# 基于相似性样本生成的深度强化学习快速抗干扰算法

周权[1,2]，牛英滔[1]
（1.国防科技大学第六十三研究所，江苏 南京 210007；2.陆军工程大学通信工程学院，江苏 南京 210007）

**摘　要：**为提高基于深度强化学习的通信抗干扰算法的学习效率，以更快适应未知干扰环境，提出一种基于相似性样本生成的深度强化学习快速抗干扰算法。该算法将基于互模拟关系的状态–动作对相似性度量与基于深度Q网络的抗干扰算法相结合，能在未知动态干扰环境下快速学习有效的多域抗干扰策略。算法在完成每步传输动作时，首先利用深度Q网络抗干扰算法与环境交互，获得实际的状态–动作对。然后，基于互模拟关系生成与之相似的状态–动作集，从而利用相似状态–动作集生成模拟的训练样本。通过上述操作，算法每步迭代能获得大量训练样本，可显著加快抗干扰算法的训练进程和收敛速度。仿真结果表明，在多路扫频干扰和智能阻塞干扰下，所提算法收敛速度快，且收敛后的归一化吞吐量均显著优于常规深度Q网络算法、Q学习算法以及基于知识复用的改进Q学习算法。

**关键词：**通信抗干扰；深度强化学习；快速抗干扰；可靠通信
**中图分类号：**TN973.3
**文献标志码：**A
**DOI：**10.11959/j.issn.1000–436x.2024131

# Fast deep reinforcement learning anti-jamming algorithm based on similar sample generation

ZHOU Quan[1,2], NIU Yingtao[1]
1. The Sixty-third Research Institute, National University of Defense Technology, Nanjing 210007, China
2. School of Communication Engineering, Army Engineering University of PLA, Nanjing 210007, China

**Abstract:** To improve the learning efficiency of anti-jamming algorithms based on deep reinforcement learning and enable them to adapt more quickly to unknown jamming environments, a fast deep reinforcement learning anti-jamming algorithm based on similar sample generation was proposed. By combining the similarity measurement of state-action pairs, derived from bisimulation, with an anti-jamming algorithm grounded in the deep Q-network, this algorithm was able to quickly learn effective multi-domain anti-jamming strategies in unknown, dynamic jamming environments. Specifically, once a transmission action was completed, the proposed algorithm first interacted with the environment using the deep Q-network to acquire actual state-action pairs. Then it generated a set of similar state-action pairs based on bisimulation, employing these similar state-action pairs to produce simulated training samples. Through these operations, the algorithm was able to acquire a large number of training samples at each iteration step, thereby significantly accelerating the training process and convergence speed. Simulation results show that under comb sweep jamming and intelligent blocking jamming, the proposed algorithm exhibits rapid convergence speed, and its normalized throughput after convergence significantly superior to the conventional deep Q-network algorithm, the Q-learning algorithm, and the improved Q-learning algorithm based on knowledge reuse.

**Keywords:** communication anti-jamming, deep reinforcement learning, fast anti-jamming, reliable communication

**收稿日期：**2024–02–02；**修回日期：**2024–06–28
**通信作者：**牛英滔，niuyingtao78@163.com
**基金项目：**国家自然科学基金资助项目（No.62371461）
**Foundation Item：**The National Natural Science Foundation of China (No.62371461)

## 0 引言

无线通信由于以电磁波为信息传播的载体，因此其传输信道具有开放性。干扰机能轻易地向目标信道注入恶意干扰信号，以降低传输信号在接收端的信干噪比（SJNR, signal-to-jamming-plus-noise ratio），使所传数据难以被正确解调，进而使传输质量下降甚至中断传输。为确保无线通信在恶意干扰威胁下的可靠性和有效性，有必要采用通信抗干扰技术。

作为现代通信抗干扰的主流技术之一，扩展频谱（以下简称扩谱）技术自20世纪50年代起应用于通信抗干扰，并在各种无线通信系统中广泛应用[1]。目前，扩谱技术已从常规扩谱发展为自适应扩谱[2]，典型代表包括频率/功率自适应跳频技术和基于窄带干扰自适应抑制/功率自适应的直接序列扩谱技术。这些技术具备一定的干扰检测及基于规则的功率/频率自适应调整能力，能有效应对单音、多音、部分频带等常规干扰，但对干扰环境的适应能力有限，难以有效应对多种动态干扰、高效干扰、灵巧干扰、认知干扰、智能化干扰等新型干扰。近年来，人工智能相关技术的兴起为通信抗干扰技术注入了新的活力。当前许多研究围绕智能通信抗干扰技术展开，特别是基于深度强化学习（DRL, deep reinforcement learning）的抗干扰技术。由于具备探索未知干扰环境并自主生成抗干扰策略的特点，基于DRL的抗干扰技术成为智能通信抗干扰的重要研究方向。

目前，基于DRL的通信抗干扰技术广泛采用深度Q网络（DQN, deep Q-network）及其改进算法解决复杂的通信抗干扰问题，特别是跨域抗干扰问题。文献[3]提出了一种基于DQN的跨域抗干扰算法，使移动节点在未知动态干扰环境下习得最佳位置调整和功率控制策略，以实现可靠传输。文献[4]提出了一种基于DQN的水下通信节点功率域和空域联合抗干扰算法。文献[5]提出了一种基于改进型DQN的多域抗干扰算法，能够在时、频、空多域动态干扰环境下，从频谱瀑布中习得干扰变换规律，进而通过选择合适的中继和传输信道躲避干扰。文献[6]提出了一种基于DQN的视频流传输系统抗干扰算法，利用安全性能指标设计奖励函数以优化传输策略，使发射机习得最佳的视频压缩编码、调制编码和发射功率选择策略。然而，以DQN为代表的DRL算法需要充分探索策略空间以习得有效的抗干扰策略。而多域抗干扰问题由于具有较大的状态–动作空间，算法需要更长时间进行策略探索，因此收敛速度较慢。为改进这一不足，现有研究提出了多种方法。文献[7]提出了一种安全强化学习抗干扰算法，该算法采用策略优先级的分层网络架构，并利用迁移学习初始化网络参数，可显著加快算法收敛速度。文献[8]提出了一种基于“赢”或“快速学习”策略爬山的强化学习抗干扰算法，采用模糊状态聚合对状态空间进行降维，显著提升了算法收敛速度。文献[9]提出了一种快速收敛的DRL抗干扰算法，采用软标签代替奖励，能够获得更多信息熵并且算法不需要随机探索，显著加快了算法收敛速度。现有研究主要从改进强化学习算法入手，以加快抗干扰策略的收敛速度。

有别于上述工作，文献[10]从抗干扰问题特性入手，基于状态–动作对的相似性实现域内知识复用，显著提升了Q学习抗干扰算法收敛速度。其核心思想是，抗干扰问题中通信方的传输动作只有传输成功或失败2种结果，因此，许多看似没有关联的状态–动作对由于传输结果相同而具有内在的相似性，这种相似性为加速学习进程提供了可能。

本文在文献[10]的基础上，研究了抗干扰问题的内在相似性如何提升DQN抗干扰算法的收敛速度，并在文献[10]考虑频率域和功率域传输动作的基础上，额外引入了速率域，使得不同域动作间产生耦合关系，加大了算法策略空间的尺寸和寻优的难度。此外，不同于文献[10]采用的基于表格的Q学习算法，本文采用DQN作为决策算法。该算法无法通过查表直接在相似状态–动作对之间复用知识，如何利用相似性加速DRL策略网络的训练进程是一个难点。

为此，本文根据DQN算法特性，提出了一种基于相似性样本生成的DRL快速抗干扰算法，其核心思想是在实际交互获得经验样本的基础上，基于相似性生成额外的模拟样本，这将显著提升算法每次迭代获取的样本数量和多样性，从而加快策略网络的训练进程。本文的主要贡献如下。

1) 提出了一种基于相似性样本生成的DQN快速抗干扰算法，在相似性度量算法的基础上，能在每步迭代期间利用相似状态–动作集生成模拟训练

样本，减少收集实际训练样本所需的迭代次数，从而加快算法学习进程。

2) 鉴于无线通信抗干扰任务与常规强化学习任务的区别，所提算法设计充分考虑了实际状态观测的滞后性以及奖励的延迟反馈，能够为强化学习抗干扰算法设计提供有益参考。

3) 验证了利用抗干扰问题内在相似性加速DRL算法收敛速度的可行性，为加速其他基于DRL的通信抗干扰算法提供了有益参考。

## 1　系统模型和问题建模

### 1.1　系统模型

无线通信系统（以下简称系统）由一对收发信机组成，双方分时隙传输，其传输受外部恶意干扰机威胁，如图1所示。为实现可靠通信，系统联合采用动态频谱接入、功率控制和速率控制进行通信，其可选信道为 $f\in\{1,2,\cdots,M\}$，可选功率为 $p\in\{p_1,\cdots,p_D\}$，可选传输速率为 $v\in\{v_1,\cdots,v_E\}$。与文献[11]类似，假设传输信道为常见的块衰落模型[12]，即信道增益在同一个时隙内保持不变，而在时隙间变化，不同衰落块的传输信道增益表示为 $g_T=g_T^p|h_T^s|$，其中，$g_T^p$ 表示 $T$ 时隙给定距离时的路径损耗，$|h_T^s|$ 表示衰落块之间服从瑞利分布，且 $h_T^s\sim\mathrm{CN}(0,1)$。

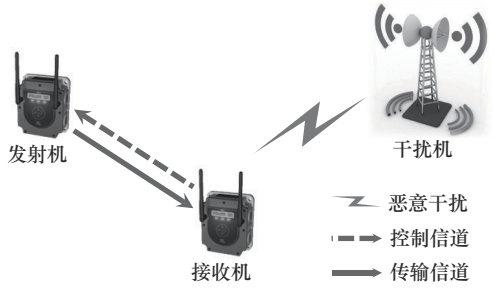



图1　系统模型

本文采用与文献[13]类似的技术，接收机能够同时进行宽带频谱感知和数据接收，并通过分析被占用信道的信号波形区分干扰信号与合法信号。接收机通过学习决策发射机的传输动作，假设干扰机主要瞄准接收机实施干扰，接收机能够通过协议加固的低容量控制信道向发射机传递动作指令。

干扰机能够发射频率–功率域动态干扰，并且可以在不同时隙间改变干扰功率 $p_k'\in\{p_1',\cdots,p_E'\}$ 和阻塞信道 $f_k'\in\{1,2,\cdots,M\}$，$k\in\{1,2,\cdots,K\}$ 表示阻塞的第 $k$ 个信道，$K\leqslant M$ 表示干扰机在同一时隙阻塞的信道数。干扰机能否成功阻断通信取决于注入传输信道的干扰信号功率能否使接收端的SJNR小于解调阈值。为不失一般性，此处只假设干扰样式，而不限定具体干扰参数。

### 1.2　问题建模

本文采用马尔可夫决策过程（MDP, Markov decision process）对通信抗干扰问题进行建模。MDP是强化学习常用的形式化模型之一，可表示为多元组 $\langle\mathcal{S},\mathcal{A},\mathcal{F},\mathcal{R}\rangle$，其中，$\mathcal{S}$ 为环境状态空间；$\mathcal{A}$ 为动作空间；$\mathcal{F}$ 为概率转移函数，表示在给定当前状态和动作时下一个状态的概率分布；$\mathcal{R}$ 为奖励函数。

1) 状态

定义 $T$ 时隙的环境状态为接收机在 $T-1$ 时隙时 $M$ 个信道上的平均接收功率组合，表示为

$$s_T=\left[\bar{P}_{1,T-1},\bar{P}_{2,T-1},\cdots,\bar{P}_{M,T-1}\right]\in\mathcal{S} \tag{1}$$

其中，平均接收功率 $\bar{P}_{m,T-1}$ 由信道 $m$ 上接收到的发射机信号功率和干扰信号功率组成，可表示为

$$\bar{P}_{m,T-1}=p_{T-1}g_{T-1}\delta(f_{T-1}=m)+\overline{P'}_{m,T-1} \tag{2}$$

其中，$\delta(\cdot)$ 为指示函数，即 $x$ 为真时，$\delta(x)=1$，否则 $\delta(x)=0$；$p_{T-1}$ 是发射机的发射功率，$g_{T-1}$ 是传输信道增益，$\overline{P'}_{m,T-1}$ 表示接收到的平均干扰加噪声功率。

环境状态取决于发射机信号和干扰信号，系统无法预知干扰信号，因此难以提前预知状态空间 $\mathcal{S}$ 中的所有状态，但能通过持续的频谱感知记录已观测状态。当接收机观测到信道环境状态 $s_T$ 并得到任意信道 $m$ 上的平均接收功率 $\bar{P}_{m,T-1}$ 后，根据式(2)可从中分离出 $\overline{P'}_{m,T-1}$。结合已知的可选发射功率 $p\in\{p_1,\cdots,p_D\}$、传输信道 $f\in\{1,2,\cdots,M\}$ 以及信道估计所得的信道增益 $\tilde{g}_{T-1}$，可推测发射机采用其他传输动作时信道 $m$ 上可能的平均接收功率 $\tilde{P}_{m,T-1}$ 为

$$\tilde{P}_{m,T-1}=p\tilde{g}_{T-1}\delta(f=m)+\overline{P'}_{m,T-1} \tag{3}$$

进而，可得到预测状态 $\tilde{s}_T=[\tilde{P}_{1,T-1},\tilde{P}_{2,T-1},\cdots,\tilde{P}_{M,T-1}]$。本文将包含 $T$ 时隙实际观测状态 $s_T$ 和所有预测状态 $\tilde{s}_T$ 的集合记作 $\tilde{\mathcal{S}}_T$，系统能通过各时隙不断累积 $\tilde{\mathcal{S}}_T$ 得到估计状态集 $\tilde{\mathcal{S}}\leftarrow\tilde{\mathcal{S}}+\tilde{\mathcal{S}}_T$。随着时间的推移，当系统充分遍历所有状态后，$\tilde{\mathcal{S}}$ 将逼近 $\mathcal{S}$。

2) 动作

接收机在$T$时隙决策的传输动作由传输信道、功率和速率组成，定义为

$$a_T=(f_{T+1},p_{T+1},v_{T+1})\in\mathcal{A} \tag{4}$$

其中，$a_T$由接收机通过控制信道反馈给发射机，由发射机在$T+1$时隙执行。

3) 奖励

奖励函数计算环境状态$s_T$下决策的传输动作$a_T$所能获得的奖励值，由于$a_T$将在$T+1$时隙执行，因此奖励函数定义为

$$r(s_T,a_T)=\zeta_{v_{T+1}}\delta(\theta_{T+1}\geqslant\theta^{\text{th}}_{v_{T+1}})-\zeta_{p_{T+1}} \tag{5}$$

其中，$\zeta_{v_{T+1}}$是传输速率相较于基准速率的权重系数，$\zeta_{p_{T+1}}$是发射功率相较于基准功率的代价因子；$\theta_{T+1}=\dfrac{\bar{P}_{m,T+1}}{\overline{P'}_{m,T+1}}$为$T+1$时隙接收端平均SJNR；$\theta^{\text{th}}_{v_{T+1}}$为接收机的SJNR解调阈值，其大小受传输速率影响；$\delta(\cdot)$为指示函数。上述奖励函数表示$T$时隙的状态–动作$(s_T,a_T)$对应的奖励，取决于$T+1$时隙接收机能否成功解码所传数据，因此只有观测到$s_{T+2}=[\bar{P}_{1,T+1},\cdots,\bar{P}_{M,T+1}]$才可计算$r(s_T,a_T)$，体现了延迟反馈的奖励机制。式(5)中$\theta_{T+1}\geqslant\theta^{\text{th}}_{v_{T+1}}$是否成立反映当前时隙的传输是否成功，指示函数$\delta(\cdot)$的值对应单位时间的吞吐量为1或0。因此，奖励函数可视为在考虑功率和速度代价情况下的单位时间吞吐量。为体现奖励在通信抗干扰问题中的物理意义，本文称其为归一化吞吐量。

4) 目标

系统目标是找到最优传输策略$\pi^*(a|s)$，系统在$T$时隙任意状态$s$下选择动作$a$后，开始执行最优传输策略，能够获得最大累积折扣奖励，可用最优$Q$函数表示为

$$Q^*(s,a)=\max_{\pi}\mathrm{E}_{\pi}\left[\sum_{\tau=0}^{\infty}\gamma^{\tau}r(s_{T+\tau},a_{T+\tau})\,\middle|\,s_T=s,a_T=a\right] \tag{6}$$

其中，$\mathrm{E}_{\pi}[\cdot]$为数学期望算子，$\gamma\in[0,1]$是反映未来奖励对当前决策影响程度的折扣因子，$\tau$表示从$T$时隙起的后续步数。如果能求得所有状态–动作$(s,a)$对应的最优$Q$值，就能按式(7)推测出最优策略$\pi^*(a|s)$。

$$\pi^*(a|s)=\begin{cases}1, a=\arg\max\limits_{a\in\mathcal{A}}Q^*(s,a)\\ 0, \text{其他}\end{cases} \tag{7}$$

本文采用基于DQN的抗干扰算法逼近最优传输策略，DQN是一种采用深度神经网络拟合$Q$函数的DRL算法[14]，通过迭代优化网络参数使深度神经网络逐渐逼近最优$Q$函数，从而在任意给定状态下输出所有可选动作的最优$Q$值，进而得到联合功率、频率、速率的最优抗干扰传输策略。

## 2 相似性样本生成

给定任意2组$(s,a),r(s,a)$和$(s',a'),r(s',a')$（分别记作$x,r,x',r'$），可根据文献[10]所提基于互模拟关系的状态–动作对相似性度量算法判断$(s,a)$是否与$(s',a')$相似，表示为

$$\Psi(x,r,x',r')=\begin{cases}1, x\sim x'\\ 0, \text{其他}\end{cases} \tag{8}$$

其中，$\Psi(\cdot)$为相似性度量算子；$x\sim x'$表示给定的状态–动作对$(s,a)$与$(s',a')$满足互模拟关系，条件为$|r(s,a)-r(s',a')|\leqslant\vartheta_{\mathcal{R}}$，$\forall x_f\in\mathcal{X}\sim\Psi$，满足$\sum\limits_{x_f}F(s_f|s,a)=\sum\limits_{x_f}F(s_f|s',a')$，$\vartheta_{\mathcal{R}}$表示相似性门限，$\mathcal{X}\sim\Psi$表示状态–动作空间$\mathcal{X}$中满足互模拟关系的状态–动作对集，$s_f$为$x_f$中的状态分量，函数$F(\cdot)$为状态转移概率函数，$F(s_f|s,a)$表示在状态$s$执行动作$a$转移至状态$s_f$的概率。

考虑通信抗干扰问题环境感知的滞后性以及奖励的延迟反馈，算法在$T$时隙根据过去的交互记录构建一组实际经验$(s_{T-2},a_{T-2},r_{T-2},s_{T-1},s_T)$，其中，$(s_{T-2},a_{T-2},r_{T-2},s_{T-1})$是用于网络训练的实际样本。如1.2节所述，已知$s_{T-1}$时可根据动作空间的其他任意可选动作$\tilde{a}_{T-2}\in\mathcal{A}$得到对应的预测状态$\tilde{s}_{T-1}=[\tilde{P}_{1,T-2},\cdots,\tilde{P}_{M,T-2}]$，记作$\tilde{s}_{T-1}\propto\tilde{a}_{T-2}$，其中$\tilde{P}_{m,T-2}=p_{T-2}\tilde{g}_{T-2}\delta(f_{T-2}=m)+\overline{P'}_{m,T-2}|_{\forall f_{T-2},p_{T-2}\in\mathcal{A}}$。同样地，已知$s_T$和$s_{T-2}$能得到对应的预测状态$\tilde{s}_T$和$\tilde{s}_{T-2}$。进一步根据式(5)可知，已知$\tilde{s}_T$可计算出任意$\tilde{x}_{T-2}$的对应预测奖励$\tilde{r}_{T-2}$。因此，在$T$时隙接收机能根据历史交互经验，得到与$x_{T-2}$相似的状态–动作集为

$$\mathcal{X}_T=\{\tilde{x}_{T-2}=(\tilde{s}_{T-2},\tilde{a}_{T-2}):\ \forall\Psi(x_{T-2},r_{T-2},\tilde{x}_{T-2},\tilde{r}_{T-2})=1\} \tag{9}$$

将上述集合$\mathcal{X}_T$中的$(\tilde{s}_{T-2},\tilde{a}_{T-2})$以及与$\tilde{a}_{T-2}$对应的预测状态$\tilde{s}_{T-1}\propto\tilde{a}_{T-2}$替换实际训练样本$(s_{T-2},a_{T-2},r_{T-2},s_{T-1})$中的对应元素，可得到相似

性样本集为

$$\mathcal{E}_T=\left\{(\tilde{s},\tilde{a},r_{T-2},\tilde{s}_{T-1}):\forall(\tilde{s},\tilde{a})\in\mathcal{X}_T,\tilde{s}_{T-1}\propto\tilde{a}\right\}\tag{10}$$

上述计算步骤可归纳为算法1。

**算法1**　相似性样本生成算法

**输入**　实际交互经验$(s_{T-2},a_{T-2},r_{T-2},s_{T-1},s_T)$

**输出**　$\mathcal{E}_T=\left\{(\tilde{s},\tilde{a},r_{T-2},\tilde{s}_{T-1}):\forall(\tilde{s},\tilde{a})\in\mathcal{X}_T\right\}$

1) 根据$s_T$得到预测状态$\tilde{s}_T$

2) 根据$\tilde{s}_T$得到预测奖励$\tilde{r}_{T-2}$

3) 根据$s_{T-1}$得到预测状态$\tilde{s}_{T-1}\propto\tilde{a}_{T-2}$

4) 根据$s_{T-2}$得到预测状态$\tilde{s}_{T-2}$

5) 按式(9)得到相似状态–动作集$\mathcal{X}_T$

6) 按式(10)得到相似性样本集$\mathcal{E}_T$

## 3　基于相似性样本生成的快速抗干扰算法

为加快DRL抗干扰算法的收敛速度，本文提出了一种基于相似性样本生成的快速抗干扰算法，算法架构如图2所示。该算法基本结构采用了常规DQN的经典架构[14]，考虑了通信抗干扰问题状态感知的滞后性和奖励的延迟反馈，并引入了算法1，使每步迭代生成额外的相似性样本，以提高算法起始训练阶段样本采集的效率，进而加速算法的学习进程，具体步骤可归纳为算法2。

**算法2**　基于相似性样本生成的快速抗干扰算法

**输入**　环境状态$s_T$

**输出**　抗干扰传输策略$\pi$

初始化　$Q$，$\hat{Q}$(令$\hat{Q}=Q$)

1) for $T=1,2,\cdots,\infty$

2)　发射机根据动作指令执行传输动作$a_{T-1}$

3)　计算出传输动作奖励$r_{T-2}$

4)　记录$(r_{T-2},s_{T-1},a_{T-1},s_T)$

5)　根据记录构建样本$(s_{T-2},a_{T-2},r_{T-2},s_{T-1})$并存入样本库

6)　执行算法1得到相似性样本集$\mathcal{E}_T$

7)　将相似样本$(\tilde{s},\tilde{a},r_{T-2},\tilde{s}_{T-1})$存入样本库

8)　从经验池批量采样$(s_i,a_i,r_i,s_{i+1})$

9)　样本输入目标网络和策略网络，前者输出$\max\limits_a\hat{Q}(s_{i+1},a;\theta)$，后者输出$Q(s_i,a_i;\theta)$

10)　随机梯度下降+反向传播法更新损失函数$(r_i+\max\limits_a\hat{Q}(s_{i+1},a;\theta)-Q(s_i,a_i;\theta))^2$中的网络参数$\theta$

11)　时隙末感知频谱得到环境状态$s_{T+1}$

12)　利用当前策略网络输出下一时隙传输

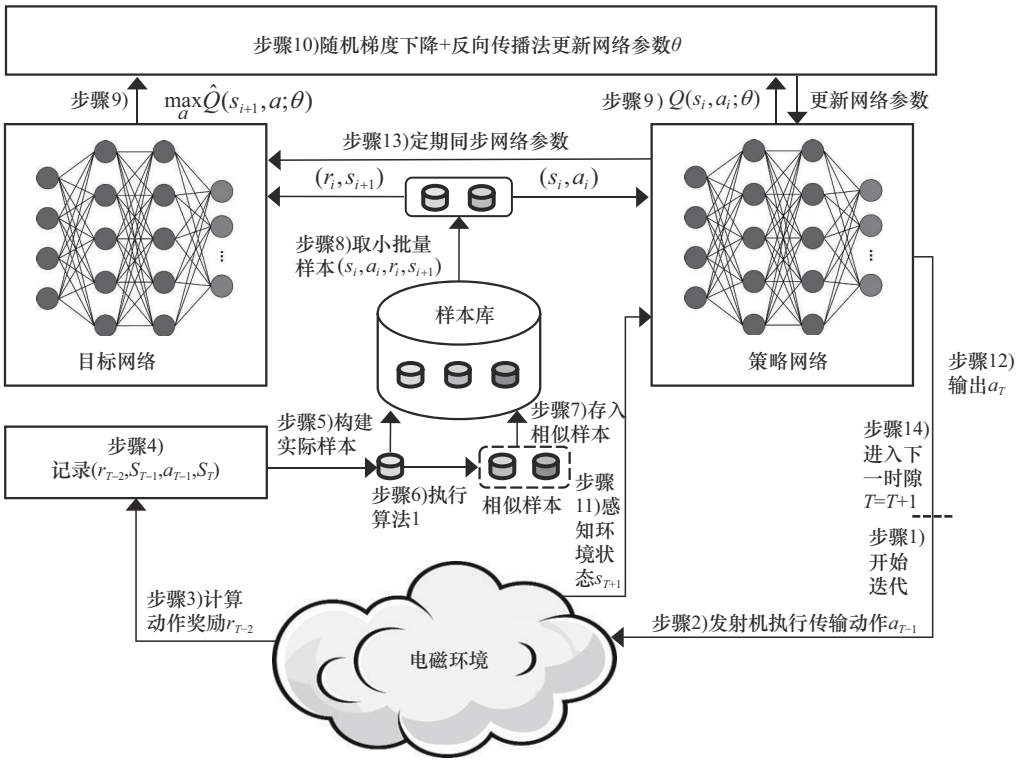



图2　算法架构

动作$a_T$并传回发射机

13)　每隔$C$个时隙令$\hat{Q}=Q$

14)　$T=T+1$

15) end for

本文算法是一种在线学习算法，传输终止前该算法将在每个时隙迭代一次，并逐渐逼近最优多域抗干扰传输策略。如图3所示，算法中频谱感知、学习和决策、通信传输3个部分在同一时隙内并行操作。当前时隙频谱感知的输出将作为下一时隙学习和决策的输入状态，而当前时隙学习和决策的输出将由下一时隙的通信传输执行。这种并行架构有助于获取更连续的频谱状态，并使学习和决策的时间更充裕，同时感知、学习和决策不会挤占通信传输时间。相比于智能抗干扰算法常用的串行结构[15-16]和部分并行结构[17-18]，在相同传输速率下，并行时隙结构能获得更大的吞吐量。

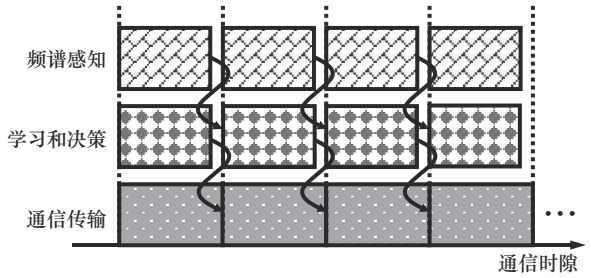



图3　算法时隙结构

## 4　算法仿真

### 4.1　仿真设置

本文算法仿真基于PyTorch实现，计算机处理器为Intel i5 12400F，显卡为NVIDIA RTX 3060。DQN的主网络和目标网络均采用3层全连接网络，其中输入层维度为5（对应状态维度），输出层维度为20（对应可选动作数），隐藏层节点数为64，样本库最大容量为10 000，小批量样本数为64，其余参数设置如表1所示。需要说明的是，折扣率的大小决定算法进行决策时对将来能获得奖励的重视程度。一般而言，过大的折扣率容易使算法因过于考虑将来的收益损失而变得保守，过小的折扣率则可能使算法因“短视”而错误评估当下决策的优劣。通信抗干扰问题通常只关注当前时隙或多个时隙所传数据包是否成功，因此算法决策只需关注当前及将来少数几步的奖励即可，故将折扣率设为相对较小的值。

**表1　仿真参数设置**

| 参数 | 数值 |
|---|---|
| 通信时隙长度$T_s$/ms | 10 |
| 信道数$M$ | 5 |
| 可选发射功率$p_{\mathrm{L}},p_{\mathrm{M}},p_{\mathrm{H}}$/dBm | 12, 16, 20 |
| 干扰信号数$K_T$ | 2 |
| 可选干扰功率$p'_{\mathrm{L}},p'_{\mathrm{H}}$/dBm | 8, 12 |
| 解调阈值$\theta^{\mathrm{th}}_{v_{\mathrm{L}}},\theta^{\mathrm{th}}_{v_{\mathrm{M}}},\theta^{\mathrm{th}}_{v_{\mathrm{H}}}$/dB | 10, 14, 18 |
| 相似性门限$\vartheta_{\mathrm{R}}$ | 0.1 |
| 速率权重系数$\zeta_{v_{\mathrm{L}}},\zeta_{v_{\mathrm{M}}}$ | 0.5, 0.75 |
| 功率代价因子$\zeta_{p_{\mathrm{M}}},\zeta_{p_{\mathrm{H}}}$ | 0.3, 0.6 |
| 折扣率$\gamma$ | 0.3 |

算法探索–利用机制采用$\varepsilon$贪心法，其中$\varepsilon$表示探索的概率，即在训练过程中以一定的概率随机选择动作，以便探索更多的状态和策略。在线抗干扰算法需要从探索平滑过渡到利用。在训练初期，需要更多地探索，因此$\varepsilon$的值应较大；在训练后期，需要更多地利用已经学到的抗干扰策略，因此$\varepsilon$的值应逐渐减小。仿真中采用式(11)的指数衰减函数计算各时隙探索率$\varepsilon$。

$$\varepsilon(T)=\varepsilon_{\mathrm{end}}+(\varepsilon_{\mathrm{start}}-\varepsilon_{\mathrm{end}})\mathrm{e}^{-\frac{T}{\varepsilon_{\mathrm{decay}}}} \tag{11}$$

其中，$\varepsilon_{\mathrm{start}}$为初始值，可设为1以确保训练开始时充分探索；$\varepsilon_{\mathrm{end}}$为终止值，可设为0.001以确保训练后期充分利用；$T=1,2,\cdots,\infty$表示时隙数；$\varepsilon_{\mathrm{decay}}$表示衰减率，$\varepsilon_{\mathrm{decay}}$值越大，算法从探索过渡到利用所需的步数越多。

本文仿真考虑以下2种典型的动态干扰样式。

1) 多路扫频干扰：采用2路窄带干扰信号（一路为高功率，另一路为低功率）对目标频段内的各信道依次进行周期性线性扫描。这种干扰方式因其干扰效率高且易于产生，在实际通信对抗场景中得到广泛应用。

2) 智能阻塞干扰：又称智能响应式干扰，是现有工作广泛研究的智能干扰方式之一[19-21]。其干扰攻击基于感知–行动环路，能够根据过去5个时隙合法信号对信道的占用情况选择占用率最高的2个信道分别施加强弱2路窄带干扰。

本文将本文算法与以下3种典型智能通信抗干扰算法进行仿真对比。

1) 常规DQN（DQN）算法：除了不采用基于

相似性度量的样本生成以外，其余算法流程、参数设置等均与本文算法一致。

2) 常规Q学习（QL）算法：传输动作决策基于广泛应用的Q学习算法（如文献[15-16]），其状态、动作、奖励定义均与本文一致。

3) 带域内知识复用的Q学习（QL-IKR）算法：在相似的状态–动作间复用每步更新的$Q$值，从而加速Q学习收敛，详见文献[10]。

为有效评估算法性能，采用式(5)定义的奖励函数作为不同算法在考虑功率、速度代价情况下的归一化吞吐量。此外，为得到一般性的结论，仿真结果均为50次独立仿真的平均值，且曲线绘制采用滑动平均法（窗口长度为50）以获得平滑曲线。为客观地分析本文算法的性能，对比算法均采用式(11)计算各时隙的探索率。在后续仿真中，常规DQN的学习率设置与本文算法一致，2种Q学习算法的学习率设置为0.1，对比算法中与强化学习相关的其余参数设置均与本文算法保持一致。

### 4.2　仿真分析

学习率$\alpha$是DQN的一个重要超参数，它控制着神经网络参数的更新速度。为设置合适的学习率参数，本文在多路扫频干扰环境下验证不同学习率对算法性能的影响。图4对比了学习率$\alpha$分别为0.01、0.001和0.000 1时本文算法的归一化吞吐量性能。当$\alpha$=0.000 1时，由于设置过小，神经网络参数的更新速度较慢，因此收敛速度较慢；当$\alpha$=0.01时，尽管神经网络参数收敛迅速，但是会导致$Q$值估计的不稳定以及爆炸性梯度等问题，从而使得归一化吞吐量性能下降；当$\alpha$=0.001时，其收敛速度与归一化吞吐量性能均表现良好，因此后续仿真中均采用此设置。

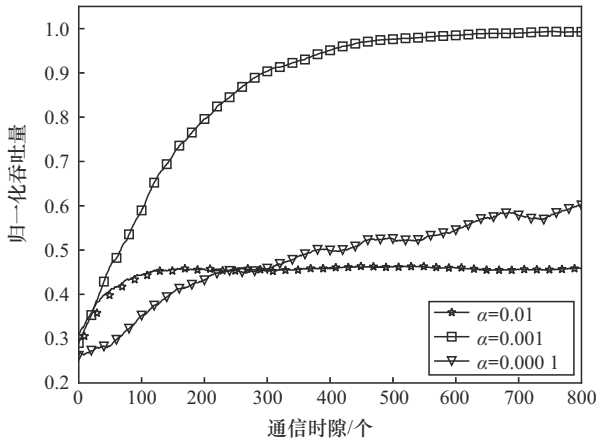



图4　不同学习率下算法的归一化吞吐量性能对比

如前所述，本文采用式(11)计算各时隙的探索率，以实现强化学习抗干扰算法从探索到利用的平滑过渡，其中衰减率$\varepsilon_{\text{decay}}$是影响算法收敛性能的一个重要超参数。为设置合理的衰减率，图5和图6分别在多路扫频和智能阻塞干扰环境下对比不同$\varepsilon_{\text{decay}}$对算法的归一化吞吐量性能的影响。从2种干扰环境下均可看出，$\varepsilon_{\text{decay}}$越大，算法从探索过渡到利用所需的步数越多，收敛所需的通信时隙数也越多，但当$\varepsilon_{\text{decay}}$较小时，算法收敛所需的通信时隙数不再显著较少，至少需500个通信时隙的迭代训练，因此后续仿真均设置$\varepsilon_{\text{decay}}$ = 50。

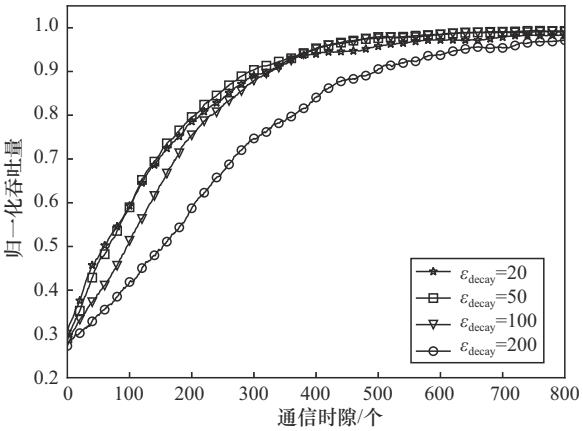



图5　多路扫频干扰下不同衰减率的算法的归一化吞吐量性能对比

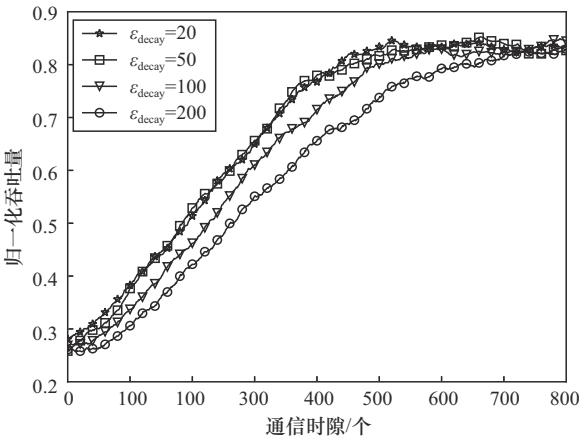
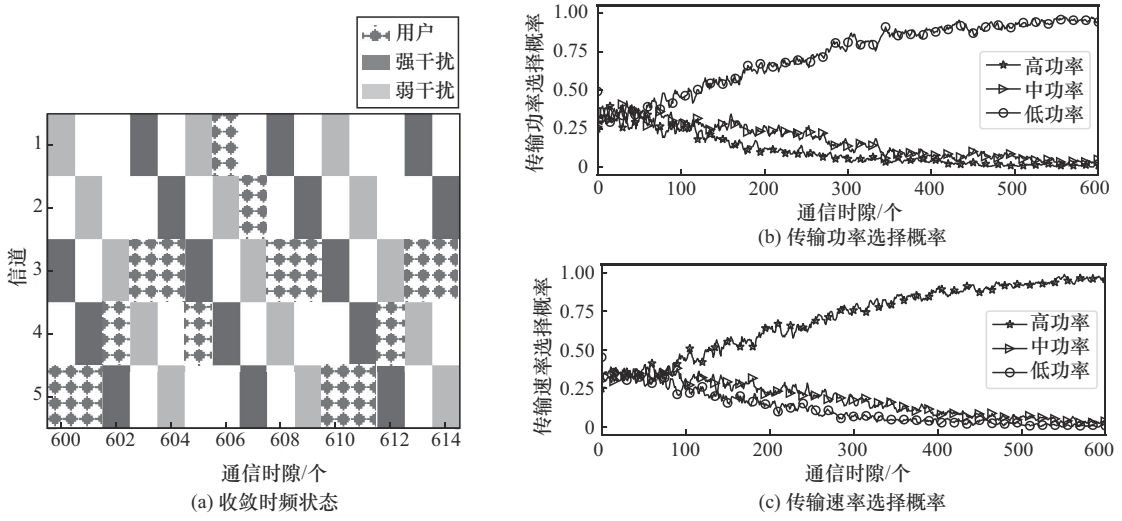



图6　智能阻塞干扰下不同衰减率的算法的归一化吞吐量性能对比

图7(a)展示了双路扫频干扰环境下本文算法的收敛时频状态，系统能在频率域完全躲避强弱2路窄带干扰信号。同时，如图7(b)和图7(c)所示，随着算法收敛，系统倾向于以低功率和高速率保持通信传输，从而获得更大的吞吐量。显然，双路扫频干扰环境下算法收敛时采用了最佳的频

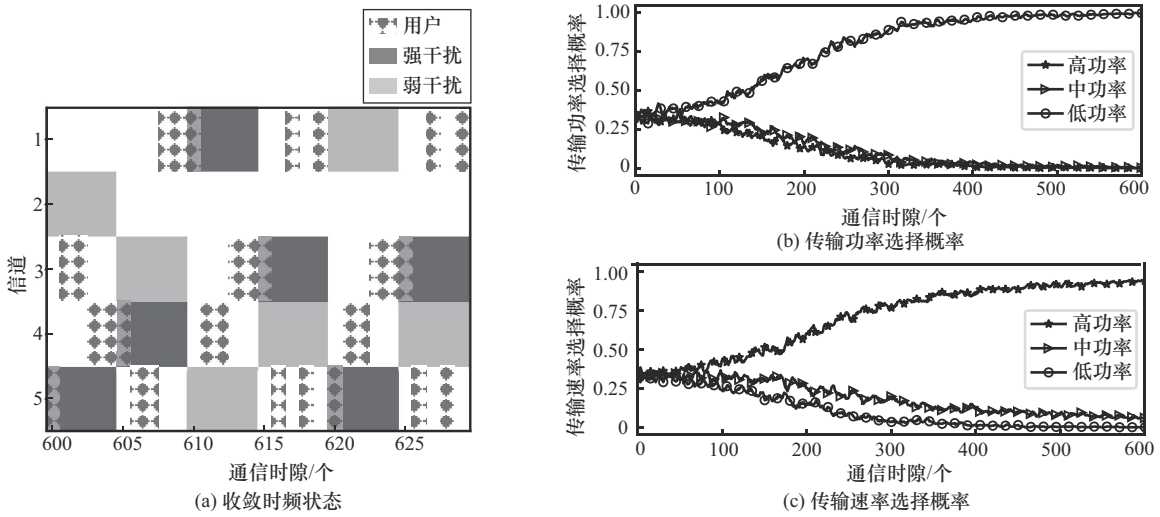
率、功率、速率多域联合抗干扰传输策略。类似地，图8(a)展示了智能阻塞干扰环境下本文算法的收敛时频状态，系统能在频率域躲避约80%的窄带干扰。同时，如图8(b)和图8(c)所示，算法追求以低功率和高速率保持通信。在智能阻塞干扰环境下，由于干扰机具备跟踪能力，系统难以完全躲避干扰，但仍能以低功率和高速率躲避约80%的窄带干扰。

图9和图10分别展示了多路扫频干扰和智能阻塞干扰环境下本文算法与DQN、QL、QL-IKR算法的归一化吞吐量性能对比。如图9所示，在多路扫频干扰下，本文算法仅需约500个通信时隙的迭代训练就能稳定收敛于归一化吞吐量趋近于1的最佳多域抗干扰策略。相同参数设置下的常规DQN算法收敛速度与本文算法大致相当，但归一化吞吐量仅收敛于0.84；QL算法的归一化吞吐量仅能收敛于0.53；QL-IKR算法性能仅次于本文算法，但受限于QL算法单步迭代的不足，归一化吞吐量收敛于0.86。如图10所示，在智能阻塞干扰下，本文算法仅需400个通信时隙的迭代训练就能收敛至平均归一化吞吐量大于0.8的多域抗干扰策略，其中无滑动平均情况下归一化吞吐量的大幅度波动源于算法仅能躲避约80%的智能阻塞干扰，与图8(a)结果分析一致。相同参数设置下的常规DQN算法经800个时隙的迭代训练归一化吞吐量仅达0.6；QL算法的归一化吞吐量仅收敛于0.25；QL-IKR算法经400个时隙的迭代训练归一化吞吐量可达0.7，性能仅次于本文算法。根据图9和图10的仿真结果可知，在多路



(a) 收敛时频状态　(b) 传输功率选择概率　(c) 传输速率选择概率

图7　双路扫频干扰下算法的收敛时频状态及功率、速率选择概率



(a) 收敛时频状态　(b) 传输功率选择概率　(c) 传输速率选择概率

图8　智能阻塞干扰下算法的收敛时频状态及功率、速率选择概率

扫频干扰和智能阻塞干扰下，本文算法收敛速度快，且收敛后的归一化吞吐量均显著优于常规DQN算法，证明了相似性样本生成能有效提高DQN抗干扰算法性能。

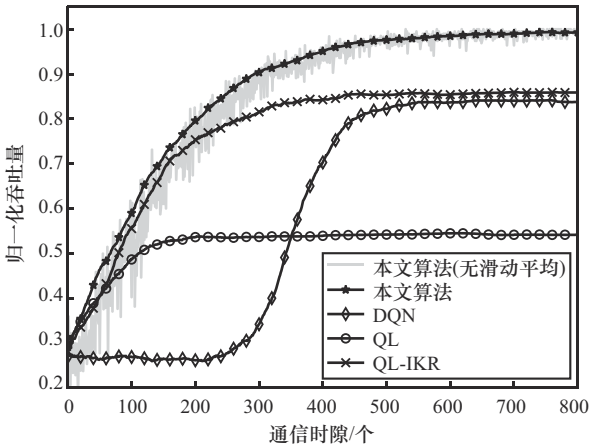



图9　多路扫频干扰下各算法的归一化吞吐量性能对比

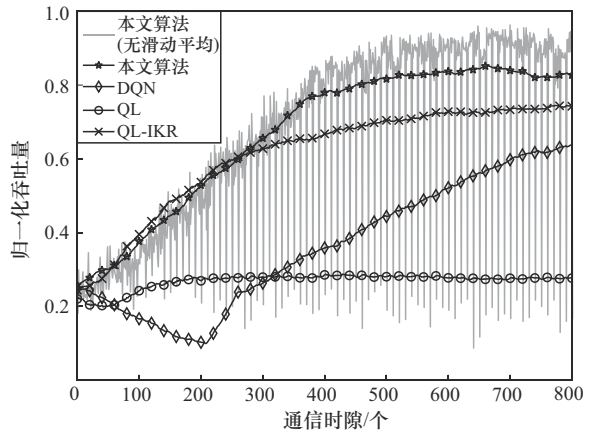



图10　智能阻塞干扰下各算法的归一化吞吐量性能对比

本文假设接收机能分离当前接收端的干扰加噪声功率，当干扰与合法信号在不同信道时容易实现，但当两者在相同信道时，这一假设可能难以实现。因此，应当考察接收机无法完全分离的同信道干扰功率情况下的算法性能。图11展示了多路扫频干扰下同信道干扰分离率分别为0、25%、50%、75%和100%情况下算法的归一化吞吐量性能。实验结果表明，干扰分离率越高，算法收敛速度越快，但随着干扰分离率降低，算法的性能损失并不明显，即使分离率为0，算法仍能经700个通信时隙的迭代训练后使归一化吞吐量达到0.9以上。

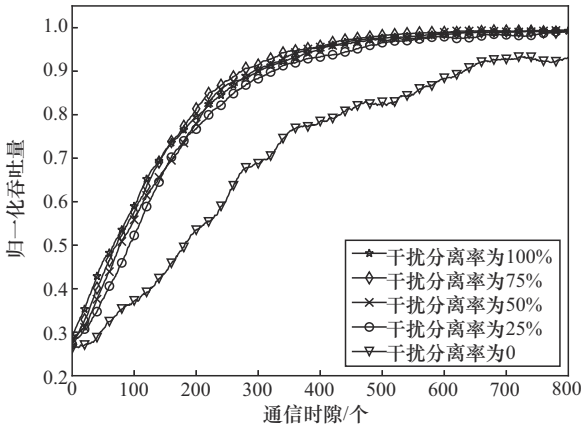



图11　不同干扰分离率下的归一化吞吐量性能对比

## 5　结束语

为加快DRL算法解决多域通信抗干扰问题的收敛速度，本文提出了一种基于相似性样本生成的DRL快速抗干扰算法。该算法利用通信抗干扰问题中状态–动作对之间的相似性，基于实际传输获得的状态–动作对生成与之相似的模拟状态–动作集，从而以较少的迭代步数获得大量训练样本，进而加快算法的学习进程和收敛速度。本文提出的针对DQN抗干扰算法的加速收敛思路具有在其他DRL抗干扰算法中应用的潜力。

## 参考文献：

[1] DON T. Principles of spread-spectrum communication systems[M]. Berlin: Springer, 2018.

[2] 姚富强. 通信抗干扰工程与实践[M]. 北京: 电子工业出版社, 2012.
YAO F Q. Communication anti-jamming engineering and practice[M]. Beijing: Publishing House of Electronics Industry, 2012.

[3] XIAO L, JIANG D H, XU D J, et al. Two-dimensional antijamming mobile communication based on reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2018, 67(10): 9499-9512.

[4] XIAO L, JIANG D H, WAN X Y, et al. Anti-jamming underwater transmission with mobility and learning[J]. IEEE Communications Letters, 2018, 22(3): 542-545.

[5] YUAN H C, SONG F, CHU X J, et al. Joint relay and channel selection against mobile and smart jammer: a deep reinforcement learning approach[J]. IET Communications, 2021, 15(17): 2237-2251.

[6] XIAO L, DING Y Z, HUANG J H, et al. UAV anti-jamming video transmissions with QoE guarantee: a reinforcement learning-based approach[J]. IEEE Transactions on Communications, 2021, 69(9): 5933-5947.

[7] LU X Z, XIAO L, NIU G H, et al. Safe exploration in wireless security: a safe reinforcement learning algorithm with hierarchical structure[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 732-743.

[8] YANG H L, XIONG Z H, ZHAO J, et al. Intelligent reflecting surface

assisted anti-jamming communications: a fast reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2021, 20(3): 1963-1974.

[9] LI Y Y, XU Y H, LI G X, et al. Dynamic spectrum anti-jamming access with fast convergence: a labeled deep reinforcement learning approach[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5447-5458.

[10] ZHOU Q, NIU Y T, XIANG P, et al. Intra-domain knowledge reuse assisted reinforcement learning for fast anti-jamming communication[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 4707-4720.

[11] YAO F Q, JIA L L, SUN Y M, et al. A hierarchical learning approach to anti-jamming channel selection strategies[J]. Wireless Networks, 2019, 25(1): 201-213.

[12] 孙岳, 李蓓蕾, 梁彩虹, 等. 块衰落信道下串联多链空间耦合LDPC码设计[J]. 西安电子科技大学学报, 2019, 46(2): 1-5, 28.
SUN Y, LI B L, LIANG C H, et al. Design of serial connecting multiple spatially coupled LDPC codes for block-fading channels[J]. Journal of Xidian University, 2019, 46(2): 1-5, 28.

[13] BOUZABIA H, DO T N, KADDOUM G. Deep learning-enabled deceptive jammer detection for low probability of intercept communications[J]. IEEE Systems Journal, 2023, 17(2): 2166-2177.

[14] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[15] ZENG L H, YAO F Q, ZHANG J Z, et al. Dynamic spectrum access based on prior knowledge enabled reinforcement learning with double actions in complex electromagnetic environment[J]. China Communications, 2022, 19(7): 13-24.

[16] YAO F Q, JIA L L. A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks[J]. IEEE Wireless Communications Letters, 2019, 8(4): 1024-1027.

[17] HUANG Y, ZHU X Y, WU Q H. Intelligent spectrum anti-jamming with cognitive software-defined architecture[J]. IEEE Systems Journal, 2023, 17(2): 2686-2697.

[18] 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.

[19] 张国敏, 张少勇, 张津威. 基于PPO算法的攻击路径发现与寻优方法[J]. 信息网络安全, 2023, 23(9): 47-57.
ZHANG G M, ZHANG S Y, ZHANG J W. Discovery and optimization method of attack paths based on PPO algorithm[J]. Netinfo Security, 2023, 23(9): 47-57.

[20] LIU X, XU Y H, JIA L L, et al. Anti-jamming communications using spectrum waterfall: a deep reinforcement learning approach[J]. IEEE Communications Letters, 2018, 22(5): 998-1001.

[21] 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.

## [作者简介]

**周权**（1991– ），男，江苏溧阳人，陆军工程大学与国防科技大学第六十三研究所联合培养博士生，主要研究方向为通信抗干扰技术。

**牛英滔**（1978– ），男，山东泰安人，博士，国防科技大学第六十三研究所副研究员、硕士生导师，主要研究方向为认知无线电、通信抗干扰技术。