣因子，用于权衡即时奖励与未来奖励的重要性。

式(3)的解对应一个最优动作价值函数 $Q^*(s_t,a_t)$，即在状态 s_t 下执行动作 a_t 后，遵循最优策略所能获得的最大期望累积奖励。$Q^*(s_t,a_t)$ 满足贝尔曼最优方程，如式(4)所示。

$$Q^*(s_t,a_t)=E_{s_{t+1}\sim P(\cdot|s_t,a_t)}\left[R(s_t,a_t,s_{t+1})+\gamma\max_{a_{t+1}\in A}Q^*(s_{t+1},a_{t+1})\right] \tag{4}$$

从式(4)可以看出，$Q^*(s_t,a_t)$ 函数具备递归结构，算法的目标是通过训练一个神经网络来近似这个 $Q^*(s_t,a_t)$ 函数。奖励函数 R 的具体设计是连接优化目标与干扰物理效果的关键，目的为当干扰动作使接收端JSR趋近于压制系数门限 α 时，给予高奖励，从而引导智能体在保证有效干扰的同时，避免不必要的功率消耗，同时降低暴露风险。

2 基于HRDQN的智能干扰决策算法

2.1 RDQN算法

深度强化学习是结合深度学习与强化学习的一种新兴技术，旨在解决传统强化学习方法在处理高维状态空间和复杂决策问题时的局限性，近年来在游戏、机器人控制、自动驾驶等复杂决策任务中取得了显著成果，其中DQN作为开创性的工作，将Q-Learning算法中的Q-Table更新问题变成一个函数拟合问题，使用神经网络得到状态动作的Q值，并通过更新参数使Q函数逼近最优Q值，为强化学习与深度学习的结合奠定了基础。然而，DQN在处理连续动作空间、长期依赖、探索–利用权衡等问题上存在局限性。为了解决这些问题，研究人员提出了一系列改进算法。Rainbow-DQN正是这样一种集成多种增强技术的深度强化学习算法，提升了DQN在复杂环境中的学习效率和性能[23]。

1) Double-DQN（DDQN）

DDQN解决了传统DQN估值过高的问题，在DQN中，选择下一时刻动作和计算下一时刻状态–动作Q值时，使用的都是目标网络（Target-Net）。通过解耦动作选择与价值估计解决过估计问题，在计算实际Q值时，动作选择由评估网络（Eval-Net）得到，而价值估计由Target-Net得到。

$$y_t^{\mathrm{DDQN}} = r_t + \gamma Q_{\mathrm{target}}\left(s_{t+1}, \arg\max_a Q\left(s_{t+1},a;\theta\right);\theta^-\right) \tag{5}$$

其中，θ是评估网络的参数，θ^-是目标网络的参数。

2) 优先级经验回放（prioritized experience replay, PER）

传统DQN的经验回放采用均匀采样，而PER则根据样本的时序差分（temporal difference, TD）误差，即预测Q值和目标Q值之间的绝对差赋予样本不同的优先级，使TD误差较大的样本被采样的概率更高，从而加速学习过程。

3) 竞争网络结构

竞争网络结构将Q网络分解为状态价值函数$V(s)$和动作优势函数$A(s,a)$两部分，这种结构使网络能够更有效地学习状态的价值，最后通过聚合层组合得到Q值。

$$Q(s,a) = V(s) + A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a') \tag{6}$$

4) 多步学习

标准DQN使用单步回报，即一步TD目标，而RDQN则使用n步回报来更新Q值，以平衡偏差和方差。

$$y_t^{\text{multi-step}} = \sum_{k=0}^{n-1}\gamma^k r_{t+k} + \gamma^n \max_a Q_{\mathrm{target}}\left(s_{t+n},a\right) \tag{7}$$

这种方法可以加速奖励的传播，在奖励稀疏的环境中尤其有效。

5) 分布式Q学习

传统DQN学习Q值的期望，而分布式Q学习则学习Q值的完整分布，可以获得更多有用的信息，从而得到更好且更稳定的结果。

6) 噪声网络

为了改进探索策略，噪声网络在网络的权重参数中加入噪声，从而在训练过程中实现自适应的探索。这替代了传统的ε-greedy探索机制，使探索策略能够根据状态的不同而自适应调整。

2.2 基于HRDQN的智能干扰决策算法设计

在通信对抗的复杂环境中，通信和干扰的参数多且维度较大，通信参数主要包括通信载波频率、通信调制方式、通信功率、通信距离等；干扰参数包括干扰载波频率、干扰调制方式、干扰功率和干扰距离等。RDQN算法在处理这样的高维问题时，其扁平决策结构难以应对由多种异构参数构成的高维混合动作空间所带来的维度灾难，收敛速度慢，无法满足干扰决策的实时性要求。为此，本文提出了HRDQN算法，该算法的分层结构设计尤其适用于跳频通信这类参数捷变系统。在跳频通信场景中，通信频率快速切换，算法通过高层策略跟踪跳变的频率，低层策略优化干扰参数，实现了对该场景的高效决策。

基于MDP模型，本文算法的总体逻辑框架如图3所示，其核心在于引入分层决策机制，将联合动作决策分解为高层频点选择与低层参数配置两个阶段。

智能体从环境中感知状态s_t，状态信息同时输入高层策略网络和低层策略网络。高层策略网络π^{H}根据当前状态输出频点选择动作a_t^{H}，该动作与感知状态s_t共同构成组合状态，输入低层策略网络π^{L}，进而输出具体的调制方式与功率等级动作a_t^{L}。高层与低层动作组合成最终执行的干扰动作a_t作用于环境。环境根据动作a_t转移到新感知状态s_{t+1}，

并产生一个奖励信号r_{t+1}。该奖励信号经过差异化分配，$r_{t+1}{}^{\mathrm{L}}$直接用于更新低层策略，而高层策略则依赖于周期累积的内在奖励$r_{t+1}{}^{\mathrm{H}}$进行更新。下面将对该框架中的各组成部分进行详细阐述。

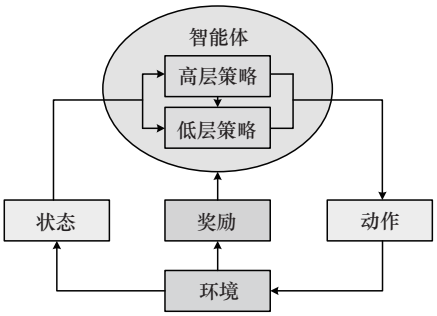


图3　总体逻辑框架

2.2.1　状态和动作设计

1) 状态空间。根据通信对抗场景的特点，将状态空间定义为包含通信频率、通信调制方式、通信功率、通信距离与干扰距离参数的集合，表示为$s \in S = \{ f_c, m_c, p_c, d_c, d_j \}$。

2) 动作空间。干扰动作空间定义为包含干扰频点、干扰调制方式与干扰功率参数的集合，表示为$a \in A = \{ f_j, m_j, p_j \}$。进一步地，将动作空间分为高层空间$a^{\mathrm{H}} \in A^{\mathrm{H}} = \{ f_j \}$和低层空间$a^{\mathrm{L}} \in A^{\mathrm{L}} = \{ m_j, p_j \}$。

3) 分层动作决策。将复杂决策过程分解为双层决策框架，如图4所示，分别为高层决策和低层决策，两个层次决策紧密协同但功能解耦，实现决策优化。

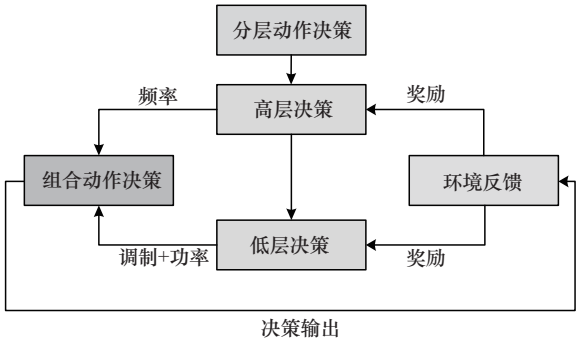


图4　分层动作决策

高层决策专注于战略决策，负责选择最优的干扰频点，本质上定义了当前干扰任务的首要目标，即确定需要集中干扰的频段。具体为接收感知的环境状态信息，通过神经网络对这些环境状态信息进行处理，输出一个离散的干扰频点选择决策。

低层决策负责战术执行，在高层决策确定的频点目标框架下，进一步优化具体的干扰参数配置。具体为使低层决策能根据高层目标自适应调整行为，通过专门设计的动作空间编码机制，将离散的调制类型和功率组合，输出调制方式与功率等级的联合决策。

两个层级的决策融合为组合动作决策，同时二者通过独立的回放缓冲区和优化器进行异步更新。高层决策通常采用较低的更新频率保持决策稳定性，而低层决策则每一步都更新以实现快速适应，这种差异化的更新机制显著提升了训练效率。整个分层架构通过目标传递和奖励反馈形成闭环决策系统，使智能体能够同时处理长期战略目标和短期战术优化，有效解决了大规模动作空间下的维度灾难问题。

2.2.2　状态转移概率

在MDP模型内，状态转移概率$P(s_{t+1} | s_t, a_t)$定义了环境的动态特性，它量化了在当前状态s_t下采取动作a_t后，环境转移到新状态s_{t+1}的似然性。本文所构建的通信对抗环境，其状态转移是一个由通信方抗干扰行为、无线信道时变衰落特性与智能干扰动作的施加共同驱动的复杂随机过程。

1) 通信方的抗干扰行为。在典型对抗通信场景中，通信方并非静态目标，其通信参数会根据通信链路质量自适应调整。例如，当通信链路性能恶化至特定门限时，通信方会主动切换通信频率或调制方案以规避干扰。这类行为使环境状态的变化呈现出强烈的对抗性与智能性，是状态转移中不确定性的一大核心。

2) 无线信道的时变衰落特性。此为环境中固有随机性的物理基础。无线信道的复增益是一个典型的随机过程，其小尺度衰落特性导致接收端信号与干扰功率均呈现随机波动。因此，即便通信与干扰双方的参数和空间位置维持不变，接收端的JSR等关键观测状态也会随信道衰落而动态变化，这为状态感知与决策带来了持续的随机扰动。

3) 智能干扰动作的施加。此为驱动状态演化的主导控制变量。智能干扰机施加的动作直接改变了干扰信号的特性，从而主动引导着环境状态的转移方向。一个有效的干扰策略不仅会立即改变接收端的瞬时JSR，更可能作为关键事件，触发通信方上述的抗干扰行为，引发通信参数的连锁变化，形成智能体与环境之间的闭环博弈。

总体而言，状态转移概率是智能干扰机在动态通信环境中的状态迁移模式，刻画了一个兼具对抗与物理随机的复杂动态环境，其核心由通信参数变化规律与干扰策略的交互机制共同决定。在通信对抗场景下，通信方的参数调整由感知模块与评估模块获取，与智能干扰机的动作执行形成闭环反馈，确保了状态迁移的马尔可夫性质。

2.2.3 结合JSR及压制系数门限的奖励函数设计

1) JSR-BER关系推导

接收端接收信号可表示为$r(t)=s(t)+j(t)+n(t)$，其中$s(t)$为信号分量，$j(t)$为干扰分量，$n(t)$为噪声分量。

那么接收端信噪比（signal-to-noise ratio, SNR）与JSR可分别表示为$\mathrm{SNR}=\dfrac{P_s}{N_0B}$与$\mathrm{JSR}=\dfrac{P_j}{P_s}$，其中，$P_s$和$P_j$为功率，$N_0$为功率谱密度，那么接收端有效信干噪比（signal-to-jamming plus noise ratio, SJNR）为

$$\mathrm{SJNR}=\frac{P_s}{P_j+N_0B}=\frac{1}{\mathrm{JSR}+\dfrac{1}{\mathrm{SNR}}} \tag{8}$$

对于二进制相移键控（binary phase shift keying, BPSK）系统来说，$E_b=\dfrac{P_s}{R_b}$，$R_b=B\Rightarrow E_b=\dfrac{P_s}{B}$，那么BPSK系统的理论BER可表示为

$$\mathrm{BER}_{\mathrm{BPSK}}=Q\left(\sqrt{2\mathrm{SJNR}}\right)=Q\left(\sqrt{\frac{2}{\mathrm{JSR}+\dfrac{1}{\mathrm{SNR}}}}\right) \tag{9}$$

同理，正交相移键控（quadrature phase shift keying, QPSK）与16进制正交幅度调制（16 quadrature amplitude modulation, 16QAM）的理论BER公式可推导为

$$\mathrm{BER}_{\mathrm{QPSK}}=Q\left(\sqrt{\frac{2}{\mathrm{JSR}+\dfrac{2}{\mathrm{SNR}}}}\right) \tag{10}$$

$$\begin{aligned}\mathrm{BER}_{\mathrm{16QAM}}=&\frac{3}{8}\mathrm{erfc}\left(\sqrt{\frac{\mathrm{SNR}}{10(1+\mathrm{JSR}\cdot\mathrm{SNR})}}\right)+\\&\frac{1}{4}\mathrm{erfc}\left(3\sqrt{\frac{\mathrm{SNR}}{10(1+\mathrm{JSR}\cdot\mathrm{SNR})}}\right)-\\&\frac{1}{8}\mathrm{erfc}\left(5\sqrt{\frac{\mathrm{SNR}}{10(1+\mathrm{JSR}\cdot\mathrm{SNR})}}\right)\end{aligned} \tag{11}$$

2) 奖励函数设计

高层策略的训练采用基于目标跟踪准确性的内在奖励机制，当选择的频点与实际通信频点匹配时获得正奖励，失配时获得负奖励，这种设计使得高层策略能快速收敛到有效的频点跟踪策略。在训练机制上，分层设计采用差异化的奖励分配策略，高层策略依赖频点跟踪策略的内在奖励，低层策略则直接使用环境反馈的干扰效果奖励，二者互相解耦，并通过独立的回放缓冲区和优化器进行异步更新。故将奖励函数设计为：当$f_j\neq f_c$时，$r^{\mathrm{L}}=-1$；当$f_j=f_c$时，$r^{\mathrm{H}}=1$，$r^{\mathrm{L}}=R$。

当一个通信系统的BER大于0.1时，就认为这个通信系统无法正常通信，干扰效果比较好[24]，可以将使BER达到0.1时所需的JSR值定义为压制系数门限α。智能干扰决策的设计在考虑达到预期干扰效果的基础上，还要考虑最小化干扰功耗，因而奖励值的设计需要与干扰效果和干扰功率相关。因为干扰效果和干扰功率都与JSR直接相关，故设计干扰效果奖励函数为

$$R=\begin{cases}-10 & ,f_j\neq f_c\\ 20\mathrm{e}^{0.1(\mathrm{JSR}-\alpha)}-10 & ,f_j=f_c\text{且}\mathrm{JSR}<\alpha\\ 10\mathrm{e}^{-0.1(\mathrm{JSR}-\alpha)} & ,f_j=f_c\text{且}\mathrm{JSR}\geqslant\alpha\end{cases} \tag{12}$$

通过奖励函数的设计，预期达到的效果为：当$f_j\neq f_c$时，认为干扰无效，智能体获得一个负奖励以惩罚智能体的错误行为；当$f_j=f_c$时，奖励值会在JSR接近α时变大，远离α时减小。这意味着JSR在α附近时，干扰效果与干扰功率处于均衡状态，即当$\mathrm{JSR}<\alpha$时，干扰效果不理想，需增大干扰功率，提高JSR以达到理想干扰效果。当$\mathrm{JSR}\geqslant\alpha$时，干扰效果较好，此时再增加干扰功率，会使干扰能效降低，奖励值减少。在式(12)中分别引入常量0.1，目的是确保奖励函数在JSR越趋近于α时，奖励值提升越快，以帮助智能体快速找到最优策略。常量10与20的目的是首先保证奖励值连续，其次确保$f_j=f_c$时的奖励值要始终高于$f_j\neq f_c$，同时奖励值变化要在α右侧变化较缓，即决策f_j的优先级高于其他动作，干扰效果的策略优先级高于干扰能效。当$f_j=f_c$时，奖励函数曲线如图5所示。

2.2.4 算法流程

基于上述设计，HRDQN算法的训练流程如图6所示。

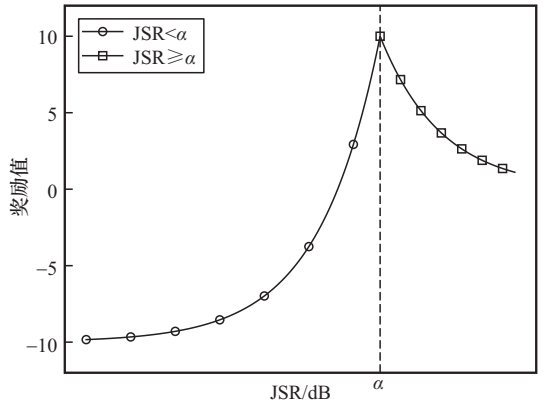


图5　奖励函数曲线

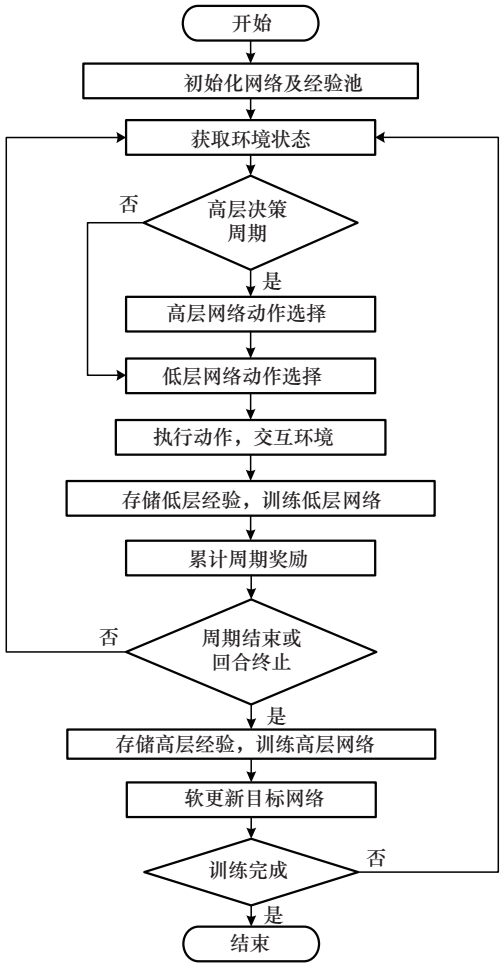


图6　HRDQN算法训练流程

该训练流程的核心在于高低层策略的异步训练机制。高层策略以一个决策周期为单位进行更新，其目标是学习稳定的频点跟踪策略；低层策略则在每一步都进行更新，以实现对干扰参数的快速自适应。为平衡探索与利用，算法采用噪声网络实现自适应的探索策略。其详细步骤如算法1所示。

算法1　基于HRDQN的智能干扰决策

输入　环境参数（状态集、动作集和奖励函数），超参数（学习率、折扣因子、网络参数、软更新系数、批量大小和经验池容量），分层决策周期

输出　高层策略，低层策略，性能指标

1) 初始化高层和低层策略的策略网络$Q(\theta)$与目标网络$Q(\theta')$

2) 初始化高层经验池D_{H}、低层经验池D_{L}，以及周期累积奖励R

3) 获取环境初始状态

4) for 时间步$t=1\rightarrow T$

5)　分层动作选择高层：若$(t-1)\bmod c=0$，观察s_t，选择并存储频率目标g_t及其决策状态s^{H}低层：根据组合状态(s_t,g_t)选择具体干扰动作c_t

6)　环境交互：组合成最终动作$a_t=(g_t,c_t)$，执行该动作后获得环境奖励r_t和下一状态s_{t+1}

7)　低层经验生成与存储计算低层奖励r^{L}将低层转移$(s_t,c_t,r_t,(s_{t+1},g_t))$存入$D_{\mathrm{L}}$

8)　低层网络训练：从D_{L}中采样，通过最小化TD误差损失$L(\theta_{\mathrm{L}})$来更新低层的策略网络$Q(\theta_{\mathrm{L}})$

9)　高层奖励累积：$r^{\mathrm{H}}\leftarrow r^{\mathrm{H}}+r_t$

10)　if 周期结束 or 回合结束 then

11)　　高层经验生成与存储将周期累积奖励作为高层奖励$r^{\mathrm{H}}\leftarrow R$将高层转移$(s^{\mathrm{H}},g_t,r^{\mathrm{H}},s_{t+1})$存入$D_{\mathrm{H}}$

12)　　高层网络训练：从D_{H}中采样，更新高层策略的策略网络$Q(\theta_{\mathrm{H}})$并重置$R\leftarrow 0$

13)　end if

14)　软更新目标网络：对高层和低层的目标网络进行软更新$\theta'\leftarrow\tau\theta+(1-\tau)\theta'$

15) end for

2.3　算法理论分析

2.3.1　可行性分析

本文所提HRDQN算法解决智能干扰决策问题在理论上是可行的，其依据主要包括以下3个方面。

首先，智能干扰决策过程可建模为马尔可夫决策过程，其状态空间复杂、动作空间高维离散。本文算法继承RDQN框架，作为一种深度强化学习算法，具备处理此类问题的能力，其深度神经网络能够对连续或高维状态进行有效表征，并基于Q值的离散动作选择机制与干扰参数的离散特性一致。此外，该算法所整合的DDQN、PER等技术，有助于提升训练稳定性与样本效率，为在动态复杂电磁环境中求解最优干扰策略提供了基础。

其次，HRDQN算法所采用的分层结构符合通信干扰环境中的决策特性。通信频率跳变是影响JSR的主要因素，而信道衰落与调制方式匹配程度则决定其具体数值。该算法将动作空间划分为高层策略与低层策略：高层策略负责干扰频点选择，以响应通信参数的宏观变化；低层策略则在选定频点基础上优化调制方式与功率等级，实现对干扰效果的精确控制。该分层结构通过将高维动作空间分解为两个低维决策阶段，有效降低了策略搜索的复杂度，从而在理论上保障了算法在复杂环境中的收敛速度与决策效率。

最后，本文所设计的奖励函数以压制系数门限α为优化目标，当JSR = α时奖励函数取得最大值，从而使强化学习的目标与JSR跟踪任务相一致。该奖励函数在JSR偏离α时产生显著负奖励，引导智能体避免无效决策。当JSR接近α时则提供更为敏感的奖励变化，促进策略的局部优化。此类设计为策略梯度提供了明确的优化方向，有助于使算法稳定收敛至近似最优策略。

综上所述，HRDQN算法在问题建模、结构设计与奖励引导方面均具备理论可行性。

2.3.2 计算复杂度分析

算法计算复杂度主要与网络结构规模、空间维度以及策略更新机制密切相关，主要基于以下关键操作计算每步计算量，采用浮点运算次数（FLOPS）作为计算量单位。

1) 前向传播：状态到动作的推理计算。

2) 反向传播：梯度计算与参数更新。

3) 经验处理：经验存储、采样与优先级更新。

4) 目标网络更新：参数软更新。

5) 探索机制：噪声生成与处理。

为进一步评估计算效率，将本文算法与原始RDQN算法及当前较为先进的PPO算法进行对比。RDQN算法在DQN的基础上进行了六大改进，性能得到了极大提升。PPO算法作为一种先进的同策略算法，其优势在于通过裁剪的概率比和专门的价值函数设计，保证了策略更新的稳定性，并凭借其算法结构简单、易于实现的优点获得了广泛应用。然而，其同策略的本质也导致了明显的劣势：无法使用经验回放，必须依赖当前策略实时采集的数据，这使其样本效率相对较低，在面对需要大量环境交互的复杂问题时，收敛速度可能受限。相比之下，HRDQN与RDQN算法均属于异策略算法，能够充分利用经验回放机制，重用历史数据，从而显著提高了样本利用效率。

当主要超参数设置相同时，3种算法单步计算量统计如表1所示。

表1 算法单步计算量统计

算法	操作	计算量/步
HRDQN（本文算法）	双策略网络前向	137 280
	双目标网络前向	137 280
	双经验采用	16 384
	双反向传播	411 840
	双优先级更新	1 024
	双噪声重置	3 840
	双目标网络更新	98 304
	总计	1 003 952
RDQN	策略网络前向	128 640
	目标网络前向	128 640
	经验采样	16 384
	反向传播	385 920
	优先级更新	512
	噪声重置	2 560
	目标网络更新	65 536
	总计	728 192
PPO	Actor前向	98 304
	Critic前向	65 536
	经验存储	256
	GAE计算	1 024/512
	策略更新	1 310 720/512
	总计	166 694

2.3.3 性能与代价分析

HRDQN算法的分层结构降低了策略搜索复杂

度。设扁平结构的动作空间维度为$|A_{\text{flat}}|=|F||M||P|$，其中，$|F|$为频率维度，$|M|$为调制维度，$|P|$为功率维度。而分层结构将其解耦为两个低维子空间，动作空间维度降低为$|A_{\text{hier}}|=|A_{\text{high}}|+|A_{\text{low}}|=|F|+|M||P|$。理论上分层结构相比扁平结构，可以提升收敛速度的倍数为

$$\beta=\frac{|A_{\text{flat}}|}{|A_{\text{hier}}|}=\frac{|F||M||P|}{|F|+|M||P|} \tag{13}$$

本文算法所采用的分层结构在带来显著性能提升的同时，也引入了代价。首先，分层决策过程存在策略振荡与局部次优的风险，由于高低层策略需协同更新，任何一方的策略漂移都可能使整个系统学习不稳定。其次，分层结构直接导致单步计算复杂度的提升，维护与训练两套网络及经验回放池增加了计算与内存开销。此外，模型结构的复杂化不可避免地带来了超参数数量的增多，高低层网络的学习率、更新频率、经验池容量以及高层决策周期等参数相互耦合，显著增加了算法的调参难度。

针对策略振荡与次优风险，通过奖励函数解耦与异步更新机制予以有效缓解。奖励函数解耦为高层策略赋予频点跟踪的内在奖励，使其学习目标与低层策略的干扰效果优化相分离，从根本上降低了两层策略的目标冲突。异步更新机制则允许低层策略高频自适应环境细节变化，而高层策略低频稳定地学习宏观战略，此种差异化的节奏作为内在稳定器，有效保障了训练过程的平稳收敛。对于单步计算复杂度的增加，分层结构所带来的探索效率与收敛速度的极大提升构成了有效的抵消，从而在实际训练时间上反而可能更具优势，这一点将在仿真分析中加以验证。

综上，HRDQN 算法通过引入分层结构，以增加单步计算复杂度和超参数调试难度为代价，换取了收敛速度的显著提升和总训练成本的降低，这一权衡在实际应用中是合理且可接受的。

3　仿真分析

在本文所构建的场景中，状态与动作均被假定为离散值。考虑通信频率为400~420 MHz，可用频点为10个，定义为1~10的离散值$f_c,f_j\in\{1,2,\cdots,10\}$，通信调制方式和干扰调制方式均用 1~3 的离散值表示，即 BPSK、QPSK 和 16QAM 表示为$m_c,m_j\in\{1,2,3\}$，通信与干扰的最大功率设置为 30 dBm，将可调功率定义为 1~10 的离散值$P_c,P_j\in\{1,2,\cdots,10\}$。通信和干扰距离考虑取[500,1 500] m 的随机值，信道考虑瑞利衰落、路径损耗与阴影衰落。

在实际应用中，当通信方遭受干扰时，会通过调整通信参数实现抗干扰效果，常规手段为变更通信频率。为模拟对抗性电磁环境，考虑通信方具有一定抗干扰能力，通过设定通信方周期性及受扰后触发式的频率跳变，模拟了一个动态的跳频通信目标。当通信方的接收 BER>0.1 时，会被动调整通信频率以躲避干扰攻击，同时每 200 步会主动调整一次通信频率。仿真参数如表 2 所示。

表 2　　仿真参数

参数类型	参数名称	参数值
通信参数	最大功率/dBm	30
	天线增益/dBi	10
	调制方式	BPSK/QPSK/16QAM
	编码方式	$\frac{1}{2}$卷积码
	最大距离/km	1.5 (m=1,n=3.5)
	总频段/MHz	400~420
	可选频点数	10
	跳频策略	200 步/BER>0.1
信道参数	信道类型	Nakagami-m 信道
	瑞利信道 m 值	1
	莱斯信道 m 值	4
	路径损耗类型	对数距离-n 模型
	自由空间 n 值	2
	阴影衰落标准差/dB	4
	噪声功率谱密度/(dBm·Hz^{-1})	−174
干扰参数	最大功率/dBm	30
	调制方式	BPSK/QPSK/16QAM
	最大距离/km	1.5 (m=1,n=4)

3.1　JSR 与 BER 关系曲线

图 7 给出了 SNR 为 15 dB 时不同调制方式下接收端 JSR 与 BER 的理论关系曲线。从图 7 可以看出，当 BER 达到 0.1 时，BPSK 调制的压制系数门限为 α = 1 dB，QPSK 调制的压制系数门限为 α = −2 dB，16QAM 调制的压制系数门限为 α = −8 dB。

BPSK 因为只有 2 种相位状态，信号区分简单，抗干扰能力最强；QPSK 有 4 种相位状态，抗干扰能力次之；16QAM 有 16 种相位状态，信号区分复杂，抗干扰能力最弱。因此，不同调制方式的压制系数门限与不同调制方式的特性相符。

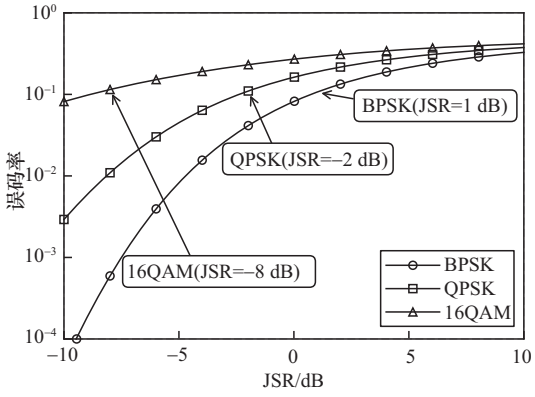


图 7　不同调制方式的 JSR 与 BER 理论关系曲线

3.2　HRDQN 算法性能分析

图 8 展示了 HRDQN 算法在单次训练中奖励值变化的原始曲线与平滑曲线，其中原始曲线用于展示算法实际性能，平滑曲线用于观察算法总体走势及平均性能。从图 8 可以看出，HRDQN 算法原始曲线呈现明显波动，这是因为采用了对抗性的通信抗干扰方式和时变的信道同时作用，而从平滑曲线来看，HRDQN 算法初期处于探索阶段，奖励处于较低状态。由于其分层动作决策机制，能够迅速学习并在 5 000 步左右完成快速收敛，说明 HRDQN 算法在实现干扰决策上具有有效性。

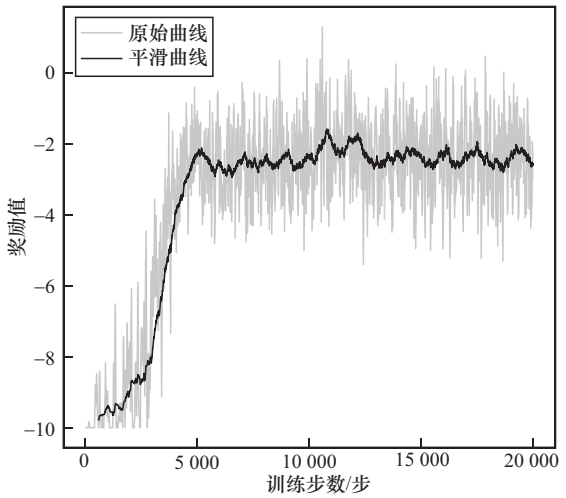


图 8　HRDQN 算法单次训练奖励值曲线

图 9 展示了不同学习率下的算法敏感性，将智能体与环境交互的次数定义为训练步数。从图 9 可以看出，本文算法对学习率相对敏感，采用 0.000 1 学习率收敛速度过慢，而采用 0.000 5 学习率虽然收敛速度更快，但是长期策略变得不稳定，出现多处明显波动。因而将学习率设置为 0.000 3 属于合理范畴。

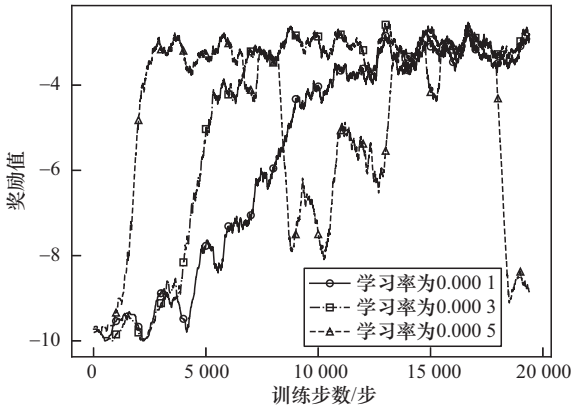


图 9　不同学习率曲线

图 10 展示了不同折扣因子下的算法敏感性。从图 10 可以看出，本文算法针对折扣因子相对不敏感，采用不同折扣因子对算法整体性能的影响并不显著，这是因为通信环境的动态性与奖励结构的设置，导致长期回报影响微弱，干扰决策算法更注重短期收益，即当前干扰动作是否有效。因此，当算法有效视野发生变化时，对整体训练效果影响并不显著。

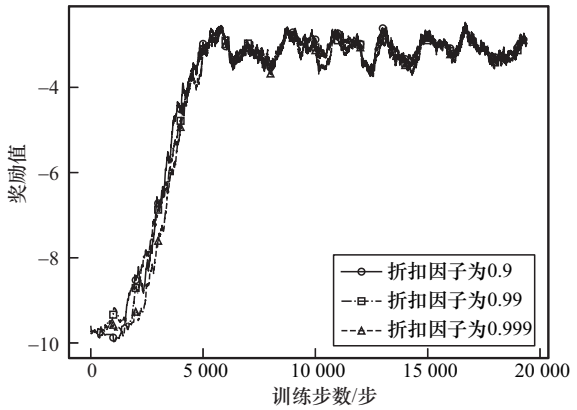


图 10　不同折扣因子曲线

3.3　智能干扰决策效果对比

为了验证本文算法的有效性，将其与原始 RDQN 和 PPO 算法进行对比。本文算法和 RDQN

算法根据经验选择了超参数的初步值，PPO算法的部分初值参数参考文献[17]，并在仿真过程中对3种算法的超参数进行了进一步的调优，RDQN与PPO算法外部奖励函数的设置与本文算法的干扰效果奖励函数相同。算法超参数设置如表3所示。

表3　　算法超参数设置

算法	参数类型	数值
HRDQN（本文算法）	学习率	3×10^{-4}
	折扣因子	0.99
	经验池容量	40 000
	训练批次大小	128
	软更新参数	0.005
RDQN	学习率	3×10^{-4}
	折扣因子	0.99
	经验池容量	40 000
	训练批次大小	128
	软更新参数	0.005
PPO	学习率	3×10^{-4}
	折扣因子	0.99
	熵奖励系数	0.01
	GAE参数	0.95
	更新间隔步数	512

对于RDQN算法，学习率决定了神经网络参数更新的步长，折扣因子用于权衡即时奖励与长期回报的重要性，经验池容量影响了用于训练的数据的多样性和相关性，训练批次大小关系到每次参数更新时的梯度估计稳定性，软更新参数则控制了目标网络参数更新的平滑程度。对于PPO算法，其熵奖励系数用以鼓励策略探索，广义优势估计（generalized advantage estimation, GAE）参数在优势估计中平衡了偏差与方差，更新间隔步数定义了收集多少新样本后才进行一次策略更新。

为方便观察算法特性及对比结果，预先生成随机信道信息参数，将3种算法在相同信道状态下进行了决策对比，并将对比结果进行取相同窗口计算均值，达到平滑曲线目的。

3.3.1　奖励值对比

图11展示了HRDQN算法与原始RDQN和PPO算法在通信对抗环境中执行干扰决策的奖励值对比。在此场景下，算法需额外适应动态环境。尽管如此，HRDQN算法在继承RDQN算法的诸多优点外，凭借动作分层机制展现出更快的探索效率。仿真结果显示，HRDQN算法在约5 000步即可收敛至最优决策动作，同时能够保持算法决策效果稳定，这得益于其高低层采用的不同决策奖励机制。RDQN算法由于扁平动作空间，在约30 000步时逼近收敛，但其在接近收敛时振荡较为剧烈，这是因为越是接近收敛，其算法经验池中的用以回放的具有一定优先级的经验相对变少，因此呈现出振荡趋势。而PPO算法虽然计算复杂度较低，但是由于采用同策略（On-Policy）机制，在复杂环境中采样开销巨大，导致在面对高维空间时收敛速度较慢，在近45 000步时才收敛。观察到奖励值最终收敛至−3左右，这是因为通信方的频率跳变机制导致干扰决策在得到较高奖励时，会触发通信方跳频，导致下一次或几次的决策奖励值较低，因此计算得到的平均奖励值会远低于所设计奖励函数的最大值。实验结果验证了HRDQN算法在通信对抗环境中具有更优的探索效率与收敛速度。

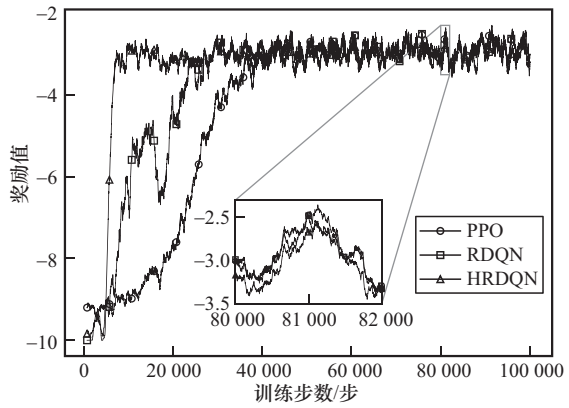


图11　算法奖励值对比

3.3.2　JSR对比

图12展示了3种算法在通信方接收端的JSR对比结果。从收敛结果为−2 dB可知，本次仿真实验通信方采用了QPSK调制方式。对比图11的奖励值结果可以看出，采用相同窗口平滑后，收敛后JSR的波动要比奖励值小，这是因为奖励函数的设计，使JSR越接近α时，奖励变化越剧烈。HRDQN算法因其动作分层机制，相比其他两种算法在探索效率与收敛速度上具有显著优势。RDQN算法因其PER机制，在即将收敛时出现振荡现象。PPO算法

因其On-Policy机制，收敛速度明显慢于RDQN算法。3种算法均能收敛到QPSK调制的α处，说明3种算法均能实现在完成一定的干扰目的的同时，最小化干扰功耗，HRDQN算法在探索效率和收敛速度上有显著优势。

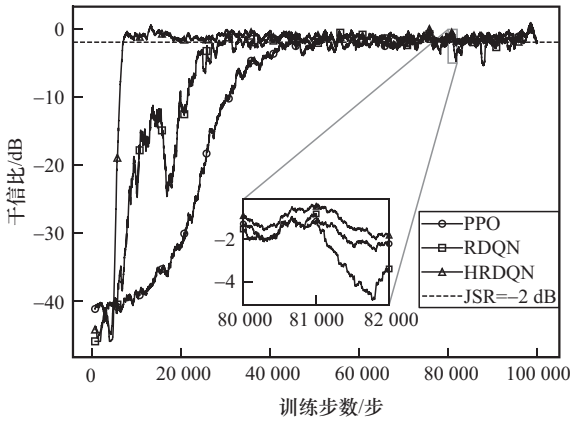


图12　接收端JSR对比

3.3.3　BER对比

图13展示了3种算法在通信对抗环境中的通信方接收端BER对比情况，以观察干扰效果。在面对具有一定抗干扰能力的通信系统时，由于通信方会按照一定的规律进行频率的跳变，同时无线信道具备时变特性，3种算法收敛时最终结果仍呈现出一定的波动，但都集中在0.10附近，且平均BER大于0.10，达到了既定干扰目标。HRDQN算法由于动作分层机制，能够更快速地收敛至最优结果，这意味着在面对复杂多变的通信环境时，HRDQN算法能够更快速地提供有效干扰决策。实验数据表明，HRDQN算法在通信对抗环境中，相比RDQN与PPO算法，在收敛速度与探索效率上有显著优势。

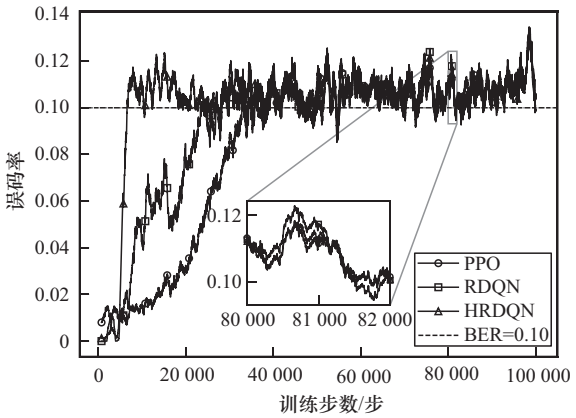


图13　接收端BER对比

3.3.4　收敛速度对比

收敛速度是评估算法效率的核心指标。在单步计算开销相近的情况下，收敛所需的训练步数可直接衡量收敛速度。然而，当算法架构差异导致单步计算量显著不同时，应以达到收敛状态所需的总计算量作为评判收敛速度的依据，收敛速度与总计算量成反比。

从图11仿真结果观察到，HRDQN算法仅需约5 000步即可收敛，远低于RDQN算法的30 000步与PPO算法的45 000步，展现了其在探索效率上的优势。结合表4中的计算量统计进行综合分析时，可计算得到HRDQN、RDQN与PPO算法的总计算量之比约为1:4.4:1.5。

表4　　计算量统计对比

算法	单步计算量/(FLOPS·步$^{-1}$)	收敛步数/步	总收敛计算量/FLOPS
HRDQN（本文算法）	1 003 952	5 000	5.01×10^{9}
RDQN	728 192	30 000	2.18×10^{10}
PPO	166 694	45 000	7.50×10^{9}

结果表明，HRDQN算法的收敛速度最快。尽管其分层结构引入了额外开销，导致单步计算量高于RDQN算法，但决策效率的显著提升使其总计算量远低于对比算法。PPO算法虽然单步计算量很低，但是其On-Policy机制导致样本效率低下，收敛所需步数最多，且参数调整需要较高时延，无法即时调整参数。而HRDQN算法虽然牺牲部分单步计算复杂度，但是却使收敛所需总计算量降低，收敛速度大幅度提升，这对智能干扰的实时性要求有利。

4　结束语

本文提出了一种基于HRDQN的智能干扰决策算法，通过模型构建、算法设计和仿真分析，验证了该算法的有效性。仿真结果表明，在通信对抗环境中，本文算法相较RDQN与PPO算法具有更高的探索效率与更快的收敛速度，能快速适应复杂电磁环境变化并选择最优干扰策略。同时，本文算法通过功率约束机制能够在保证干扰效果的前提下有效降低功耗，为隐蔽型微平台通信对抗系统提供了创新性解决方案。

参考文献：

[1] 姚富强 . 通信抗干扰工程与实践[M]. 3 版 . 北京: 电子工业出版社, 2025.
Yao F Q. Communication anti-jamming engineering and practice[M]. 3rd ed. Beijing: Publishing House of Electronics Industry, 2025.

[2] Zhang Y J, Huo W B, Zhang C, et al. Smart noise jamming power adjustment using exploratory deep deterministic policy gradient[C]//Proceedings of the 2023 IEEE Radar Conference (RadarConf23). Piscataway: IEEE Press, 2023: 1-6.

[3] 韩晨, 刘爱军, 牛英滔, 等 . 智能干扰: 目的、方式、反馈[J]. 指挥与控制学报, 2022, 8(2): 133-140.
Han C, Liu A J, Niu Y T, et al. Smart jamming: purpose, method and feedback[J]. Journal of Command and Control, 2022, 8(2): 133-140.

[4] 宋佰霖, 许华, 齐子森, 等 . 一种基于深度强化学习的协同通信干扰决策算法[J]. 电子学报, 2022, 50(6): 1301-1309.
Song B L, Xu H, Qi Z S, et al. A collaborative communication jamming decision algorithm based on deep reinforcement learning[J]. Acta Electronica Sinica, 2022, 50(6): 1301-1309.

[5] Zhuan S S, Yang J A, Liu H, et al. A novel jamming strategy-greedy bandit[C]//Proceedings of the 2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN). Piscataway: IEEE Press, 2017: 1142-1146.

[6] Yang H J, Shi M, Xia Y Q, et al. Security research on wireless networked control systems subject to jamming attacks[J]. IEEE Transactions on Cybernetics, 2019, 49(6): 2022-2031.

[7] 许华, 宋佰霖, 蒋磊, 等 . 一种通信对抗干扰资源分配智能决策算法[J]. 电子与信息学报, 2021, 43(11): 3086-3095.
Xu H, Song B L, Jiang L, et al. An intelligent decision-making algorithm for communication countermeasure jamming resource allocation[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3086-3095.

[8] 张君毅, 张冠杰, 杨鸿杰 . 针对未知通信目标的干扰策略智能生成方法研究[J]. 电子测量技术, 2019, 42(16): 148-153.
Zhang J Y, Zhang G J, Yang H J. Research on intelligent interference strategy generation method for unknown communication target[J]. Electronic Measurement Technology, 2019, 42(16): 148-153.

[9] Han B W, Yang X P, Wu X C, et al. Smart noise jamming suppression method based on fast fractional filtering[J]. The Journal of Engineering, 2019(19): 6201-6205.

[10] Shi Y, Davaslioglu K, Sagduyu Y E. Generative adversarial network in the air: deep adversarial learning for wireless signal spoofing[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 7(1): 294-303.

[11] 饶宁, 许华, 宋佰霖 . 融合动作剔除的深度竞争双Q网络智能干扰决策算法[J]. 空军工程大学学报(自然科学版), 2021, 22(4): 92-98.
Rao N, Xu H, Song B L. An intelligent jamming decision algorithm based on action elimination dueling double deep Q network[J]. Journal of Air Force Engineering University (Natural Science Edition), 2021, 22(4): 92-98.

[12] 饶宁, 许华, 宋佰霖 . 融合有效方差置信上界的Q学习智能干扰决策算法[J]. 哈尔滨工业大学学报, 2022, 54(5): 162-170.
Rao N, Xu H, Song B L. Q-learning intelligent jamming decision algorithm based on efficient upper confidence bound variance[J]. Journal of Harbin Institute of Technology, 2022, 54(5): 162-170.

[13] Wang L G, 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] 刘铮, 冯永新, 钱博 . 基于DQN的跳频信号干扰决策方法研究[J]. 沈阳理工大学学报, 2023, 42(3): 9-15.
Liu Z, Feng Y X, Qian B. Research on frequency hopping signal interference decision method based on DQN[J]. Journal of Shenyang Ligong University, 2023, 42(3): 9-15.

[15] 陆永安, 陈杰豪, 张琪露, 等 . 基于全并行深度Q网络的通信干扰资源快速分配算法[J]. 现代电子技术, 2024, 47(13): 47-54.
Lu Y A, Chen J H, Zhang Q L, et al. Communication jamming resource fast allocation algorithm based on fully parallel deep Q-network[J]. Modern Electronics Technique, 2024, 47(13): 47-54.

[16] Zhang C D, Yang B, Wang L, et al. A cognitive jamming decision-making method based on heuristic improved A2C algorithm[J]. IEEE Transactions on Vehicular Technology, 2025, 74(2): 2871-2883.

[17] 张静克, 杨凯, 李超, 等 . 基于先验知识嵌入LSTM-PPO模型的智能干扰决策算法[J]. 通信学报, 2024, 45(12): 227-239.
Zhang J K, Yang K, Li C, et al. Intelligent interference decision algorithm with prior knowledge embedded LSTM-PPO model[J]. Journal on Communications, 2024, 45(12): 227-239.

[18] 周成, 林茜, 马丛珊, 等 . 通信干扰信道和功率智能决策算法[J]. 电子与信息学报, 2024, 46(10): 3957-3965.
Zhou C, Lin Q, Ma C S, et al. Intelligent decision-making for selection of communication jamming channel and power[J]. Journal of Electronics & Information Technology, 2024, 46(10): 3957-3965.

[19] Tan H C, Wei P, Xiao S, et al. Personalized recognition for distributed jamming in dynamic environments[J]. IEEE Wireless Communications Letters, 2024, 13(12): 3603-3607.

[20] Liu M Q, Liu Z L, Lu W D, et al. Distributed few-shot learning for intelligent recognition of communication jamming[J]. IEEE Journal of Selected Topics in Signal Processing, 2021, 16(3): 395-405.

[21] Feng Z B, Xu Y H, Jiao Y T, 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.

[22] 刘全, 翟建伟, 章宗长, 等 . 深度强化学习综述[J]. 计算机学报, 2018(1): 1-27.
Liu Q, Zhai J W, Zhang Z C, et al. Summary of deep reinforcement learning[J]. Chinese Journal of Computers, 2018(1): 1-27.

[23] Hessel M, Modayil J, Hasselt H V, et al. Rainbow: combining improvements in deep reinforcement learning[J]. Proceedings of the AAAI

Conference on Artificial Intelligence, 2018, 32(1): 3215-3222.

[24] 王世练, 骆俊杉, 魏鹏, 等. 认知通信抗干扰[M]. 北京: 国防工业出版社, 2023.

Wang S L, Luo J S, Wei P, et al. Cognitive communication for anti-jamming[M]. Beijing: National Defense Industry Press, 2023.

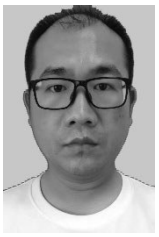
[作者简介]

刘天一（1993– ），男，黑龙江哈尔滨人，中国人民解放军 63861 部队工程师，主要研究方向为通信对抗、可重构智能表面。

吴宣利（1980– ），男，黑龙江哈尔滨人，博士，哈尔滨工业大学教授、博士生导师，主要研究方向为超密集网络、可重构智能表面、物理层安全等。

许涛（2000– ），男，贵州毕节人，哈尔滨工业大学博士生，主要研究方向为可重构智能表面、加权分数阶傅里叶变换、物理层安全。

王吉彬（1982– ），男，辽宁凤城人，中国人民解放军 63861 部队工程师，主要研究方向为移动通信、通信抗干扰。



李广华（1986– ），男，河南漯河人，中国人民解放军 63861 部队工程师，主要研究方向为卫星通信、通信抗干扰。