基于样本信息熵辅助的深度强化学习抗干扰策略

李刚[1]，吴麒[1]，王翔[1]，罗皓[1]，李良鸿[2]，景小荣[2]，陈前斌[2]
（1.中国西南电子技术研究所，四川 成都 610036；2.重庆邮电大学通信与信息工程学院，重庆 400065）

摘　要：针对深度强化学习驱动的智能化干扰，提出了一种基于样本信息熵辅助的通信抗干扰策略。首先，基于神经网络对抗干扰策略网络和熵预测网络进行设计；接着，利用短时傅里叶变换对接收信号处理所形成的频谱瀑布图作为样本，对抗干扰策略网络和信息熵预测网络进行训练；之后，利用信息熵预测网络对抗干扰策略网络的训练样本进行精细化筛选，以提高训练样本的质量，最终提高抗干扰策略的在线决策能力和泛化性能。仿真结果表明，在干扰方干扰策略更新频率不超过通信方40倍且最大干扰通道数为3的极端条件下，基于样本信息熵辅助的通信抗干扰策略仍可取得至少61%的成功率；同时，与其他几种对比抗干扰策略相比，所提通信抗干扰策略具有更快的收敛速度。

关键词：抗干扰；深度强化学习；样本信息熵；智能干扰
中图分类号：TN975
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024161

Deep reinforcement learning-empowered anti-jamming strategy aided by sample information entropy

LI Gang[1], WU Qi[1], WANG Xiang[1], LUO Hao[1], LI Lianghong[2], JING Xiaorong[2], CHEN Qianbin[2]
1. Southwest Institute of Electronic Technology, Chengdu 610036, China
2. School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: For the deep reinforcement learning (DRL)-empowered intelligent jamming, an anti-jamming strategy aided by sample information entropy was proposed. Firstly, the anti-jamming strategy network and entropy prediction network were designed based on neural networks. Then, the anti-jamming strategy network and entropy prediction network were trained with the samples of the spectrum waterfall, which were formed by performing the short-time Fourier transform to the received signals. The information entropy prediction network was utilized for fine-grained selection of training samples of the anti-jamming strategy network to improve the quality of training samples, thereby enhancing the ultimate online decision-making capability and generalization performance of the anti-jamming strategy. The simulation results indicate that under the extreme condition where the jamming strategy update frequency does not exceed forty times that of the communication anti-jamming strategy and the maximum number of jamming channels is 3, the proposed anti-jamming strategy, aided by sample information entropy, can still achieve a success rate of at least 61%. Moreover, compared to several other anti-jamming strategies, the proposed strategy demonstrates faster convergence.

Keywords: anti-jamming, deep reinforcement learning, sample information entropy, intelligent jamming

收稿日期：2024–02–05；修回日期：2024–08–06
通信作者：李良鸿，leelianghong@outlook.com.
基金项目：国家自然科学基金资助项目（No.U23A20279）；中电天奥创新理论技术群基金资助项目（No.2022-1193-04-04）
Foundation Items: The National Natural Science Foundation of China (No.U23A20279), China Electronics Tian'ao Innovation Theory and Technology Group Fund (No.2022-1193-04-04)

0 引言

无线通信以电磁波作为信息传递的媒介，具有在空间开放传播的特性，而该特征使无线通信更容易受到恶意干扰的攻击。在早期，恶意干扰的样式相对固定，通信方很容易挖掘出干扰的变化规律，并能够有效地应对这类干扰。随着人工智能与干扰技术的融合，干扰设备逐渐具备感知通信环境、学习通信方行为并智能地调整干扰策略的能力[1]。干扰设备的智能化给无线通信带来了严重的威胁，因此，开展智能抗干扰技术的研究变得尤为紧迫。

近年来，学术界在智能抗干扰领域展开了广泛深入的研究。按照抗干扰所采用的模型，这些研究可分为2类：基于传统强化学习模型的智能抗干扰和基于深度强化学习模型的智能抗干扰。在基于传统强化学习模型的智能抗干扰研究方面，文献[2]基于三维Q学习，提出一种适用于无线通信多通道随机干扰场景的多参量智能抗干扰方法；通过学习信道状态、时隙选择和干扰功率之间的随机变化规律，该方法引导发射机在每个时隙选择最佳通道和功率配置，以达到抗干扰的目的。基于强化学习策略，Zhang等[3]提出了一种智能抗干扰中继通信系统，根据对通信环境特征的实时学习，该系统利用信噪比等关键参量对通信频率进行优化选择，以避免盲目跳频带来的额外开销。针对多用户场景，Yao等[4]采用马尔可夫博弈模型对通信抗干扰问题进行建模，进而基于强化学习提出了一种多智能体协作抗干扰算法。在非完美信道状态信息（CSI, channel state information）条件下，文献[5]基于多智能体Q学习方法，提出一种联合通道、功率和带宽等资源来实现通信抗干扰的优化模型，并通过引入信道重构技术和睡眠机制，来避免通信冲突问题。基于多臂赌博机（MAB, multi-armed bandit）模型，在太赫兹通信场景，文献[6]通过联合优化通道和传输时隙，来对抗智能化干扰。上述基于传统强化学习模型的抗干扰研究成果，当动作空间维度较低时，具有良好的抗干扰性能；然而，当动作空间非常庞大或连续时，该模型将面临维度灾难问题，且无法确保模型的收敛性。

针对传统强化学习模型在智能抗干扰方面的缺陷，部分学者开始研究如何利用深度强化学习模型来实现通信抗干扰。借助电磁环境感知所生成的频谱瀑布图，文献[7]利用深度卷积神经网络来近似各状态-动作对的Q函数，提出了一种基于深度Q网络（DQN, deep Q-network）的通信抗干扰方案。为了应对宽带通信中的动态频谱干扰，Li等[8]研究者提出了一种基于分层深度强化学习的高效抗干扰算法；首先，该算法利用宽带选择网络选取适当频带，然后通过频率选择网络在所选频带中进一步选取具体的频点，避免了抗干扰过程中可选频率数量过大的问题。为了有效应对多用户通信场景中外部的恶意干扰，同时避免用户间通道竞争所引起的互干扰问题，文献[9]提出了一种基于深度强化学习的智能抗干扰决策方法；在该方法中，基站将感知到的包含多用户和干扰源在内的频谱信息作为深度强化学习网络的输入，然后基于动态贪婪算法通过选择联合动作帮助通信用户智能地选择可用频段。文献[10]利用对手建模（OM, opponent modeling）研究了针对智能干扰的动态频谱接入抗干扰问题；在该研究中，利用最小最大深度Q网络来近似抗干扰效用，同时应用模仿学习来推理干扰器的策略，以提升抗干扰性能。通过引入经验回放和基于爬山策略的动态机制，文献[11]提出了一种基于深度强化学习的通信抗干扰智能决策方法。文献[12]针对异构卫星互联网中的智能干扰，以最小化抗干扰路由成本为目标，基于斯塔克尔伯格（Stackelberg）博弈和强化学习，提出一种空间抗干扰决策方案。文献[13]利用深度双Q学习，通过对信道访问和传输功率的联合优化，设计了一种更有效地规避各种干扰模式的通信传输策略。为了适应快速变化的干扰环境，Li等[14]基于标签化的深度强化学习，提出一种动态频谱抗干扰接入方案；该方案将强化学习周期分为2个阶段，即训练阶段和应用阶段，并以各可用通道的信干噪比作为软标签；该方案与现有基于深度强化学习的算法相比，收敛速度更快。在非完美的频谱感知环境下，文献[15]提出了一种智能抗干扰频谱接入方案，包括频谱感知、频谱补全、频谱学习和频谱接入等关键步骤。为了应对移动边缘计算网络中的智能干扰，Chen等[16]提出一种具有后决策状态的多用户智能博弈模型。文献[17]为有效对抗智能化干扰，提出了干扰主动防御技术体系架构，通过主动调整通信行为扰乱干扰学习过程，有效降低干扰效能，为实现“理解对手”“克制对手”“战胜对手”的目标提供了新途径。上述基于深度强化学习的抗干扰策略，尽管无须估计干

扰参数且解决了高维动作空间问题，然而，其通常在学习阶段将所有样本无差异地输入到网络进行训练。因此，这些策略会导致如下2个问题：首先，样本标签代价通常都比较高，尤其是在样本数量相对比较大的情况下；其次，从对深度强化学习网络进行训练的角度考虑，由于不同训练样本在确定性上存在差异，导致它们的信息熵不同，信息熵越低的样本，网络学到的知识就越有限，从而降低了网络的学习效率和泛化性能。

鉴于上述抗干扰策略的缺陷，本文针对深度强化学习驱动的智能干扰，提出了一种基于信息熵辅助的通信抗干扰策略。在该抗干扰策略中，主要贡献包括。

1) 所提通信抗干扰策略中，包括基于神经网络设计的抗干扰策略网络和信息熵辅助网络，前者主要负责在线抗干扰决策，后者主要负责筛选抗干扰策略网络训练所用的样本。

2) 所提通信抗干扰策略中，为了提高抗干扰策略网络的学习效率和泛化性能，根据熵预测网络对样本特征的信息熵预测结果，引入一种精细样本选择机制；同时，以最小化决策效益损失和熵预测损失的联合均方误差为目标，对抗干扰决策网络和熵预测网络的参数进行更新。通过如此处理，有效提高了网络训练样本的质量和关联性，使网络的抗干扰在线决策能力和泛化性能得以改善，并在一定程度上节省网络训练的时间成本。

3) 所提通信抗干扰策略中，采用深度强化学习来驱动2种智能化干扰，包括智能跟踪干扰和智能梳状干扰；同时，为确保智能干扰的水平和对抗的公平性，干扰策略网络采用与抗干扰策略网络相同的结构。此外，令干扰方策略更新频率为通信方的1倍、10倍、20倍、30倍、40倍、50倍、100倍，使通信/干扰双方之间的对抗过程呈现出更为复杂和动态的特征，以更准确地模拟真实通信对抗场景中的复杂电磁环境。

4) 仿真结果表明，在干扰通道数小于或等于3且干扰策略更新频率不超过通信方40倍的极端情况下，所提抗干扰策略仍可取得至少61%的平均成功率；同时，与文献[7]和文献[10]中的基于深度强化学习的抗干扰策略相比，当干扰通道数为3且干扰策略更新频率为通信方10倍时，所提抗干扰策略的平均成功率分别提高了10%和5%，从而验证了所提策略的可靠性和有效性。

1　系统模型

系统模型如图1所示，通信对抗场景包括一对通信收发机和一个智能干扰机，其中通信接收机端和智能干扰机均部署有智能体。通信接收机端智能体通过对电磁环境的实时感知和学习，获取干扰信号的特征，进而智能地选择通信通道。干扰机端的智能体同样具备感知电磁环境的能力，基于对通信方行为和规律的学习结果，进而通过干扰链路发送干扰信号来破坏通信收发机间的正常通信。通信收发机通过通信链路和控制链路分别完成数据传输和控制信息交换，其中，控制信息主要包括接收机确认收到发送机发送的信号而返回给发送机的ACK信息和通信策略交换信息等。

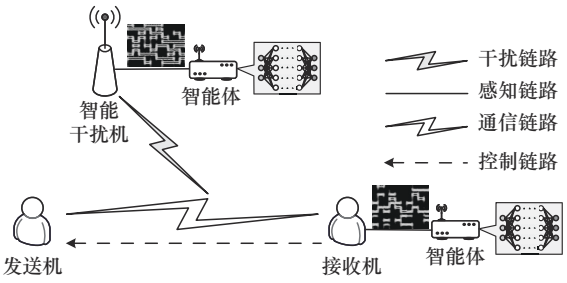


图1　系统模型

干扰链路和通信链路对应的信道系数用$h_j(t)$和$h_c(t)$表示；通信收发机与干扰机共享相同的频段$[f_L f_U]$，以带宽B将整个频段均匀地划分为$Z=\left\lfloor\frac{f_U-f_L}{B}\right\rfloor$个互不重叠的通道（或频带），由各通道中心频率所组成的集合$\Xi=\{f_0,f_1,\cdots,f_{Z-1}\}$，其中，$f_z=f_L+\left(z+\frac{1}{2}\right)\times B$表示第$z$个通道的中心频率。通信方和干扰方的动作集分别用$\Omega_c$和$\Omega_j$表示，$\Omega_c,\Omega_j\subset\Xi$；设定通信收发机和干扰机工作的总时隙数为$K$，且通信收发机在每时隙只选择一个通道进行通信。鉴于通信方和干扰方对抗过程中每个时隙都非常短暂，故根据文献[18]，假设$h_j(t)$和$h_c(t)$在K个时隙内保持不变，均服从平坦块衰落信道模型，即令$h_i(t)=\sqrt{d_i^{-\xi}}g_i(t)$，其中，$i\in\{c,j\}$，$d_i$为发送设备（智能干扰机或通信发送机）到通信接收机之间的距离，ξ为路径衰减指数，$g_i(t)$表示服从0均值单位方差的复高斯分布。

在第k个时隙，通信接收机端智能体感知到的

时域信号可建模为

$$r(t)=s(t)+j(t)+n(t) \tag{1}$$

其中，$s(t)$和$j(t)$分别表示通信信号部分和干扰信号部分；$n(t)$表示均值为0、方差为σ_n^2的复高斯背景白噪声；$t\in[(k-1)t_s,kt_s]$，t_s表示每个时隙的持续时间，$k=1,2,\cdots,K$。通信信号部分可进一步表示为

$$s(t)=h_c(t)\otimes\sqrt{P_c}\,\mathrm{rect}\left(\frac{t}{t_s}\right)m(t)\cdot \exp(\mathrm{j}(2\pi f_k t+\theta_c)) \tag{2}$$

其中，⊗表示卷积运算，P_c为通信信号发射功率，$m(t)$表示发送的通信符号，$f_k\in\Omega_c$和θ_c分别为第k个时隙通信载波的中心频率和初始相位，$\mathrm{rect}\left(\frac{t}{t_s}\right)=\begin{cases}1, & |t|\leqslant\frac{t_s}{2}\\ 0, & \text{其他}\end{cases}$表示矩形方波，$t_s$表示每个时隙的持续时间。干扰信号部分又可写作

$$j(t)=h_j(t)\otimes\sum_{m=0}^{M-1}\frac{\sqrt{P_j}}{M}\mathrm{rect}\left(\frac{t}{t_s}\right)\cdot \exp(\mathrm{j}(2\pi f_{k,m}t+\theta_j)) \tag{3}$$

其中，P_j为干扰信号的发射功率，M为同时干扰的频带（或通道）数，$f_{k,m}\in\Omega_j$和θ_j分别表示干扰信号的中心频率和相位。

对$r(t)$进行离散采样，得到$r(n)$，然后对其进行短时傅里叶变换（STFT, short-time Fourier transform），于是有

$$\mathrm{STFT}_r(m,f)=\sum_{n=m}^{n+N_L-1}r(n)w(n-m)\exp\left(-\mathrm{j}\frac{2\pi}{N_F}nf\right) \tag{4}$$

其中，$w(n)$表示长度为N_L的分析窗函数，N_F为频点数。根据式(4)，感知信号的时频功率分布为

$$\mathrm{TF}_r(m,f)=\left|\mathrm{STFT}_r(m,f)\right|^2 \tag{5}$$

根据式(2)，定义时频点(m,f)处的信号与干扰加噪声比（SJNR, signal-to jamming and noise ratio）为

$$\mathrm{SJNR}(m,f)=\frac{\mathrm{TF}_c(m,f)}{\mathrm{TF}_j(m,f)+\overline{N}} \tag{6}$$

其中，$\mathrm{TF}_c(m,f)$、$\mathrm{TF}_j(m,f)$和$\overline{N}$分别表示接收信号中的通信信号功率、干扰信号功率和噪声功率。

根据式(5)计算时频点(m,f)处功率，并定义矢量$\boldsymbol{v}_s=[\mathrm{TF}_r(m,f_0),\cdots,\mathrm{TF}_r(m,f_{N_F-1})]^{\mathrm{T}}$，将$\boldsymbol{v}_s$回溯$K-1$个历史时刻，构建时-频谱状态矩阵$\boldsymbol{W}_s=[\boldsymbol{v}_s,\boldsymbol{v}_{s-1},\cdots,\boldsymbol{v}_{s-K+1}]$，将其定义为时刻$t$的状态$s_t$，结果形成如图2(a)所示的时-频状态图，其中横轴和纵轴分别表示时间t和频率f；进一步，将其转化为热力图，即如图2(b)所示的频谱瀑布图。由于在该研究领域各研究小组对数据样本的保密性，与文献[10]采用快速傅里叶变换（FFT, fast Fourier transform）获得的频谱瀑布图不同，本文则根据STFT生成的频谱瀑布图来生成本研究中的数据样本。观察频谱瀑布图中的灰度特征，明显存在受干扰的通信信号、干扰信号和未受干扰的通信信号及背景噪声。干扰机端智能体采用与通信接收机端智能体类似的操作来确定干扰机的状态。

2　智能抗干扰策略设计

与模式化干扰相比，深度强化学习驱动的智能干扰更复杂，且具有动态性，因此，在通信和干扰双方博弈的过程中，快速做出正确的抗干扰决策对通信方至关重要。为了实现该目标，本节将利用深度强化学习，提出一种基于样本信息熵（SIE, sample information entropy）辅助的通信抗干扰策略。在该策略中，基于神经网络对抗干扰策略网络和信息熵辅助网络进行设计，旨在借助信息熵辅助网络对抗干扰策略网络的训练样本进行精细化筛选，从而达到降低网络训练时间成本、

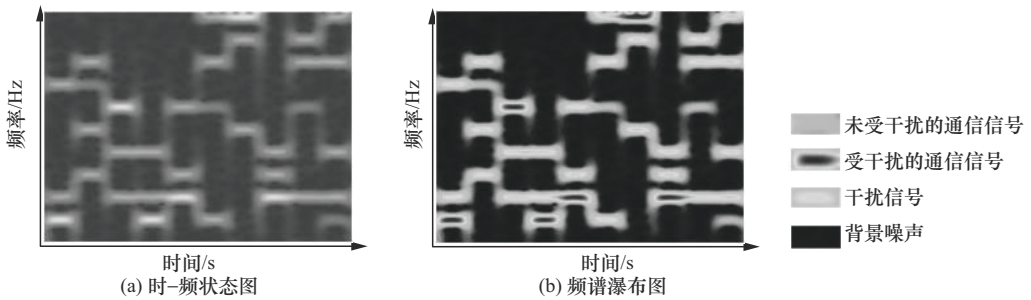


图2　时-频状态图与频谱瀑布图

提高抗干扰策略的在线决策能力和泛化性能的目的。此外，为确保智能干扰的智能化水平，干扰机的干扰策略网络采用了与抗干扰策略网络相同的结构。

为了描述复杂通信场景下的智能干扰对抗问题，本文将通信和干扰双方之间的对抗过程建模为马尔可夫博弈模型，即将对抗过程视为2个博弈者参与的系统，对应的博弈六元组为 $g=\left\langle S,\Omega_c,\Omega_j,\boldsymbol{P},R,\lambda\right\rangle$，其中，$S$表示状态集，本文将历史频谱瀑布图作为状态集；$\boldsymbol{P}$表示状态转移概率矩阵，表示对抗系统从一个状态转移到另一个状态的概率分布，反映状态之间的相互关系；R表示用户奖励函数，本文选用效益函数作为奖励函数，用于评估博弈中各个状态执行完动作后的效用或价值；λ表示奖励折扣因子，用于衡量博弈过程中对未来奖励的重视程度。

设定通信解调门限为β_{th}，即当$\text{SJNR}(m,f)\geqslant\beta_{\text{th}}$时，通信方则成功避开干扰。根据$t$时刻$\text{SJNR}^t$，设计通信效益函数$\mu_c(t)$，即

$$\mu_c(t)=\begin{cases}\mu_0, \text{SJNR}^t\geqslant\beta_{\text{th}}\\ -\mu_0, \text{SJNR}^t<\beta_{\text{th}}\end{cases}\tag{7}$$

其中，μ_0表示通信成功对应的奖励。本文设定智能干扰方和通信方之间的博弈为零和博弈，即博弈双方的总体收益为零，即干扰方的效益函数$\mu_j(t)=-\mu_c(t)$。

令通信策略和干扰策略分别为π_c和π_j，则通信方最佳策略π_c^*可通过最大化长期折扣奖励得到

$$\pi_c^*=\arg\max_{\pi_c}\text{E}_{s\sim P,a_c\sim\pi_c,o_j\sim\pi_j}\left[\sum_{t=0}^{\infty}\lambda_c(t)r_t(s,a_c,o_j)\right]\tag{8}$$

在实际干扰对抗过程中，通信方和干扰方都采用非透明的策略，有学者通过寻找纳什均衡（NE, Nash equilibrium）策略求解上述问题[19]，当采用NE策略时，累积奖励不再发生变化。然而，由于观测状态信息不完美或学习能力有限，干扰者可能会偏离NE策略[10]。

为获取最优策略π_c^*，可利用深度强化学习来挖掘频谱瀑布中的特征信息，根据这些特征信息来拟合最大累计奖励对应的Q函数[7,10]。然而，对于深度学习网络，不同样本的确定性不同，即信息熵不同，信息熵越低的样本，网络学习到信息越小，若直接从样本池中随机采样批量样本进行网络训练，既增加了样本训练成本，又降低了网络学习效率。

为此，本文提出一种基于样本信息熵辅助的抗干扰策略，其网络模型如图3所示，包括抗干扰策略网络和熵预测网络，其中，抗干扰策略网络用于在线输出最佳抗干扰策略，熵预测网络则在网络训练过程中用于筛选信息熵较大的训练样本。通过联合训练这2个网络，确定抗干扰决策网络参数θ和熵预测网络参数ϕ；需要强调的是，熵预测网络仅用于辅助抗干扰网络的有效训练，训练结束后，熵预测网络将不参与在线抗干扰决策过程。在时刻t，

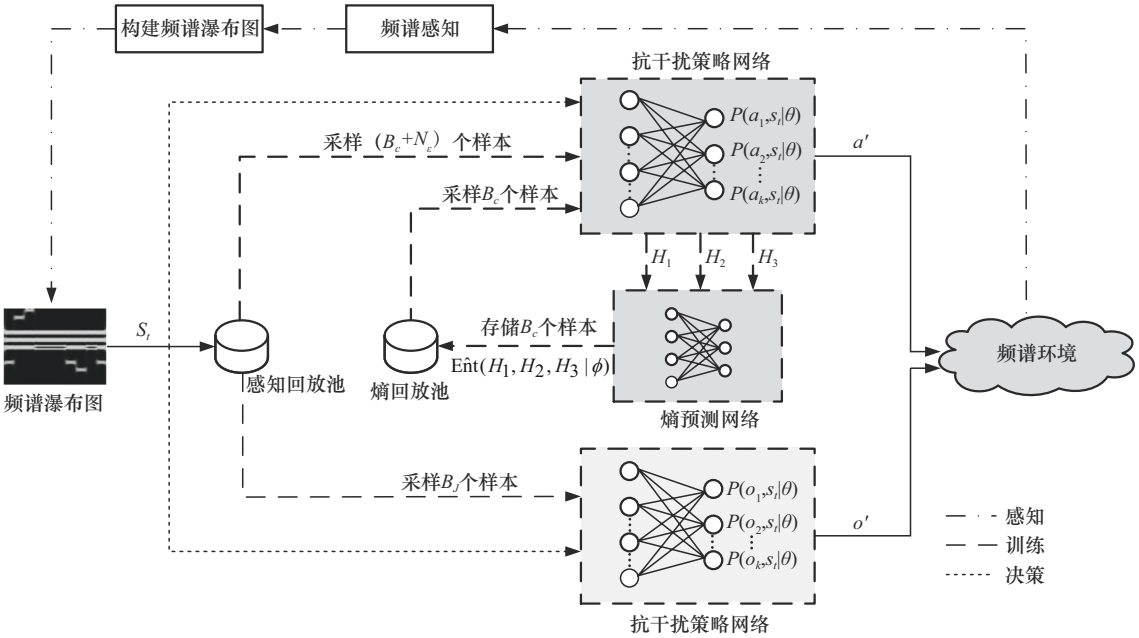


图3　基于样本信息熵辅助的抗干扰策略的网络模型

通信接收机端的智能体通过对电磁环境的感知确定当前状态s_t，进而利用ε-贪婪算法执行动作a_t，获得奖励r_t，同时感知下一时刻环境状态s_{t+1}，构成样本$e(s_t,a_t,r_t,s_{t+1})$，将其存放在感知回放池Γ中；当Γ中样本满足批量采样要求时，从Γ中随机抽取(B_c+N_ε)个样本，作为抗干扰策略网络的输入样本；经特征提取层输出的特征图则被送入熵预测网络，根据预测结果，仅保留B_c个信息熵较大的样本，存放于熵回放池Ψ；随后训练过程中，从Ψ中随机抽取批量样本用于抗干扰策略网络训练；实际上，熵预测网络结构的复杂度远低于抗干扰决策网络。

干扰策略网络采用与抗干扰策略网络相同的结构制定干扰策略，企图干扰通信收发机间的正常通信。需要指出的是，在通信方和干扰方双方博弈过程中，其对应的抗干扰策略网络和干扰策略网络，依据各自对电磁环境的感知和推理结果来判断对方是否进行了策略调整，进而决定是否需要对各自网络的参数进行更新。下面对图3中各类网络进行详细说明。

2.1 抗干扰策略网络设计

抗干扰策略网络由3个卷积层（CL, convolutional layer）、2个全连接层（FCL, fully connected layer）和一个softmax层（SL, softmax layer）组成，分别负责特征提取、线性表示以及动作概率输出等关键任务。3个CL分别用CL_1、CL_2和CL_3表示，2个FCL分别用FCL_1和FCL_2表示。

2.2 熵预测网络设计

在抗干扰网络训练阶段，借助于信息熵辅助网络对抗干扰策略网络的训练样本进行精细化筛选，以达到降低网络训练时间成本、提高抗干扰策略的在线决策能力和泛化性能的目的。熵预测网络结构如图4所示。

熵预测网络由3个基本块单元和一个包含10个神经元的SL组成。3个基本块单元包括全局平均池化层（GAPL, global average pooling layer）、FCL，以及激活层（AL, activation layer）。由于抗干扰策略网络中处于不同深度的CL从输入频谱瀑布图中提取的特征信息存在差异，为了充分利用这些差异性特征信息，将抗干扰策略网络中3层CL提取出中间特征图H_1、H_2和H_3分别作为熵预测网络中3个基本块单元的输入，由每个基本块单元独立完成特征提取任务。该方式避免了熵预测网络对原始频谱瀑布图执行重复性特征提取操作，有效降低了熵预测网络的运算量。同时，为了充分利用3个基本块单元提取的不同特征信息，对其进行特征拼接操作，即特征融合，以便SL利用融合后的特征进行准确的信息熵预测，其预测值用符号Ent表示。此外，熵预测网络的训练标签，由训练过程中抽取的批量样本经抗干扰策略网络SL输出的预测熵来确定。

需要指出的是，图4中包含3个基本块单元的熵预测网络结构与抗干扰策略网络中的3层CL相对应；实际中，熵预测网络中基本块单元的数量可根据抗干扰策略网络中CL数量的变化灵活调整，与之对应，后续式(14)等也据此进行相应调整。

2.3 抗干扰策略网络和熵预测网络训练

通信方与干扰方之间的对抗过程属于零和博弈问题，在这种情况下，通信方会选择策略以最小化干扰方获得的价值，同时最大化自身的价值，以达到一种博弈平衡。为此采用min-max深度强化学习方法，通过训练网络模型来逼近状态值函数，其对应的值函数表示为

$$V_{c,i}^{*}(s_i)=\max_{\pi_c}\min_{a_{j,i}\in\Omega_c}\sum_{a_{c,i}\in\Omega_c}\pi_c(a_{c,i}|s_i)Q(s_i,a_{c,i},a_{j,i},\bar{\theta}) \tag{9}$$

其中，$Q(s_i,a_{c,i},a_{j,i},\bar{\theta})$表示在状态$s_i$执行动作对

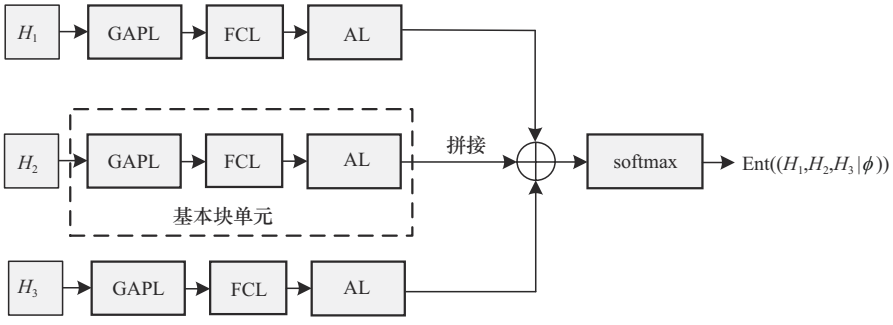


图4 熵预测网络结构

$(a_{c,i},a_{j,i})$抗干扰策略网络需要逼近的Q值函数，$\bar{\theta}$代表目标网络的参数。对上述问题采用线性规划方法来求解，以便在最小化干扰收益的同时最大化自身收益。对问题式(9)完成线性规划后，即可联合动作奖励确定目标Q值，即

$$y_{c,i}(a_{c,i},r_{c,i},\bar{\theta},s_i)=r_{c,i}+\lambda_c V_{c,i}^{*}(s_i) \tag{10}$$

其中，$r_{c,i}$表示通信方执行动作$a_{c,i}$所获得的奖励。

根据目标Q值和抗干扰策略网络中FCL2输出的估计Q值$Q(s_i,a_{c,i},\pi_c|\theta)$，确定决策效益损失；在此基础上，联合熵预测损失确定联合MSE损失，进而对抗干扰策略网络和熵预测网络进行联合训练。这不但减少了训练时间和资源消耗，提高了训练效率，而且利用正则化处理防止单一网络过拟合，从而提高了模型的泛化能力。

$$L(s|\theta,(H_1,H_2,H_3)|\phi)=\underbrace{\frac{1}{B_c}\sum_{i=1}^{B_c}\left[y_{c,i}(a_{c,i},r_i,\theta_i,s_i)-Q(s_i,a_{c,i},\pi_c|\theta)\right]^2}_{\text{决策效益损失}}+\underbrace{\frac{\gamma}{B_c}\sum_{i=1}^{B_c}\left[\mathrm{Ent}(s_i|\theta)-\mathrm{Ent}\left((H_{i,1},H_{i,2},H_{i,3})|\phi\right)\right]^2}_{\text{熵预测损失}} \tag{11}$$

其中，$H_{i,j}$表示输入样本i经抗干扰策略网络中CL_j后的输出特征图，$j=1,2,3$；$\gamma\in(0,1]$表示损失权重因子；$\mathrm{Ent}(s_i|\theta)$表示第i个样本作为抗干扰策略网络的输入时该网络SL输出的预测概率所对应的信息熵，简称为预测熵；$\mathrm{Ent}((H_{i,1},H_{i,2},H_{i,3})|\phi)$表示将$H_{i,1}$、$H_{i,2}$和$H_{i,3}$作为熵预测网络的输入，该网络SL输出的预测概率所对应的信息熵。具体地，

$$Q(s_i,a_c,\pi_c|\theta)=f_\theta(s_i,a_c|\pi_c) \tag{12}$$

其中，$f_\theta(s_i,a_c|\pi_c)$表示在状态s_i，根据通信策略π_c执行动作a_c后由抗干扰策略网络拟合出的估计值。

$$\mathrm{Ent}(s_i|\theta)=-\sum_{i'}P_\theta(y_{i'}|s_i)\log P_\theta(y_{i'}|s_i) \tag{13}$$

其中，$P_\theta(y_{i'}|s_i)$表示在状态s_i和网络参数为θ时，抗干扰策略网络输出的预测Q值为$y_{i'}$的概率。

$$\mathrm{Ent}((H_{i,1},H_{i,2},H_{i,3})|\phi)=-\sum_{i'}P_\phi(\hat{y}_{i'}|(H_{i,1},H_{i,2},H_{i,3}))\log P_\phi(\hat{y}_{i'}|(H_{i,1},H_{i,2},H_{i,3})) \tag{14}$$

其中，$P_\phi(\hat{y}_{i'}|(H_{i,1},H_{i,2},H_{i,3}))$表示在状态$s_i$和网络参数为$\phi$时，将$H_{i,1},H_{i,2},H_{i,3}$作为熵预测网络的输入所得Q值为$\hat{y}_{i'}$的概率。

通过最小化联合MSE损失，对抗干扰策略网络和熵预测网络进行训练。为此，构建如下优化问题。

$$\arg\min_{\theta,\phi}L(s|\theta,(H_1,H_2,H_3)|\phi) \tag{15}$$

对该问题采用随机梯度下降（SGD, stochastic gradient descent）算法进行求解，对应地，相关参数迭代更新如式(16)所示。

$$\begin{cases}\theta_d=\theta_{d-1}-\eta_\theta\sum_{i=1}^{B_c}\nabla_{\theta_{d-1}}L(s_i|\theta_{d-1},(H_{i,1},\cdots,H_{i,3})|\phi_{d-1})\\ \phi_d=\phi_{d-1}-\eta_\phi\sum_{i=1}^{B_c}\nabla_{\phi_{d-1}}L(s_i|\theta_{d-1},(H_1,\cdots,H_3)|\phi_{d-1})\end{cases} \tag{16}$$

其中，θ_d和ϕ_d分别表示第d次迭代更新后抗干扰策略网络和熵预测网络的参数，η_θ和η_ϕ分别表示抗干扰策略网络和熵预测网络参数的学习率。

根据上述分析，所提基于样本信息熵辅助的抗干扰策略步骤如算法1所示，包含样本预采集和网络训练2个阶段。在样本预采集阶段，首先智能体根据当前环境感知信号，确定环境状态s_t，通过ε-贪心算法选择通信通道（或动作）a_t，同时计算奖励值$r_t=\mu_c(t)$；接着，感知下一时刻频谱状态s_{t+1}，并将对应的样本$e(s_t,a_t,r_t,s_{t+1})$存储在感知回放池中；当感知回放池中样本满足批量采样要求时，从中随机抽取(B_c+N_ε)个样本，作为抗干扰策略网络的输入样本；经特征提取层输出的特征图则被送入熵预测网络，根据熵预测结果，仅保留B_c个信息熵较大的样本，存放于熵回放池中。当熵回放池中的样本数量达到最低要求时，算法进入网络训练阶段。在训练过程中，首先利用熵预测网络完成样本筛选；接着采用SGD算法求解联合优化问题式(15)，直至训练误差满足门限要求或达到最大迭代次数时，当前训练过程停止；然后保存训练完成的模型，用于在线抗干扰决策。随着对抗过程的持续进行，感知回放池和熵回放池随着电磁环境的变化不断地更新，网络参数因而不断地调整，从而在对抗中实现动态训练。

需要说明的是，网络训练过程也是一个与环境交互的过程。熵回放池和感知回放池中的样本按照先进先出（FIFO, first in first out）的原则存入各自的回放池。当样本数量超过各自回放池设定的阈值

时，最早存入的样本将被最新获得的样本替换，因此，用于网络训练的批量样本并非固定不变。

算法 1 所提基于样本信息熵辅助的抗干扰策略步骤

初始化 设置 θ，ϕ，η_θ，η_ϕ 初始值和最大迭代次数 D，令 $\Gamma=\varnothing$，$\Psi=\varnothing$

1) 感知当前频谱环境状态 s_t；

2) 根据 ε-贪婪算法执行动作 a_t，其中

$$a_t=\begin{cases}\arg\max Q(s_t,a_c|\theta), & 1-\varepsilon\\ \text{随机选择} a_c\in\Omega_c, & \varepsilon\end{cases}$$

3) 根据通信效益函数式(7)计算奖励值 $r_t=\mu_c(t)$；

4) 感知下一时刻频谱状态 s_{t+1}；

5) 存储样本至感知回放池，即令 $\Gamma=e(s_t,a_t,r_t,s_{t+1})\cup\Gamma$；

6) if $|\Gamma|\geqslant B_c+N_\varepsilon$

7) 从 Γ 中随机采样 B_c+N_ε 个样本输入到抗干扰策略网络；并将 3 个 CL 的输出特征图输入到熵预测网络；

8) end if

9) 根据熵预测网络结果，存储 B_c 个具有较高预测熵的样本到熵回放池中，即 $\Psi=e(s_i,a_i,r_i,s_i')\cup\Psi$；

10) 重复步骤 1)~步骤 9)生成训练样本；

11) 从 Ψ 中随机采样 B_c 个样本训练抗干扰策略网络；

12) 根据式(10)计算目标 Q 值，并由式(11)计算 MSE 损失；

13) 采用 SGD 算法实现 MSE 损失 $L(s|\theta,(H_1,H_2,H_3)|\phi)$ 最小化，同时根据式(16)更新网络参数 θ 和 ϕ；

14) 重复步骤 11)~步骤 13)，直到 $L(s|\theta,(H_1,H_2,H_3)|\phi)\leqslant\xi$ 或达到最大迭代次数 D；

输出 训练结束的抗干扰策略网络参数 θ。

在算法 1 中，$e(s_t,a_t,r_t,s_{t+1})$ 中下标 t 表示时刻；$e(s_i,a_i,r_i,s_i')$ 中下标 i 表示回放池中的样本索引号。

2.4 干扰策略网络设计与训练

为了确保通信方与干扰方在对抗过程中的公平性，干扰策略网络采用与抗干扰策略网络相同的结构，因此，其网络参数参照算法 1 进行配置。这样处理并非意味着通信方需要借助于训练来获得干扰网络结构，实际上，通信方和干扰方之间的对抗过程属于非合作马尔可夫博弈，通信方是无法获取干扰策略网络结构及相关参数的。对干扰机的更新策略，通信方也是无法提前获得的，只能根据对当前时隙电磁环境的感知和推理结果以一种非直接方式来判断干扰机的策略。

为模拟智能化干扰的动态变化过程，干扰机处的智能体内部嵌入深度强化学习模型，以最大化长期折扣效益为目标对其进行训练。为此，按照式(17)来确定最佳干扰策略。

$$\pi_j^*=\arg\max_{\pi_j}\mathrm{E}_{s\sim P,o_j\sim\pi_j,a_c\sim\pi_c}\left[\sum_{t=0}^{\infty}\lambda_j(t)r_{j,t}'(s_t,\bar{o}_j,a_c)\right]\tag{17}$$

其中，$r_{j,t}'(s_t,\bar{o}_j,a_c)$ 表示时刻 t 在状态 s_t 执行动作集 $\bar{o}_j$ 后干扰机所获得的奖励值。为确保公平性，干扰方采用与通信方目标 Q 值计算方式相同的方法，即

$$y_{j,i}(a_{j,i},r_{j,i},\bar{\theta},s_i)=r_{j,i}+\lambda_jV_{j,i}^*(s_i)\tag{18}$$

其中，s_i 表示干扰机的状态（本文假设干扰机处的智能体和通信接收机处的智能体具有相同的环境感知能力，因此均采用 s_i 来表示状态）。

干扰网络亦采用 MSE 作为损失函数，定义为

$$L(s_{t,i};a_{j,i}|\omega)=-\frac{1}{B_j}\sum_{i=1}^{B_j}\Big[y_{j,i}(a_{j,i},r_{j,i},\bar{\omega},s_i)-Q(s_i,a_{j,i},\pi_j|\omega)\Big]^2\tag{19}$$

其中，B_j 表示干扰策略网络的批量训练样本数，$a_{j,i}$ 表示在第 i 轮训练结束后，模型所预测的干扰动作。通过最小化上述损失函数，构建以下优化问题

$$\arg\min_{\omega}L(s_i;a_{j,i}|\omega)\tag{20}$$

对问题式(20)使用 SGD 算法，来完成干扰网络训练，其网络参数按照式(21)迭代更新。

$$\omega_k=\omega_{k-1}-\eta_\omega\sum_{i=1}^{B_j}\nabla_{\omega_{k-1}}L(s_i;a_{j,i}|\omega_{k-1})\tag{21}$$

其中，ω_k 表示第 k 次迭代更新后干扰策略网络的参数，η_ω 表示干扰策略网络参数学习率。

当干扰策略网络的损失值满足门限要求或达到最大迭代次数时，训练过程停止，并保存模型。随后，干扰方即可调用该模型来实现在线干扰。

3 仿真分析

本节对本文提出的基于样本信息熵辅助的抗干扰策略进行数值仿真。硬件环境为 Intel(R) Core (TM) i9-10900X CPU 和 NVIDIA GeForce RTX 3090 GPU。软件环境为 Pytorch 1.12 框架和 CUDA

11.6 计算平台。操作系统为 64 位 Windows 10 专业版。仿真中，通信收发机采用端到端传输模式，干扰机具备发射智能跟踪干扰和智能梳状干扰的能力。智能梳状干扰与传统梳状干扰的频谱瀑布图类似，均呈梳齿状；但智能梳状干扰由基于深度强化学习的干扰策略网络根据学习到的通信收发机的规律，有目的地选择某几个等间隔频带（通道）进行干扰。智能追踪干扰则选择非等间隔的频带进行干扰。

经多次优化和验证，抗干扰策略网络参数配置如表 1 所示。对于熵预测网络，3 个 FCL 的神经元数量分别设置为 16、32 和 32，AL 采用 ReLU 函数；抗干扰策略网络、熵预测网络和干扰策略网络的学习率分别设为 0.0 001、0.0 002 和 0.0 001；抗干扰策略和干扰策略采用相同的贪婪率（训练初期和后期分别设为 0.9 和 0.01）；$B_C = 32$，$N_e = 16$，奖励折扣因子 $\lambda_c = \lambda_j = 0.95$，感知回放池和熵回放池大小分别设置为 15 000 和 10 000，干扰策略网络对应的训练样本池大小设为 10 000。

表 1　　抗干扰策略网络参数配置

层	核尺寸	步长	通道数/个	神经元数/个
CL_1	2×2	2	16	—
CL_2	2×2	4	32	—
CL_3	2×2	2	64	—
FCL_1	—	—	—	512
FCL_2	—	—	—	256
SL	—	—	—	10

为了模拟动态复杂的干扰环境，干扰机干扰策略的更新频率等于通信收发机的 10 倍；在双方通信干扰的对抗过程中，每经 50 轮干扰，干扰机将以 50% 的概率在智能跟踪干扰和智能梳状干扰之间进行切换。干扰信号功率、通信信号功率和通信接收机解调门限分别设定为 30 dBm、0 和 10 dBm。通信带宽设定为 200 MHz，被分成 10 个互不交叠的频带（或通道），每个频带 20 MHz；频谱状态包含 $K = 10$ 个时隙；干扰机处智能体上部署的干扰策略网络将根据频谱感知和推理结果确定干扰动作，即选择单个或多个频带对通信收发机进行干扰。

为了验证所提抗干扰策略的有效性，将其与传统的基于能量检测（ED, energy detection）的经典抗干扰策略、基于 Q 学习的抗干扰策略，以及 2 种经典的基于深度强化学习的抗干扰策略，即基于 DQN 的抗干扰策略[7]和基于 OM 抗干扰策略[10]进行对比（仿真图中分别以 ED 策略、Q 学习策略、DQN 策略、OM 策略进行标识）。其中，基于 ED 的抗干扰策略根据实时感知频谱状态，选择能量最低的通道作为下一时隙通信收发机的通信通道；基于 Q 学习的抗干扰策略以瞬时频谱数据作为输入，智能体通过贪婪算法执行通信动作探索频谱状态直到 Q 表收敛，最后将收敛后的 Q 表作为抗干扰策略。

所提抗干扰策略与智能梳状干扰之间的对抗过程及样本信息熵的变化情况如图 5 所示。由图 5 可知，随着对抗过程的推进，频谱状态处于动态变化过程中；同时，随着通信方抗干扰策略网络对干扰机发射的智能梳状干扰规律的逐渐洞悉，样本信息熵的值在逐渐减小，特别地，在第 12 回合后，样本信息熵达到最小值 0.28，这表明通信方已基本掌握了智能梳状干扰的特征和规律。同样地，干扰机处智能体中的干扰策略网络在对抗过程中也学习到了通信方的通信策略，于是对干扰通道进行了切换，此时通信方的抗干扰策略网络需要重新学习干扰机的干扰模式和规律，导致第 13 回合的样本信息熵的值陡增。也就是说，在通信收发机与干扰机动态对抗过程中，感知回放池中样本的信息熵亦在动态变化。而所提抗干扰策略利用熵预测网络筛选出信息熵相对较大的样本，对后续抗干扰策略网络可进行更为有效的训练，从而使通信收发机能够更快地适应动态变化的干扰环境。

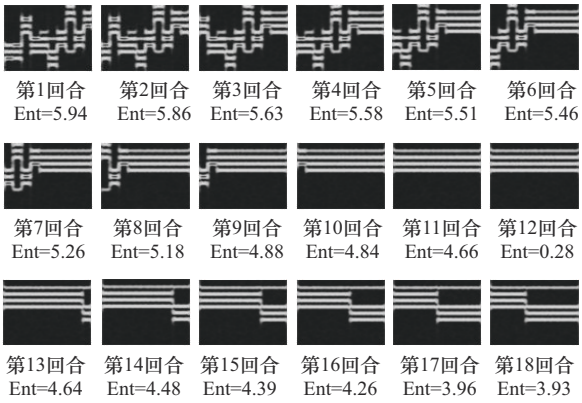


图 5　所提抗干扰策略与智能梳状干扰之间的对抗过程及样本信息熵的变化情况

所提抗干扰策略与智能跟踪干扰之间的对抗过程及样本信息熵的变化情况如图6所示。从图6中可明显看出，在首个回合对抗中，干扰机成功干扰到通信收发机的通信通道；随后，干扰机和通信收发机分别根据各自策略网络确定的干扰策略和通信策略对干扰通道和通信通道进行了切换。在第2回合，通信收发机利用通信策略成功地躲避过智能跟踪干扰。在随后的几个对抗回合中，随着通信接收机端智能体对智能跟踪干扰特征和规律的掌握，样本信息熵逐渐减小，直到第13回合，样本信息熵达到最小值0.29，表明通信收发机已基本掌握了智能跟踪干扰的特征和规律。同时，干扰机端智能体根据前几个对抗回合中所学习到的通信方的规律，在第14回合，干扰机选择通道3、4、6进行干扰。然而，通信收发机根据抗干扰策略的决策结果将通信通道切换至通道8，从而成功地躲过了干扰。此时，由于新的干扰特征信息出现，频谱状态图发生了变化，导致样本信息熵从0.29骤增至5.94。在随后的对抗回合中，通信收发机需要重新学习干扰特征和规律。随着对智能跟踪干扰的规律和特征的逐渐掌握，样本信息熵逐渐减小。

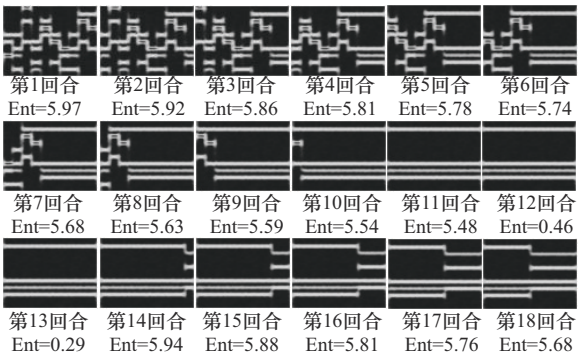


图6　所提抗干扰策略与智能跟踪干扰之间的对抗过程及样本信息熵的变化情况

所提抗干扰策略与2种智能干扰之间的对抗过程及样本信息熵的变化情况如图7所示。由图7可知，在前3个对抗回合，通信收发机成功地躲开了智能追踪干扰；在第4回合，干扰机发射的智能追踪干扰成功地干扰到通信收发机的正常通信，此时信息熵有所增加；在第5回合，通信收发机利用抗干扰策略进行了通道切换，从而避免受到干扰。在随后的对抗回合中，样本信息熵逐渐减小，直到第15回合减至0.38，表明通信收发机已基本掌握了智能跟踪干扰的样式和规律。在第16回合，干扰机将干扰样式切换为智能梳状干扰。同时，通信方根据抗干扰策略网络的决策将通信通道切换至8，即干扰机发射的智能梳状干扰并未成功；但由于新的干扰样本的出现，样本信息熵骤增至5.92。在随后的对抗回合中，通信方需要重新学习智能梳状干扰特征和规律；随着通信方对智能梳状干扰规律和特征的逐步学习，对应的样本信息熵也在逐渐减小。

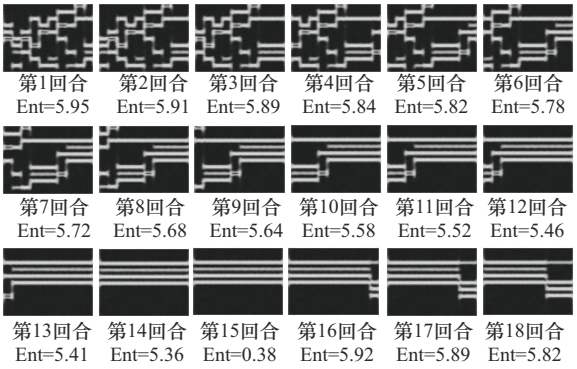


图7　所提抗干扰策略与2种智能干扰之间的对抗过程及样本信息熵的变化情况

图5~图7的仿真结果表明，随着通信方与干扰方对抗过程的延续，通信方的智能体在与环境交互过程中，获得的样本一直处于动态更新中，其对应信息熵亦处于动态变化中，进一步说明要提高网络的学习效率，对网络训练样本采取精细化筛选的必要性。由于采用无差别样本进行训练常规的深度网络训练，需要对每个用于训练的样本打标签，因而样本标签代价通常都比较高，而本文样本形成方式及网络训练过程规避了该问题。

当干扰机同时干扰的通道数为3时，所提抗干扰策略与4种对比抗干扰策略的平均成功率随迭代次数变化曲线如图8所示。由图8可知，经过10 000次的迭代，OM策略的平均成功率约为80%，DQN策略的成功率约为75%，而所提抗干扰策略实现了85%的成功率，这主要是所提抗干扰策略借助信息熵预测网络过滤掉了训练效率低的冗余样本，从而使其在抗干扰方面表现出更佳的性能。Q学习策略和ED策略抗干扰的平均成功率波动性较大，同时，相比于其他3种抗干扰策略，它们的平均成功率也相对较低，这表明这2种策略在对抗干扰方面能力非常有限。

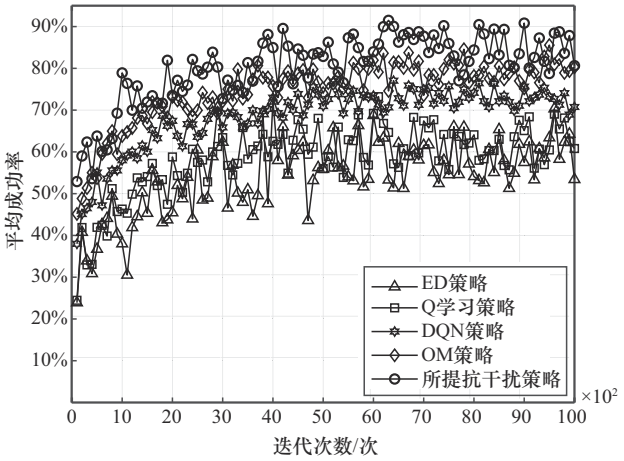


图8　所提抗干扰策略与4种对比抗干扰策略的平均成功率随迭代次数变化曲线

MSE 损失随迭代次数变化曲线如图 9 所示。由图 9 可知，随着迭代次数的增加，3 个抗干扰策略网络训练的 MSE 损失逐渐减小，虽然存在一定波动，但最终均趋于稳定。在稳定状态下，所提抗干扰策略的网络相比于其他 2 种抗干扰策略网络展现出更小的 MSE 损失值；同时，所提抗干扰策略网络训练的收敛速度更快，大约在 400 次迭代后趋于收敛状态，这主要归功于熵预测网络对训练样本的精细化选择，提高了抗干扰决策网络训练样本的质量，进而提高了抗干扰策略网络的学习效率。

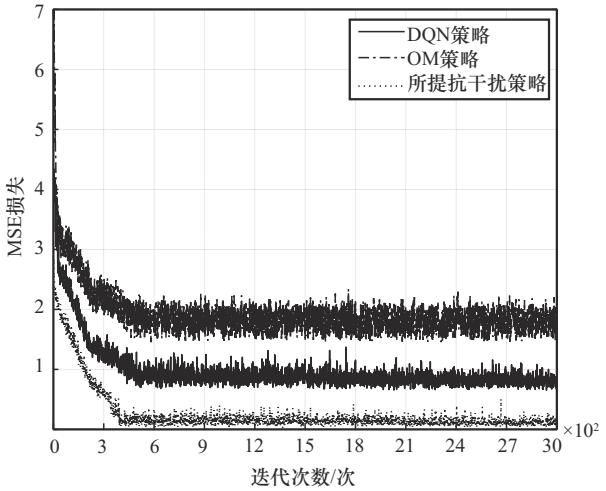


图9　MSE损失随迭代次数变化曲线

所提抗干扰策略与 4 种对比策略的平均成功率随干扰通道数变化曲线如图 10 所示，图中每个点为 10 000 次对抗结果的统计平均值。由图 10 可知，当干扰机只干扰一个通道时，几种抗干扰策略均表现出良好的抗干扰性能，其中所提抗干扰策略的平均成功率高达 96%，其次为 OM 策略，其平均成功率约为 92%。当干扰机同时干扰的通道数为 3 时（即通信方受到相对非常恶劣的干扰），所提抗干扰策略仍然保持约 85% 的平均成功率；当干扰的通道数增至 4 时，所提抗干扰策略的平均成功率下降至 68%，而 OM 策略的平均成功率为 62%，DQN 策略、Q 学习策略和 ED 策略的平均成功率已不足 60%。这是因为随着干扰通道数量的增加，抗干扰策略网络推荐的通信通道被干扰的概率也增加，因而导致抗干扰成功率降低。然而，就性能而言，所提抗干扰策略仍然明显优于其他抗干扰策略。

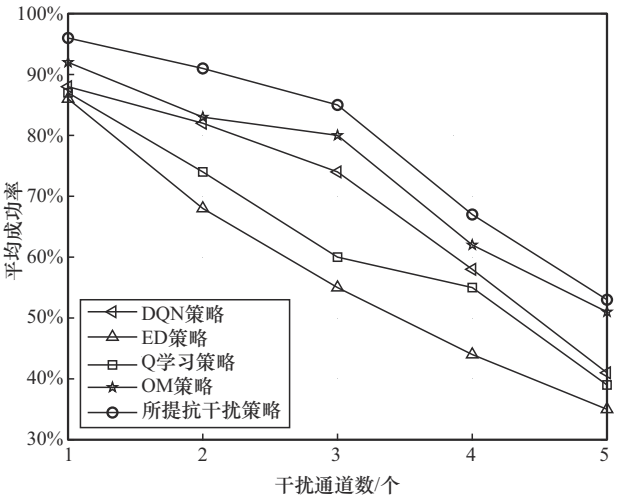


图10　所提抗干扰策略与4种对比策略的平均成功率随干扰通道数变化曲线

所提抗干扰策略与 3 种对比策略的归一化累积长期折扣效益随迭代次数变化曲线如图 11 所示。由图 11 可以看出，所提抗干扰策略具有最高的归一化累积长期折扣效益，相比 OM 策略、DQN 策略及 Q 学习策略，分别提升了近 5%、15% 和 30%；其次，相比其他 3 种抗干扰策略，所提抗干扰策略的归一化累积长期折扣效益随迭代次数变化曲线的波动幅度相对平稳。这是由于所提抗干扰策略中的熵预测网络对抗干扰策略网络的训练样本实现精细筛选，从而在抗干扰策略网络训练过程中，避免训练效率低的冗余样本对网络参数训练造成的波动。

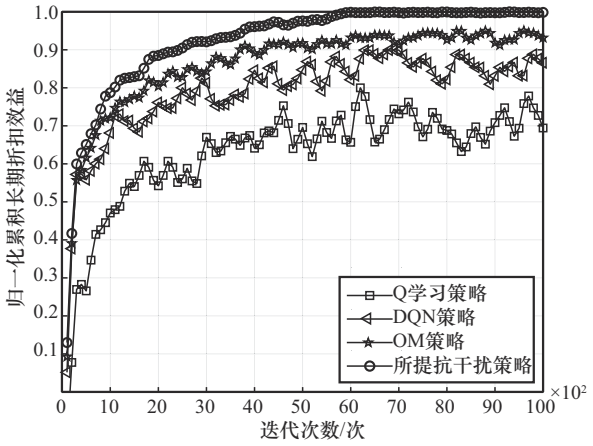


图11　所提抗干扰策略与3种对比策略的归一化累积长期折扣效益随迭代次数变化曲线

所提抗干扰策略在不同干扰策略更新频率下的平均成功率随迭代次数变化曲线如图12所示，分别设定干扰策略的更新频率为抗干扰策略更新频率的1倍、10倍、20倍、30倍、40倍、50倍及100倍。由图12可以看出，随着干扰策略更新频率的增加，所提抗干扰策略的平均成功率逐渐下降；当干扰策略更新频率和抗干扰策略更新频率相等时，所提抗干扰策略可实现94%的平均成功率；当干扰策略的更新频率为抗干扰策略更新频率的10倍时（由于人工智能技术的发展，通信方和干扰方对对方规律的学习水平都在提升，双方10倍更新频率差距是一种非常极端的设置），所提抗干扰策略实现了约85%的平均成功率；当干扰策略更新频率分别为抗干扰策略更新频率的20倍、30倍、40倍和50倍时，所提抗干扰策略的平均成功率分别约为75%、70%、61%和52%；当干扰策略的更新频率为抗干扰策略更新频率的100倍时，仍可达到30%的平均成功率。综上所述，所提抗干扰策略在干扰策略更新频率不超过抗干扰策略更新频率的40倍时，可取得60%的平均成功率，从而说明所提抗干扰策略的有效性。

以网络参数量和抗干扰决策时间作为指标，抗干扰策略的网络模型相关指标对比如表2所示，将基于样本信息熵辅助的抗干扰策略与2种基于深度强化学习的抗干扰策略，即DQN策略和OM策略的网络模型进行了对比。根据表2，DQN策略网络模型的网络参数量最小，同时其抗干扰决策时间也最短，这是因为DQN策略的网络模型结构相对简单；然而，其抗干扰的平均成功率却远不及OM策略和所提抗干扰策略（如图10所示）。此外，所提抗干扰策略的决策时间尽管比OM策略长1.22 ms，但其平均成功率不但比OM策略高5%，而且网络参数量少0.64 MB（所提抗干扰策略的网络模型相比OM策略少一个CL）；同时，所提抗干扰策略的网络模型训练的收敛速度和MSE损失明显优于OM策略。综上分析，所提抗干扰策略在抗干扰成功率、网络训练速度及MSE损失、网络参数量和决策时间之间取得了相对更优的折中。

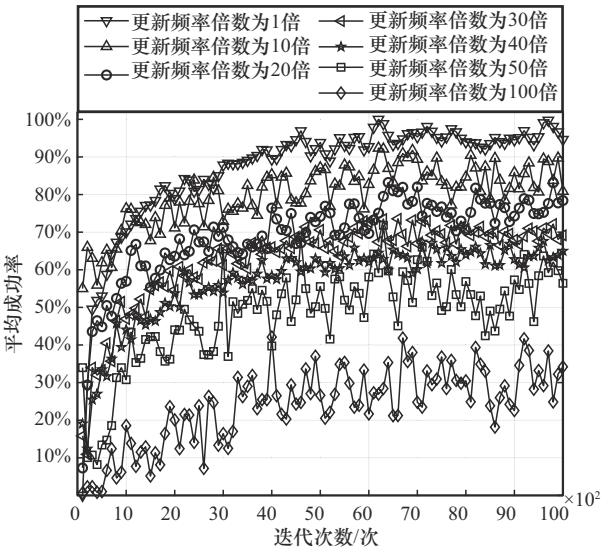


图12　所提抗干扰策略在不同干扰策略更新频率下的平均成功率随迭代次数变化曲线

表2　抗干扰策略的网络模型相关指标对比

抗干扰策略	网络参数量大小/MB	决策时间/ms
DQN策略	4.35	10.16
OM策略	18.13	15.12
所提抗干扰策略	17.49	16.34

4　结束语

针对深度强化学习驱动的智能干扰场景，本文提出了一种基于样本信息熵辅助的通信抗干扰策略。在该抗干扰策略中，首先，基于神经网络设计了抗干扰策略网络和熵预测网络；接着，为了增强抗干扰网络的决策能力，借助熵预测网络对训练样本进行精细化筛选，提高了训练数据的质量和相关性，从而改善了抗干扰策略网络的泛化性能。仿真结果表明，相较于其他几种抗干扰策略，所提抗干扰策略在抗干扰成功率、网络训练收敛速度及MSE损失、网络参数量和在线决策

时间等指标之间取得了相对更优的折中。同时，为了保证通信方和干扰方对抗过程的公平性，所提抗干扰策略中所设计的抗干扰策略网络与干扰策略网络采用相同的结构，但由于通信方和干扰方之间的对抗过程属于非合作马尔可夫博弈，通信方没必要也无法确知干扰策略网络的结构，因此，即使采用未知的不同干扰策略网络，所提抗干扰策略仍将适用和有效。此外，该研究主要针对端到端通信干扰场景，且考虑理想的电磁环境状态与干扰机状态完全吻合的理想状态；接下来笔者计划将相关研究拓展到多个通信用户通过协作机制来对抗智能干扰的场景，同时考虑实际电磁环境感知所确定状态与实际干扰机的状态存在偏差等特殊情形，对此展开深入研究。

参考文献：

[1] AMURU S, TEKIN C, VAN DER SCHAAR M, et al. Jamming bandits—a novel learning method for optimal jamming[J]. IEEE Transactions on Wireless Communications, 2016, 15(4): 2792-2808.

[2] PU Z M, NIU Y T, ZHANG G L. A multi-parameter intelligent communication anti-jamming method based on three-dimensional Q-learning [C]//Proceedings of the 2022 IEEE 2nd International Conference on Computer Communication and Artificial Intelligence (CCAI). Piscataway: IEEE Press, 2022: 205-210.

[3] ZHANG Z X, WU Q H, ZHANG B, et al. Intelligent anti-jamming relay communication system based on reinforcement learning[C]//Proceedings of the 2019 2nd International Conference on Communication Engineering and Technology (ICCET). Piscataway: IEEE Press, 2019: 52-56.

[4] 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.

[5] ZHANG X B, WANG H, RUAN L, et al. Joint channel, power and bandwidth optimization for Anti-jamming communications: a multi-agent Q-learning approach[C]//Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway: IEEE Press, 2021: 1-6.

[6] DING Y M, YANG F H, FENG J X, et al. Intelligent Anti-jamming algorithm based on time-frequency domain joint[C]//Proceedings of the 2021 6th International Symposium on Computer and Information Processing Technology (ISCIPT). Piscataway: IEEE Press, 2021: 163-167.

[7] 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.

[8] LI Y Y, XU Y H, XU Y T, et al. Dynamic spectrum anti-jamming in broadband communications: a hierarchical deep reinforcement learning approach[J]. IEEE Wireless Communications Letters, 2020, 9(10): 1616-1619.

[9] ZHANG L, MA L, TIAN F, et al. An anti-jamming intelligent decision-making method for multi-user communication based on deep reinforcement learning[C]//Proceedings of the 2022 IEEE 22nd International Conference on Communication Technology (ICCT). Piscataway: IEEE Press, 2022: 1335-1339.

[10] LI W, XU Y H, CHEN J, et al. Know thy enemy: an opponent modeling-based anti-intelligent jamming strategy beyond equilibrium solutions[J]. IEEE Wireless Communications Letters, 2023, 12(2): 217-221.

[11] SONG B L, XU H, JIANG L, et al. An intelligent decision-making method for anti-jamming communication based on deep reinforcement learning[J]. Journal of Northwestern Polytechnical University, 2021, 39 (3): 641-649.

[12] 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.

[13] NGUYEN P K H, NGUYEN V H, DO V L. A deep double-Q learning-based scheme for anti-jamming communications[C]//Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO). Piscataway: IEEE Press, 2021: 1566-1570.

[14] 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.

[15] HAN H, WANG X M, GU F L, et al. Better late than never: GAN-enhanced dynamic anti-jamming spectrum access with incomplete sensing information[J]. IEEE Wireless Communications Letters, 2021, 10(8): 1800-1804.

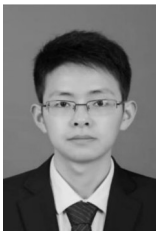
[16] CHEN M J, LIU W, ZHANG N, et al. GPDS: a multi-agent deep reinforcement learning game for anti-jamming secure computing in MEC network[J]. Expert Systems with Applications, 2022, 210: 118394.

[17] 冯智斌, 徐煜华, 杜智勇, 等. 对抗智能干扰的主动防御技术[J]. 通信学报, 2022, 43(10): 42-54.
FENG Z B, XU Y H, DU Z Y, et al. Active defense technology against intelligent jammer[J]. Journal on Communications, 2022, 43(10): 42-54.

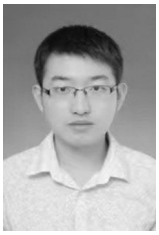
[18] HAN H, LI W, FENG Z B, et al. Proceed from known to unknown: jamming pattern recognition under open-set setting[J]. IEEE Wireless Communications Letters, 2022, 11(4): 693-697.

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

[作者简介]



李刚（1989– ），男，重庆人，中国西南电子技术研究所工程师，主要研究方向为分布式通信系统设计及人工智能等。



吴麒（1985– ），男，四川眉山人，博士，中国西南电子技术研究所高级工程师，主要研究方向为通信与系统总体设计、智能通信技术等。

王翔（1988– ），男，湖南常德人，博士，中国西南电子技术研究所工程师，主要研究方向为智能通信与网络技术等。

罗皓（1997– ），男，四川永川人，中国西南电子技术研究所工程师，主要研究方向为干扰对抗、信号处理等。

李良鸿（1992– ），男，四川巴中人，重庆邮电大学博士生，主要研究方向为通信抗干扰。

景小荣（1974– ），男，甘肃平凉人，博士，重庆邮电大学教授、博士生导师，主要研究方向为无线通信及通信对抗等。

陈前斌（1967– ），男，四川营山人，博士，重庆邮电大学教授、博士生导师，主要研究方向为无线通信、多媒体信息传输与处理。