DOI:10.20079/j.issn.1001-893x.240715002

基于多奖励值 DDQN 智能通信抗干扰决策方法*

凌 耀[1,2],谢世珺[2],梁 豪[2],冯 姣[1],高伟杰[1,2]

(1.南京信息工程大学 电子与信息工程学院,南京 210044;2.国防科技大学第六十三研究所,南京 210007)

摘 要:在动态干扰环境下的卫星通信系统中,各信道的质量和干扰功率存在差异。有限的频谱资源和复杂的干扰环境对抗干扰通信决策提出了资源分配和业务需求的挑战,即如何在避开干扰频率和优化功率的同时,实现资源的高效利用。为解决这一问题,提出了一种基于多奖励值函数的深度强化学习抗干扰算法。该算法将发送方、接收方与干扰方之间的交互建模为马尔可夫决策过程。通过优化信道切换与功率切换代价的奖励函数,引入频率切换与功率切换机制,分析相邻时隙频谱中的干扰特征,并将交互过程中采集到的干扰信号特征与信道信息结合,用于训练抗干扰策略。该策略实现了频率域与功率域的联合抗干扰决策。仿真结果表明,该算法能够有效降低系统的受干扰概率,加快算法收敛速度,并优化功率资源的利用效率。

关键词:智能通信抗干扰;联合抗干扰决策;深度强化学习;多奖励值函数

开放科学(资源服务)标识码(OSID):

中图分类号:TN973 **文献标志码**:A **文章编号**:1001-893X(2025)11-1820-08

An Intelligent Communication Anti-interference Decision Algorithm Based on Multiple Reward Value DDQN

LING Yao[1,2], XIE Shijun[2], LIANG Hao[2], FENG Jiao[1], GAO Weijie[1,2]

(1. School of Electronic and Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China; 2. The 63rd Research Institute, National University of Defense Technology, Nanjing 210007, China)

Abstract: In satellite communication systems operating in dynamic interference environments, the quality of channels and the interference power vary. Limited spectrum resources and complex interference environments pose challenges for anti-interference communication decisions, particularly in terms of resource allocation and service demands. Specifically, the challenge lies in efficiently utilizing resources while avoiding interference frequencies and optimizing power. To address this issue, a deep reinforcement learning-based anti-interference algorithm with multiple reward functions is proposed. The algorithm models the interaction between the transmitter, receiver, and interferer as a Markov decision process. By optimizing the reward function associated with the costs of channel and power switching, it introduces mechanisms for both frequency and power switching, analyzes the interference characteristics in the spectrum of adjacent time slots, and integrates the interference signal features collected during the interaction with channel information to train an anti-interference strategy. This strategy enables joint anti-interference decision-making in both the frequency and power domains. Simulation results demonstrate that the algorithm effectively reduces the probability of interference, accelerates convergence, and optimizes the utilization of power resources.

Key words: intelligent communication anti-interference; joint anti-interference decision; deep reinforcement learning; multiple reward value functions

* **收稿日期**:2024-07-15;**修回日期**:2024-11-08
基金项目:国家自然科学基金资助项目(62201596);国防科技大学学校科研计划资助项目(ZK22-45)
通信作者:谢世珺 Email:xsjxsj_520@163.com

0　引　言

卫星通信作为一种空间信息基础设施,具有传播距离远、覆盖范围广、部署速度快、不受地理环境限制和可用通信频带宽等特点,被广泛应用于军事、水利和远洋航行等领域[1]。在卫星通信网络中,空间段、地面段和用户段采用电磁波作为信息传输的媒介,并在暴露的无线信道中进行传输。这使得干扰设备可以轻易对通信链路进行干扰,因此在恶劣电磁环境中实现抗干扰的可靠传输成为一个重要的研究课题。随着软件无线电、人工智能等技术的飞速发展,抗干扰正向着智能抗干扰方向发展[2-3]。深度学习、强化学习等技术作为实现智能抗干扰的有效手段,得到了大量研究。

文献[4]提出了一种基于深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)的频率间隔选择和跳频设置算法,在干扰环境下找到最优的频率间隔和跳频策略,以最大限度地提高信噪比,并将参数决策问题表述为马尔可夫决策[5]过程(Markov Decision Process, MDP)。文献[6]则对动态干扰环境中的跳频通信系统进行建模,针对多信道干扰的情况下发射机和接收机未知干扰模式,提出了一种深度强化学习(Deep Q Network, DQN)跳频算法来解决跳频决策问题,实现抗干扰。文献[7]通过随机博弈[8]的方法解决干扰问题,提出了一种强化学习(Q-learning)算法,获得干扰机和目标发射机的智能信道跳频序列,高发射功率目标用户利用智能信道跳变来迫使低发射功率用户同时使用目标用户未跳变的信道来混淆干扰器,低发射功率目标用户通过信道跳变来避开干扰器。

然而,上述文献大多只考虑了频率域或功率域的单域干扰,没有考虑到实际信道环境中存在的频率域和功率域的联合干扰[9-10],因此,这些算法通常只能针对单个干扰域制定抗干扰策略。为了提高发送方的通信质量,一方面需要躲避干扰频率,另一方面在遭遇干扰功率时需要以最优的资源效率适当提高发射功率来克服干扰。本文的抗干扰决策算法在考虑功率域和频率域联合抗干扰的同时,优化了资源利用。

本文主要贡献如下:

1)研究了在动态频率域和功率域联合干扰下的抗干扰策略,结合干扰的二维拓扑图建立了马尔可夫抗干扰决策模型,设计了相应的状态集、动作集和代价奖励函数。

2)提出了一种多奖励值函数的深度强化学习(Multiple Reward Value-Double Deep Q Network, MRV-DDQN)抗干扰算法。与以往的 DQN 抗干扰算法不同,该算法将信干噪比、功率切换代价和频率切换代价等信息分别作为多奖励值函数,通过奖励值函数引导算法以最少的频率切换和最小的功率开销,实现频率域和功率域的联合抗干扰及资源的合理利用。

1　系统模型

1.1　抗干扰通信系统模型

如图 1 所示,本文研究发送方向接收方发送数据时的抗干扰通信方法,构建了一个由发送方、接收方、若干干扰方和抗干扰决策代理组成的通信系统模型。该模型中,前向通信链路包括多个可用信道。干扰方在每个通信时隙内对多个前向链路信道进行干扰。抗干扰决策代理根据每个通信时隙内信道受到干扰的情况,对下一通信时隙的频率切换和功率切换[11-13]动作做出决策,并通过未被干扰的通信决策控制链路通知发送方。本文仅对单向通信链路的抗干扰进行研究,并假设信息反馈控制链路中没有干扰存在。发送方在下一传输时隙执行该决策,以最优的子带和功率通过前向链路发送信息。

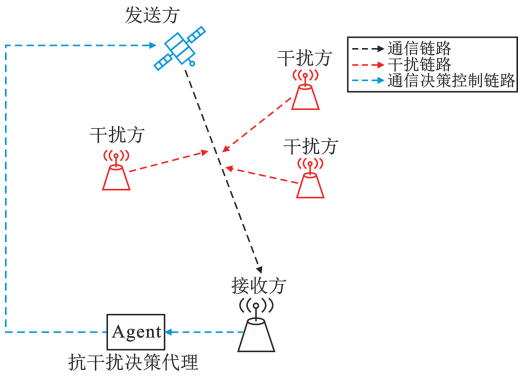


图 1　抗干扰通信系统示意

图 2 模拟了前向通信链路中的干扰,横坐标代表系统通信时隙,纵坐标代表通信链路按频率从低到高划分的若干信道,频率中心为$\{f_1, f_2, f_3, \cdots, f_N\}$。若干随机分布的干扰机产生 N 种不同功率的干扰,干扰随时隙和频率变化的二维拓扑图如图 2 所示。发送方在传输数据时,根据通信链路的动态干扰情况选择相应的频率和功率切换策略,从而实现抗干扰。抗干扰策略的最优目标是通过最少的频率切换

和最小的发送功率开销进行通信。

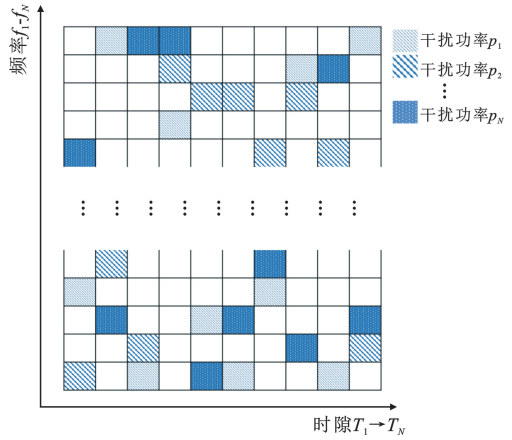


图 2 干扰动态变化二维拓扑

1.2 状态及动作建模

状态空间定义为 $S_t=[f_n,t]$,表示在第 t 个传输时隙占用第 n 个信道,$n\in[1,N]$,$t\in[1,T]$。因此,S_t 在时间和频率维度上具有 $N\times T$ 个的可能状态,并且每个状态之间相互独立。由于信道中存在干扰功率,在抗干扰通信系统中,发送机需要根据抗干扰决策代理发送的频率切换动作,决定是否在当前信道上继续传输或跳转到其他 $N-1$ 个信道之一,同时根据功率切换动作决定当前的发射功率。将频率切换动作和功率切换动作定义为二维元组 $\boldsymbol{A}=[A_{Fn},A_{Pn}]$,$A_{Fn}$ 表示跳转到第 n 个子带,A_{Pn} 表示发射功率切换为 P_n。$A_t\in A$ 是第 t 个时隙中发送方采用的频率和功率切换动作。

1.3 奖励值函数建模

奖励值函数是引导算法实现频率域和功率域联合抗干扰的重要参数,在本节中建立了奖励初值、信噪比奖励值、频率及功率切换代价奖励值和功率切换成功奖励值。算法训练完成后的奖励值也作为该抗干扰决策算法性能的评价标准。

为了便于仿真,设置了奖励初值 $\text{reward}_{\text{origin}}$,用于表示没有进行频率切换且子带没有干扰功率时的奖励值。接收方的信干噪比和传输速率分别表示为

$$\text{SINR}=\frac{g_{\text{u}}P_{\text{U},t}}{n+\sum_{j=1}^{N}g_jp_{j,t}\delta(f_{\text{j},t}=f_{\text{U},t})} \tag{1}$$

$$C_{\text{U}}=B_{\text{U}}\text{lb}\left(1+\frac{g_{\text{u}}P_{\text{U},t}}{n+\sum_{j=1}^{N}g_jp_{j,t}\delta(f_{\text{j},t}=f_{\text{U},t})}\right) \tag{2}$$

式中:$f_{\text{U},t}$ 为发送方在 t 时刻的发送频率,发射功率为 $P_{\text{U},t}$;$f_{\text{j},t}$ 为干扰方在 t 时刻选择干扰的通信频率,干扰功率为 $P_{\text{j},t}$;n 为信道噪声功率;g_{u} 表示发送方到接收方之间的信道增益;g_{j} 表示干扰方和接收方之间的信道增益;B_{U} 是用户发送信道带宽;C_{U} 是发送速率。发送方如果选择和干扰方相同的频率即 $(f_{\text{j},t}=f_{\text{U},t})$,则 $\delta(f_{\text{j},t}=f_{\text{U},t})=1$,否则 $\delta(f_{\text{j},t}=f_{\text{U},t})=0$。为了表征接收端信号质量的好坏,将信干噪比建立为奖励值函数 $\text{reward}_{\text{SINR}_n}$,根据式(1)计算不同信干噪比设置 $\text{reward}_{\text{SINR}_n}$,表示为

$$\text{reward}_{\text{SINR}_n}=\begin{cases}\text{reward}_{\text{SINR}_1}, & p=p_1\\ \text{reward}_{\text{SINR}_2}, & p=p_2\\ \vdots & \\ \text{reward}_{\text{SINR}_n}, & p=p_n\end{cases} \tag{3}$$

为了衡量功率切换和频率切换对通信过程的影响,将功率域和频率域的干扰量化为抗干扰决策算法中的奖励值函数。系统在频率切换后需要重新同步通信链路,因此产生频率切换代价,根据频率切换代价设置奖励值 reward_F,当频率没有发生切换时则代价为 0;发射功率大于干扰功率 6 dBm 以上即功率域抗干扰成功,根据功率切换成功代价设置奖励值 reward_{P_s},否则奖励值为 0。

当选择有干扰功率的信道时,发射功率会根据信道中的干扰功率进行提高。在发射功率提高的过程中会产生功率切换代价。根据不同的干扰功率,设置了功率切换代价奖励值 reward_{P_n},表示为

$$\text{reward}_p=\begin{cases}\text{reward}_{p_1}, & p=p_1\\ \text{reward}_{p_2}, & p=p_2\\ \vdots & \\ \text{reward}_{p_n}, & p=p_n\end{cases} \tag{4}$$

这样的多奖励值函数机制控制 Q 值的更新以及神经网络权值的优化,引导算法尽可能切换到空闲信道或干扰功率低的信道,通过发射功率的切换保证传输的可靠性。抗干扰策略的目标是最大化长期累积奖励值,使得发送方能在最小的频率切换和功率切换的代价下实现抗干扰。

1.4 状态动作值建模

在本文中将状态动作值定义为 $Q(S,A)$,表示在当前状态 S 经过神经网络计算输出的 Q 值元组,Q 值元组的索引位置与动作所在集合的索引位置相对应。在 MRV-DDQN 算法中,Q 值元组的更新表示为

$$\begin{aligned}Q^*(S,A)=&\ \text{reward}_{\text{orgin}}+\text{reward}_{\text{SINR}_n}+\text{reward}_F+\\ &\text{reward}_{P_s}+\text{reward}_{p_n}+\\ &\gamma Q(S_{t+1},\underset{A}{\text{argmax}}\,Q(S_{t+1},A;\theta);\theta')\end{aligned} \tag{5}$$

式中:$Q^*(S,A)$是状态 S_t 和 A 对应的状态动作更新值;γ 是折扣因子,在 0~1 范围之间;θ 为目标网络权值;θ'是策略网络权值;$Q(S_{t+1},\underset{A}{\operatorname{argmax}}\ Q(S_{t+1},A;\theta);\theta')$中先由$\underset{A}{\operatorname{argmax}}\ Q(S_{t+1},A;\theta)$取得目标网络中最大 Q 状态动作值对应的动作 A_{t+1},再根据下一时刻的 S_{t+1}、A_{t+1} 动作更新策略网络的 Q 值。

本节对抗干扰策略的状态、动作、奖励值和状态动作值进行建模,为下一节的算法实现提供理论基础。通过不断优化 Q 状态动作值,确定最优抗干扰策略的问题可以等效于发射机在每个时隙选择最优动作。

2　基于多奖励值的 DDQN 智能抗干扰算法

2.1　基于 MRV-DDQN 抗干扰算法框架

网络结构如图 3 所示,由输入、输出层和两个隐藏层组成,每层包含 30 个单元。3 种层之间用权重参数 $\theta=\{\theta_{\text{Input}},\theta_{\text{Hidden}},\theta_{\text{Onput}}\}$ 全连接。特征输入层获取状态空间元组里的信道信息。隐藏层中使用 relu 激活函数如式(6):

$$\text{relu}(x)=\max(x,0) \tag{6}$$

式中:x 为每层的输出。激活函数使神经网络[14]能够拟合复杂的非线性 Q 状态动作函数值。输出层的 30 个 Q 状态动作函数值单元对应不同的频率和功率切换动作。

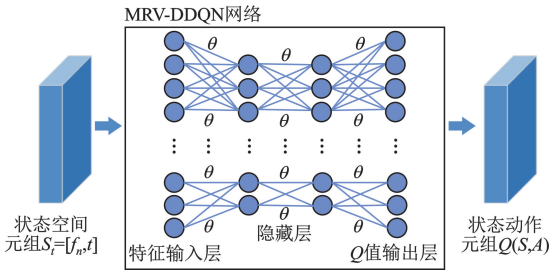


图 3　MRV-DDQN 策略神经网络的结构

基于 MRV-DDQN 抗干扰算法框架如图 4 所示。框图中展示了算法的执行过程,其中包含两张结构相同的 MRV-DDQN 网络分别是策略网络(policy net)和目标网络(target net)。

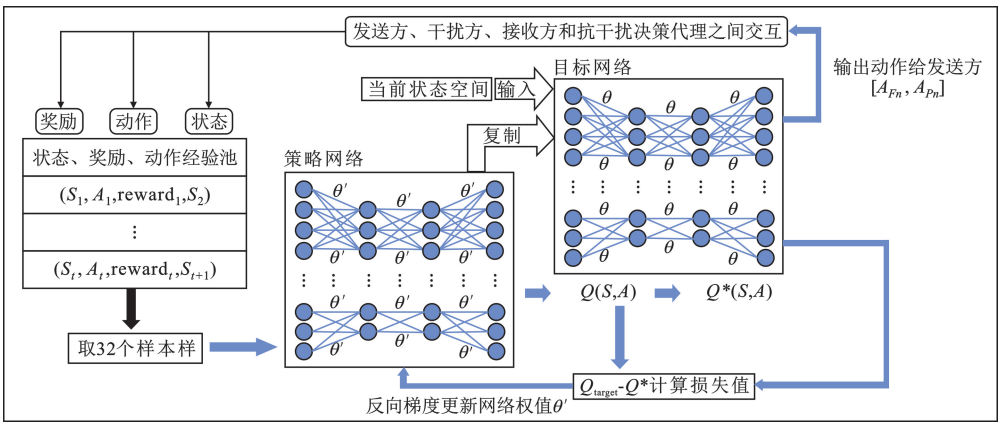


图 4　MRV-DDQN 抗干扰算法框架

在发送方、干扰方、接收方交互过程中产生的状态、奖励、动作和下一状态组成的元组保存在经验池中,训练时从经验池中周期性随机取样 32 个样本来训练策略网络。目标网络输出的 $Q_{\text{target}}(S,A)$与 $Q^*(S,A)$的差值作为损失值:

$$\text{Loss}=Q_{\text{target}}(S,A)-Q^*(S,A) \tag{7}$$

损失越小,MRV-DDQN 神经网络的收敛性越好,抗干扰策略掌握得越准确,从而输出最优的抗干扰动作。损失值通过反向梯度计算更新网络中隐藏层的权值,如式(8) 所示:

$$\theta'\leftarrow\theta'-\alpha\frac{\partial\text{Loss}}{\partial\theta'} \tag{8}$$

抗干扰策略随着网络权值的动态更新不断优化。为了加快网络收敛速度,周期地将该网络的权值复制给目标网络。在循环动态更新的过程中,目标网络作为最佳抗干扰策略的输出网络,根据 $Q(S,A)$状态动作值来获得最佳动作。

2.2　MRV-DDQN 抗干扰算法

MRV-DDQN 抗干扰算法执行流程如下:

输入:

1)初始化算法参数:各奖励值 reward,学习率 α,折扣因子 γ,贪婪策略选择概率 ε,目标网络更新周期 δ,经验池大小 η,用于训练的样本数 sample。

2)初始化环境参数:干扰功率 $P_1,P_2,\cdots,P_n$,初始化 policy 和 target 网络权值,训练回合数(episode)初始值为 0,最大值为 400。

输出:最优抗干扰动作$[A_{Fn},A_{Pn}]$。

循环:for episode = 0 to 400;

1)发送方、接收方和干扰方之间交互形成的状态动作空间存储到经验池 η 中;

2)从上述的经验池 η 中随机取出 32 样本元组输入到 policy 神经网络中得到状态动作的 $Q(S,A)$ 值;

3)根据式(7)来更新 $Q(S,A)$ 得到 $Q^*(S,A)$,并根据式(8)计算损失值 Loss;

4)根据式(8)反向梯度算法更新 policy 网络中神经网络节点的权值 θ';

5)直到损失值收敛,得到最优抗干扰动作$[A_{Fn},A_{Pn}]$,此时发送方使用最佳通信频率和发射功率。

开始之前,需要初始化算法参数和环境参数。算法参数包括第 1.3 节介绍的 5 种奖励值、用于控制网络更新幅度的学习速率 α,以及用于控制未来奖励在当前决策中权重的折扣因子 γ。在算法初期,发送方、干扰方和接收方需要相互交互,以探索信道中的干扰情况。贪婪策略的选择概率 ε 决定了算法进行探索的概率,概率越大,干扰情况统计得越全面,经验池也会越丰富。为防止算法出现 Q 值高估,设置了更新周期 δ,该参数控制策略网络向目标网络更新权重的周期数。训练样本数 sample 表示每次从经验池中提取的用于训练网络的样本数量。环境参数包括信道中设置的动态干扰($P_1,P_2,\cdots,P_n$)、初始神经网络权值和训练回合数。

在算法执行过程中,进入训练循环。首先,在算法初期,贪婪策略选择概率 ε 控制发送方、干扰方和接收方之间的交互,以探索信道中的干扰情况,并将探索得到的状态-动作对存储在经验池中。同时,算法从经验池中随机抽取 32 个训练样本输入到策略神经网络中,计算得到状态-动作值 $Q(S,A)$。根据第 1.4 节中的状态-动作值更新公式计算出 $Q^*(S,A)$,并通过计算 $Q(S,A)$ 和 $Q^*(S,A)$ 之间的差异确定算法的损失值。由于 $Q^*(S,A)$ 是依据奖励值引导的结果,相比之前的 $Q(S,A)$ 更加符合算法在频率域和功率域的抗干扰需求。通过两者差值进行反向梯度更新,优化策略网络的权重,从而增强算法的抗干扰性能。最终,随着算法的收敛,输出的抗干扰动作$[A_{Fn},A_{Pn}]$达到最优。

3 仿真与分析

3.1 场景及模型参数设置

本节对基于 MRV-DDQN 抗干扰算法的性能进行分析,并展示 MRV-DDQN 与 DQN 抗干扰算法性能的对比。在训练过程中,定义每 10 个传输时隙为一个回合,总回合数为 400 次。为了更好地比较两种算法的性能,在前向链路的 10 个信道中设置了 3 个随机分布且不同功率的干扰机对通信链路进行干扰,干扰机以 5 个时隙为周期不断切换干扰频段,对通信子带动态干扰。仿真环境基于 OpenAI 提供的 gym 库[15]。为了模拟实际通信过程中的抗干扰性能,算法中模拟了一些参数,如表 1 所示。

表 1 仿真参数

参数	值
通信频谱带宽/MHz	400
通信子带宽/MHz	40
干扰带宽/MHz	40
通信子带数	10
每回合时隙数	10
仿真干扰功率 p_1,p_2,p_3/dBm	10,19,26
子带间隔/MHz	40

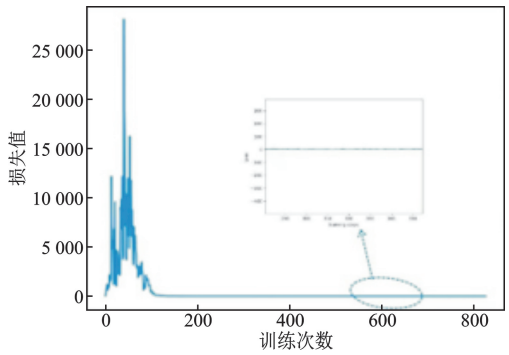
在表 2 中,参数 δ 为策略网络向目标网络复制权重的周期数。为了避免 MRV-DDQN 算法的 Q 值高估,目标网络的更新周期应适当延长。折扣因子 γ 决定未来奖励在当前决策中的权重,设为 0.9 能确保奖励有效引导算法输出最优策略。贪婪因子 ε 决定算法运行初期的探索力度,以积累足够的数据存储在经验池中,便于神经网络的训练。本文将经验池大小设为 2 000,每次从中提取 32 个样本用于训练,能够满足算法需求。初始奖励值设为正值,在频率或功率切换时将奖励值设为负值,以增强算法的抗干扰性能。学习速率控制参数更新幅度,速率过大可能导致参数过度更新和发散;过小则会减缓模型的收敛速度,延长训练时间。在调试过程中,逐步增大学习速率,最终确定 0.1 为合理值,算法的收敛效果较好。

表 2　算法参数

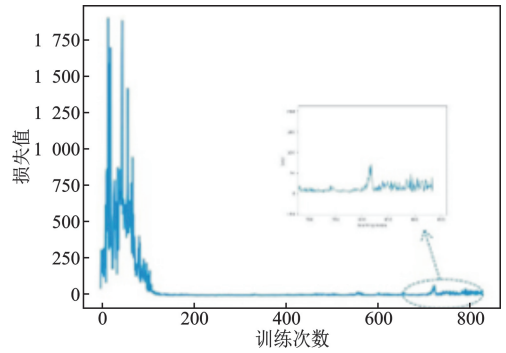
参数	值
训练回合数 episode	400
更新目标网络的周期 δ	200 个频率切换时隙
折扣因子 γ	0.9
贪婪因子 ε	0.8
经验池大小 η	2 000
初始奖励值$\text{reward}_{\text{origin}}$	10
信干噪比奖励值$\text{reward}_{\text{SINR}_n}$	-4,-6,-8
功率切换代价奖励值reward_{P_n}	-3,-4,-5
频率切换代价奖励值reward_F	-2
功率切换成功奖励reward_{P_s}	3,4,5
学习率 α	0.1
每次用于训练的样本数 sample	32 个元组

3.2　性能对比

图 5 展示了本文 MRV-DDQN 抗干扰算法与传统 DQN 抗干扰算法的收敛性能对比,损失值越小,表明神经网络收敛得越好。在算法逐渐收敛的过程中,MRV-DDQN 算法损失值最终达到 0 左右,DQN 算法损失值在 25 附近。从图中可以看出,MRV-DDQN 的损失值更低,能够掌握更好的抗干扰策略,输出更优的频率切换和功率切换动作。



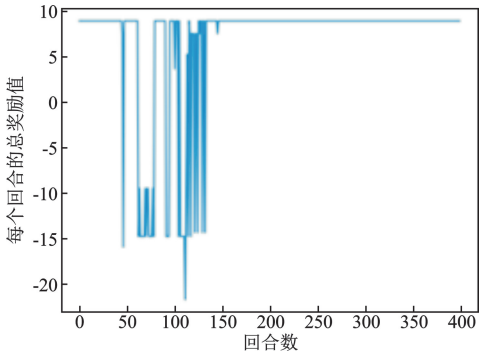
(a) MRV-DDQN 算法



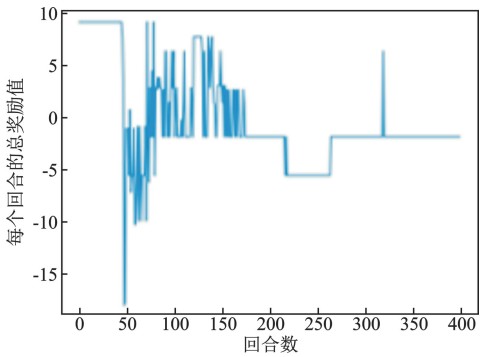
(b) DQN 算法

图 5　算法收敛性能对比

由于奖励值函数包含无干扰的初始奖励和若干受干扰的奖励代价,它也反映了系统的数据吞吐能力,奖励值越大,单位时间内的数据吞吐量越高。图 6 展示了 MRV-DDQN 和 DQN 抗干扰算法每回合总奖励值的对比。随着回合数的增加,MRV-DDQN 抗干扰算法的总奖励值不断增加,最终达到一个固定值,表明系统的数据吞吐能力不断提高并最终达到稳定状态。相比之下,DQN 抗干扰算法每回合的总奖励值较低,传输数据能力较差。



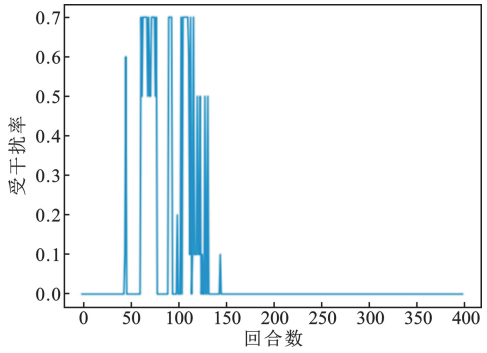
(a) MRV-DDQN 算法



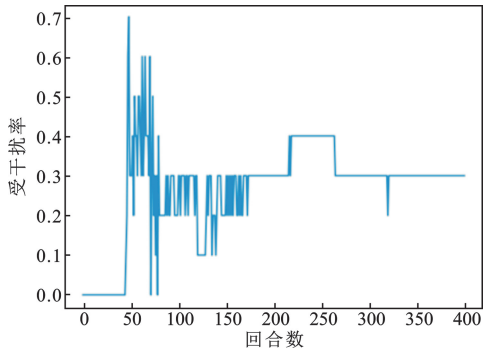
(b) DQN 算法

图 6　算法总奖励值对比

图 7 展示了 MRV-DDQN 和 DQN 抗干扰算法在每个回合中的归一化受干扰率。结果显示,MRV-DDQN 抗干扰算法在收敛后每回合的受干扰率为 0,表明其能够有效摆脱周期性的动态干扰,而 DQN 抗干扰算法在收敛后每回合的受干扰率为 0.3 左右。因此,MRV-DDQN 抗干扰算法在频率域的抗干扰性能明显优于 DQN 抗干扰算法。



(a) MRV-DDQN 算法



(b) DQN 算法

图 7 算法受干扰率对比

在训练的 400 回合中,系统生成了 4 000 个时隙,展示效果对比不明显。因此,在这 4 000 个时隙中,每隔 100 个时隙进行一次干扰功率和发射功率的抽样,如图 8 所示。算法在运行过程中与环境不断交互,当抗干扰策略未完全优化时,频率切换动作可能导致发送方通信频段切换到带有干扰功率的信道中,并且可能出现发射功率低于干扰功率的情况(如抽样 8、抽样 12)。此时,算法仅掌握部分抗干扰策略,而功率切换动作也可能导致发射功率高于干扰功率 6 dBm(如抽样 2、抽样 5、抽样 9),实现部分功率域抗干扰。当算法收敛并完全掌握抗干扰策略时,频率切换动作将使发送方通信频段切换到空闲的子带,并以最低的发射功率(6 dBm)进行通信,以实现资源的最佳利用。

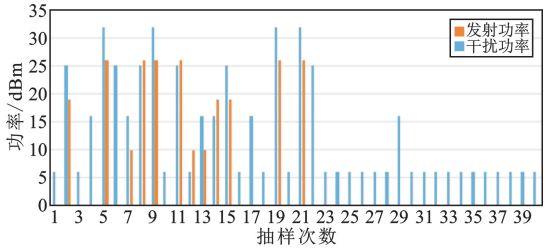


图 8 发射功率与干扰功率对比

4 结束语

本文研究了无线通信网络中功率域和频率域的联合抗干扰问题,综合考虑了资源利用、频率切换代价和功率切换代价等因素,重点解决了在有限频谱资源内如何选择合适的信道及发射功率以实现有效抗干扰。本文将发送方、接收方和干扰方之间的交互建模为马尔可夫决策过程,通过改进信道切换和功率切换代价的奖励函数,并利用干扰特征和信道信息训练抗干扰策略,实现了频率域和功率域的联合抗干扰决策。在有限资源的情况下,通过引入频率切换和功率切换机制提升了信道和功率资源的利用率,引入双神经网络则提升了系统的收敛性能。

在通信抗干扰领域仍然有许多问题有待分析和探讨:①在通信抗干扰的上下游领域包括智能干扰感知、智能抗干扰波形重构也是值得深入研究的方向;②未来抗干扰领域将由单域、多域向全域发展,这使得抗干扰的策略空间急剧膨胀,因此深度强化学习方面的策略优化是未来研究的重点。

参考文献:

[1] 陈书恒,莫嘉倩,莫小欣. 机载低轨卫星通信发展及关键技术综述[J]. 电讯技术,2024,64(1):149-157.

[2] 魏鹏. 卫星通信智能抗干扰决策技术研究[D]. 长沙:国防科技大学,2021.

[3] FOURATI F, ALOUINI M S. Artificial intelligence for satellite communication: a review [J]. Intelligent and Converged Networks,2021,2(3):213-243.

[4] ZHANG Y P,ZHAO Z J,ZHENG S L,et al. Intelligent anti-jamming decision with continuous action and state in bivariate frequency agility communication system[J]. IEEE Transactions on Cognitive Communications and Networking,2023,9(6):1579-1595.

[5] LI X C,CHEN J N,LING X,et al. Deep reinforcement learning-based anti-jamming algorithm using dual action network [J]. IEEE Transactions on Wireless Communications,2023,22(7):4625-4637.

[6] QI J,ZHANG H M,QI X L,et al. Deep reinforcement learning based hopping strategy for wideband anti-jamming wireless communications [J]. IEEE Transactions on Vehicular Technology, 2024, 73(3): 3568-3579.

[7] NOORI H,SADEGHI VILNI S. Jamming and anti-jamming in interference channels: a stochastic game approach[J]. IET Communications,2020,14(4):682-692.

[8] HAN C, HUO L Y, TONG X H, et al. Spatial anti-jamming scheme for Internet of satellites based on the deep reinforcement learning and Stackelberg game[J].

IEEE Transactions on Vehicular Technology, 2020, 69(5):5331-5342.

[9] ZHOU Q, LI Y G, NIU Y T. Intelligent anti-jamming communication for wireless sensor networks: a multi-agent reinforcement learning approach[J]. IEEE Open Journal of the Communications Society, 2021, 2: 775-784.

[10] 王桂胜,董淑福,黄国策. 无人系统认知联合抗干扰通信研究综述[J]. 计算机工程与应用,2022,58(8):1-11.

[11] ZENG X Y, CAI H, TANG X H, et al. Optimal frequency hopping sequences of odd length[J]. IEEE Transactions on Information Theory, 2013, 59(5): 3237-3248.

[12] XU H, CHENG Y F, WANG P Y. Jamming detection in broadband frequency hopping systems based on multi-segment signals spectrum clustering[J]. IEEE Access, 2021, 9: 29980-29992.

[13] WU J L, GUO F C. Time-frequency parameter estimation method of frequency hopping signal based on morphology method under low SNR [C]//2021 IEEE 6th International Conference on Signal and Image Processing. Nanjing: IEEE, 2022: 734-738.

[14] DOANIS P, SPYROPOULOS T. Sample-efficient multi-agent DQNs for scalable multi-domain 5G+ inter-slice orchestration [J]. IEEE Transactions on Machine Learning in Communications and Networking, 2024, 2: 956-977.

[15] LIU G S, DENG W J, XIE X R, et al. Human-level control through directly trained deep spiking Q-networks [J]. IEEE Transactions on Cybernetics, 2023, 53(11): 7187-7198.

作者简介:

凌　耀　男,1998 年生于江苏盐城,2021 年获工学学士学位,现为硕士研究生,主要研究方向为智能抗干扰决策。

谢世珺　女,1980 年生于贵州凯里,2005 年获工学硕士学位,现为副研究员,主要研究方向为卫星通信。

梁　豪　男,1993 年生于山东泰安,2020 年获工学博士学位,现为助理研究员,主要研究方向为卫星通信。

冯　姣　女,1984 年生于吉林延吉,2014 年获工学博士学位,现为副教授,主要研究方向为无线通信。

高伟杰　男,2000 年生于江苏无锡,2022 年获工学学士学位,现为硕士研究生,主要研究方向为智能抗干扰决策。