一种基于深度强化学习的动态自适应干扰功率分配方法

彭 翔,许 华,蒋 磊,张 悦,饶 宁
(空军工程大学信息与导航学院,陕西西安 710077)

摘 要: 针对传统干扰功率分配方法在干扰目标策略未知的情况下容易造成资源浪费和干扰效费比低的问题,本文提出一种基于深度强化学习的动态自适应干扰功率分配方法. 在目标通信功率及功率控制策略完全未知的情况下,该方法将空间分布的侦察节点的观测值作为连续状态输入,利用深度强化学习方法进行干扰功率的辅助决策,可通过对目标策略的有效学习实现自适应稳定干扰. 为进一步提升算法性能,本文设计了基于时序误差的优先经验回放机制和自适应探索策略. 仿真结果表明,所提方法在与传统干扰功率分配方法干扰效果相当的情况下可节约42.5%的功率资源,提升了干扰效费比,且成功率和功率损耗皆优于对比的智能算法.

关键词: 电子对抗;通信对抗;干扰资源分配;干扰决策;功率分配;深度强化学习;优先经验回放
基金项目: 国家自然科学基金(No.61906156)
中图分类号: TN975 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1223-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220391

A Dynamic Adaptive Jamming Power Allocation Method Based on Deep Reinforcement Learning

PENG Xiang, XU Hua, JIANG Lei, ZHANG Yue, RAO Ning
(*Information and Navigation School, Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*)

Abstract: To solve the problem that traditional jamming power allocation methods are prone to waste resources and low jamming effectiveness-cost-ratio when the jamming target strategy is unknown, a dynamic adaptive jamming power allocation method based on deep reinforcement learning is proposed. When the communication power of the target and its power control strategy is completely unknown, the method takes the observation values of spatially distributed reconnaissance nodes as continuous state input and uses the deep reinforcement learning method to assist the decision-making of jamming power. It can achieve the adaptive stable jamming by the effective learning of target strategy. To further improve the performance of the algorithm, a prioritized experience replay mechanism based on temporal-difference error and an adaptive exploration strategy are designed. The simulation results show the proposed method can save 42.5% of power resources and improve the jamming effectiveness-cost-ratio when the jamming effect is equivalent to that of the traditional jamming power distribution method. The success rate and power cost of the proposed algorithm are better than those of the comparative intelligent algorithms.

Key words: electronic countermeasures; communication countermeasures; jamming resource allocation; jamming decision-making; power allocation; deep reinforcement learning; prioritized experience replay
Foundation Item(s): National Natural Science Foundation of China (No.61906156)

1 引言

随着无线电通信技术的快速发展,电磁频谱的争夺越来越激烈,资源短缺成为限制电子对抗双方整体效能发挥的主要因素之一,因此如何高效地进行资源分配是电子对抗及其相关领域的核心问题. 干扰资源分配是资源分配在电子对抗领域的实现形式之一,不同领域的资源分配方法很多都存在可相互借鉴之处,近些年关于认知通信的通信资源分配问题的研究,涉

收稿日期:2022-04-12;修回日期:2022-08-23;责任编辑:覃怀银

及时间[1]、空间[2]、频谱[3~5]和能量[5~7]等已经取得了丰富的成果,这为干扰资源分配的研究提供了参考. 当前电子对抗领域关于干扰资源分配问题的研究主要分为雷达干扰资源分配和通信对抗干扰资源分配两方面. 其中,雷达干扰资源分配方面的研究较多,文献[8,9]分别对遗传算法和蚁群算法进行改进,提高了组网雷达的干扰效率和干扰资源分配效率;文献[10]将强化学习应用到雷达干扰资源分配当中,实现了对雷达干扰策略分配的认知决策. 而通信对抗干扰资源分配方面的成果相对较少,现有研究主要集中在跳频频率[11]和干扰功率[12]的分配上. 因此,进一步开展通信对抗干扰资源分配问题的研究是必要的.

功率管理关系到未来通信对抗网络的可持续性,同时"精确电子对抗"[13]对干扰的隐秘性提出了更高的要求. 但是在现实对抗中,干扰方通常采用大功率压制干扰方式,这不仅造成资源的浪费,不利于通信对抗网络可持续性,同时也增大了己方暴露的概率,降低了己方战场生存率,多数情况下效费比(指干扰成功率与消耗功率的比值)不高. 为了提高干扰设备的战场生存能力和干扰效费比,利用低功率干扰信号在敌对目标甚至未曾察觉的情况下对其实施干扰是一个可行的方向,开展动态自适应干扰功率分配研究具有重要现实意义.

现有的大多数工作都是从静态优化的角度出发解决功率分配和控制问题[12,14~16]. 文献[14]证明了功率分配与信道选择问题属于NP-hard,传统优化算法难以适用. 为了克服传统优化方法在决策维度过大时失效或者陷入局部最优的问题,文献[15]提出一种基于分布式深度强化学习的方法,仅利用局部信息和过去的非局部信息来自动优化信道选择和传输功率,表现出较强可伸缩性. 为进一步提高决策效率,文献[16]采用分布式多智能体深度强化学习方法实现了资源的联合分配,最大限度地提高了频谱效率和能源效率. 与以上方法不同,本文从通信对抗场景下干扰功率的动态自适应调整角度展开研究,着眼于保证干扰成功率的同时最小化功率损耗,从而提高干扰效费比.

本文为提高通信对抗过程的干扰效费比,首先构建了"侦察-干扰"通信对抗模型;而后提出一种基于深度强化学习的动态自适应干扰功率分配(Dynamic Adaptive Jamming Power Allocation based on Deep Reinforcement Learning,DAJPA-DRL)方法,设计了基于时序误差的优先经验回放机制和自适应探索策略;最后通过仿真实验和算法对比验证了算法的有效性.

2 系统模型

本文构建如图1所示"侦察-干扰"模型,通信网络采用中心组网,每对发射机和接收机组成一条通信链路,组网中心统一向各链路下发通信策略(功率控制策略和频率分配策略),各链路按照接收到的功率控制策略P_x调整其通信功率P_T,依据频率分配策略选择通信频率f_T. 为了简化模型,假设各通信链路相互正交,本文仅研究某一通信链路上的自适应干扰功率分配问题,多通信链路的情况可类比扩展.

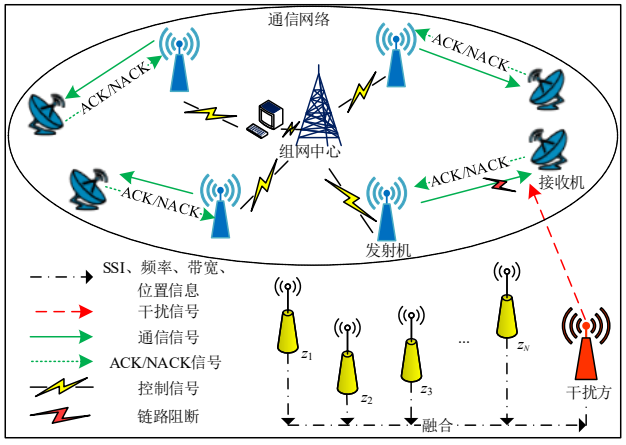


图1 "侦察-干扰"模型

以图1中通信网络中右下方发射机和接收机组成的通信链路为例,已知干扰方按实际需求事先布设N个侦察节点在发射机周围一定范围内,所有侦察节点共同组成侦察网络,$Z=\{z_1,z_2,z_3,\cdots,z_N\}$为侦察节点的集合. 侦察网络可以实时获取干扰方和发射机的信号强度信息(Signal Strength Information,SSI)并利用快速傅里叶变换等技术测量通信信号频率和带宽[17]. 同时,由于现代通信设备多是收发一体,具有全/半双工工作模式,例如通信接收机不仅接收信号,也会向发射机发送ACK/NACK信号进行应答,因此侦察网络可利用测向交叉定位和多普勒频率定位等辐射源定位技术获得目标位置信息[17]. 随后,侦察网络将获取到的SSI,通信信号频率、带宽信息和目标位置信息等进行融合并利用Zigbee[18]等传统技术实时反馈给干扰方(此过程不会影响干扰信号),干扰方进一步对信息进行综合处理后决策出干扰方案,然后对接收机实施干扰. 假设侦察网络可以准确测量和跟踪通信信号频率及带宽,保证干扰信号与通信信号的载频重合、带宽近似,实现频域对准;干扰方实施瞄准式干扰.

干扰方的目标是学会在时刻t根据收集到的SSI调整

其干扰功率,使用尽可能小的功率立刻或经过几个时隙的调整后成功干扰目标. 干扰功率为有限集P_{J}中的元素,频率为f_{J},波长为λ_{J},m为元素个数,即待选干扰功率等级.

$$\begin{aligned} &P_{\mathrm{J}} \triangleq \left\{P_{\mathrm{J}}^1, P_{\mathrm{J}}^2, \cdots, P_{\mathrm{J}}^m\right\} \\ &P_{\mathrm{J}}^1 \le P_{\mathrm{J}}^2 \le \cdots \le P_{\mathrm{J}}^m \end{aligned} \tag{1}$$

理想自由空间中信号的路径传播损耗L只与波长λ和传播距离r有关. 可以表示为

$$L = \left(\frac{4\pi r}{\lambda}\right)^2 \tag{2}$$

忽略极化损失和带宽失配损耗,第i条通信链路接收机处的信干噪比应当满足:

$$\mathrm{SINR}(i) = \frac{P_{\mathrm{T}i} G_{\mathrm{T}i} L_{ji}}{P_{ji} G_{ji} L_{\mathrm{T}i} + \Omega} \le K_i \tag{3}$$

其中,通信链路的通信功率为$P_{\mathrm{T}i}$;链路增益为$G_{\mathrm{T}i}$;干扰功率为P_{ji};干扰信号链路增益为G_{ji};$L_{\mathrm{T}i}$为发射机和接收机间的传播损耗;L_{ji}为干扰机和通信接收机间的传播损耗;Ω为环境噪声;K_i为信干噪比阈值. 由于非合作对抗条件下干扰方难以直接获取通信接收机处的信干噪比,因此无法通过式(3)直接评估干扰效果,但是综合功率准则,频率准则以及在特定通信协议下监听接收机向发射机发送的ACK/NACK信号推断出通信方的丢包率等方法可以间接判断干扰效果,因此本文假设干扰方可以得到干扰是否成功的反馈.

t时刻侦察节点z_n的观测值$P_{\mathrm{Z}}^n(t)$(即SSI)为

$$P_{\mathrm{Z}}^n(t) = \frac{P_{\mathrm{T}i}(t)}{L_{in}} + \frac{P_{\mathrm{J}}(t)}{L_{\mathrm{J}n}} + \delta_n(t) \tag{4}$$

其中,$P_{\mathrm{T}i}(t)$和$P_{\mathrm{J}}(t)$分别表示t时刻的通信功率和干扰功率;L_{in}表示第i条通信链路发射机和侦察节点z_n间的路径传播损耗;$L_{\mathrm{J}n}$表示干扰信号发射机和侦察节点z_n之间的路径传播损耗,L_{in}和$L_{\mathrm{J}n}$分别由式(2)计算得到;$\delta_n(t)$为由阴影效应和估计误差引起的观测噪声,服从方差为σ_n^2的零均值高斯分布,满足$\sigma_n^2 = \left[\left(P_{\mathrm{T}i}^1/L_{in} + P_{\mathrm{J}}^1/L_{\mathrm{J}n}\right)/\kappa + v\right]^2$;$\kappa$表示方差控制因子,$\kappa$越大,方差$\sigma_n^2$越小;$P_{\mathrm{T}i}^1$和$P_{\mathrm{J}}^1$分别为通信功率和干扰功率下界;$v$为一个极小正数,避免通信功率和干扰功率下界为零时出现σ_n^2等于零的情况.

鉴于通信方与干扰方之间的非合作关系,已知通信方以独立的策略进行功率调整,本文给出由两种功率控制策略组成的策略集$P_{\mathrm{x}} = \left\{P_{\mathrm{x}1}, P_{\mathrm{x}2}\right\}$用于算法检验[19],$P_{\mathrm{x}1}$、$P_{\mathrm{x}2}$分别为策略1和策略2,定义如下:

策略1

$$\begin{aligned} &P_{\mathrm{T}i}(t+1) = f\left(\frac{K_i \cdot P_{\mathrm{T}i}(t)}{\mathrm{SINR}(i)}\right) \\ &P_{\mathrm{T}} \triangleq \left\{P_{\mathrm{T}i}^1, P_{\mathrm{T}i}^2, \cdots, P_{\mathrm{T}i}^l\right\}, P_{\mathrm{T}i}^1 \le P_{\mathrm{T}i}^2 \le \cdots \le P_{\mathrm{T}i}^l \end{aligned} \tag{5}$$

其中,$\mathrm{SINR}(i)$为t时刻第i条通信链路接收机处的信干噪比;$P_{\mathrm{T}i}(t)$表示t时刻第i条通信链路的通信功率,通信功率以时隙为单位动态调整;K_i为信干噪比阈值;l表示集合元素个数,即待选通信功率等级;$f(x) = \underline{x}$是一个离散化函数,它将连续值映射到由离散值组成的通信功率集P_{T}中,$\underline{x}$表示最接近x但不超过x的离散值,如果$x > P_{\mathrm{T}i}^l$则令$f(x) = P_{\mathrm{T}i}^l$.

策略2

$$P_{\mathrm{T}i}(t+1) = \begin{cases} P_{\mathrm{T}i}^{j+1}, P_{\mathrm{T}i}^j \le \eta \le P_{\mathrm{T}i}^{j+1}, j+1 \le l \\ P_{\mathrm{T}i}^{j-1}, \eta \le P_{\mathrm{T}i}^{j-1}, j-1 \ge l \\ P_{\mathrm{T}i}^j \ , \text{otherwise} \\ \eta \triangleq \dfrac{K_i \cdot P_{\mathrm{T}i}(t)}{\mathrm{SINR}(i)} \end{cases} \tag{6}$$

假设t时刻通信功率$P_{\mathrm{T}i}(t) = P_{\mathrm{T}i}^j, P_{\mathrm{T}i}^j \in P_{\mathrm{T}}$. 不难看出,策略2采取逐步更新的规则,即$t$+1时刻的功率只能在相邻的两个功率等级上变动或者维持原功率,相比策略1更为保守. 通信方根据链路被干扰的程度按照组网中心下发的策略调整其通信功率,即当前时刻干扰方的行动以一种隐式的方式影响着通信方下一步的动作.

3　基于深度强化学习的动态自适应干扰功率分配方法

强化学习(Reinforcement Learning,RL)方法可以在没有先验知识的条件下,通过"试错"的方式与环境进行交互,训练出具有突出决策能力的智能体,被广泛用于智能决策领域. 传统的强化学习方法(例如Q-Learing、SARSA)适用于离散的低维动作、状态空间,在解决具有连续状态空间的问题时,其表格式存储价值的方法不再适用. 深度强化学习(Deep Reinforcement Learning,DRL)方法在RL的基础上引入神经网络,通过网络拟合状态动作价值函数,克服了传统强化学习方法的高维难题,适用于解决连续状态输入的实际问题,实现了从感知到行动的端到端的学习,深度Q-Learning算法(Deep Q-Network,DQN)[20]便是典型代表. 特别的,在本文所构建的问题模型当中,干扰方对目标链路的通信功率及功率控制策略完全未知,且侦察节点信号强度测量中的随机误差使得状态具有连续性,因此本文基于深度强化学习方法设计自适应干扰功率分配

方法.

3.1 动态干扰功率分配问题的马尔可夫决策模型

马尔可夫决策过程是单智能体强化学习方法的基础理论,用于在系统状态具有马尔可夫性质的环境中模拟智能体的随机性策略与回报,包括决策过程中的状态、动作、策略和奖赏等因素. 分析本文模型可知:在任意时刻,干扰方通过侦察节点获取环境状态信息,决策出干扰动作后实施干扰并通过情报获知干扰是否成功的反馈,通信方按照己方策略调整通信功率,环境状态发生改变. 对干扰方而言,新的状态仅取决于当前状态和干扰动作,与过去所有的状态无关,当前时刻的干扰动作以一种隐式的方式影响通信方的下一步行动,下一个状态有条件的独立于过去所有的状态和动作,因此干扰方与通信方组成的对抗系统状态具有马尔可夫性质,若干扰方作为智能体则动态干扰功率分配问题可建模为马尔可夫决策过程.

基于此,本文将干扰功率分配问题构建成由四元组$<\boldsymbol{S},\boldsymbol{A},R,\gamma>$描述的马尔可夫决策过程. 其中,$\boldsymbol{S}$为状态空间;$\boldsymbol{A}$为动作空间;$R$为当前状态下采取动作获得的即时奖励;$\gamma\in[0,1]$表示折扣因子,用来表示未来收益对当前状态的影响程度,γ越大表示越注重未来收益. 具体物理含义如下:

(1) 状态空间$\boldsymbol{S}$:$\boldsymbol{S}$为N维张量,由N个侦察节点的观测值组成. 即

$$\boldsymbol{S}(t)\triangleq\left[P_{\mathrm{Z}}^{1}(t),P_{\mathrm{Z}}^{2}(t),\cdots,P_{\mathrm{Z}}^{N}(t)\right]^{\mathrm{T}} \tag{7}$$

(2) 动作空间$\boldsymbol{A}$:$\boldsymbol{A}$为m维张量,由t时刻干扰功率集P_{J}内各功率状态决定. 即

$$\begin{aligned}&\boldsymbol{A}(t)\triangleq\left[P_{\mathrm{J}}^{1}\cdot E^{1},\cdots,P_{\mathrm{J}}^{m}\cdot E^{m}\right]^{\mathrm{T}}\\&E^{i}\in\{0,1\},i=1,2,\cdots,m\end{aligned} \tag{8}$$

(3) 环境奖励R:R设置的合理性很大程度上决定了整个模型的可行性. 结合模型特性,本文定义的奖励函数由成功干扰奖励R_1和优化功率分配奖励R_2两部分组成. 定义R_1为

$$R_1=\begin{cases}c,\mathrm{SINR}(i)\le K_i\\c',\text{otherwise}\end{cases} \tag{9}$$

式(9)中,c为正常数,表示成功干扰通信链路后获得的正收益;c'为负常数,表示干扰失败后获得的负奖励,负的奖励将使决策网络在更新网络参数时获得更大的梯度更新值,促进决策网络的优化. c和c'的取值可进行合理调整,需满足$|c|\gg|c'|$. 本文中c为10,c'为−1. 定义R_2为

$$R_2=-\mathrm{index}(P_{\mathrm{J}}(t)) \tag{10}$$

式(10)中,$\mathrm{index}\left(P_{\mathrm{J}}(t)\right)$为$P_{\mathrm{J}}(t)$在干扰功率集$P_{\mathrm{J}}$中的索引,$R_2$用于鼓励干扰方尽可能使用小的干扰功率成功干扰目标,从而优化功率分配.

因此,总的奖励函数为

$$R=R_1+R_2=\begin{cases}c-\mathrm{index}(P_{\mathrm{J}}(t)),\mathrm{SINR}(i)\le K_i\\c'-\mathrm{index}(P_{\mathrm{J}}(t)),\text{otherwise}\end{cases} \tag{11}$$

在强化学习架构下,上述马尔可夫决策过程中干扰方与通信方的交互过程如图2. 干扰方为智能体Agent,通信链路构成Environment,干扰方通过侦察网络获得当前环境状态State,实施干扰动作Action,并根据干扰成功与否获得奖励值Reward,下一时刻通信方依据功率控制策略调整信号发射功率,环境状态发生改变. 通过不断地对抗交互,干扰方最终学到通信方的功率控制策略,可以得到不同环境状态下的最佳干扰功率分配方案,从而实现累积奖励期望的最大化,即干扰方具备自适应干扰功率分配能力.

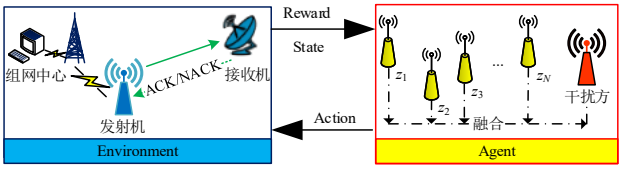


图2 干扰方与通信方的交互过程

3.2 算法描述

实际功率分配问题中,动作空间往往由有限、离散的元素构成,而深度强化学习当中的DQN算法在解决连续状态,离散动作的问题上表现突出,因而得到广泛应用. DQN算法是一种基于值的强化学习方法,而基于值的强化学习方法通过当前价值函数不断拟合目标价值函数实现参数更新,即用价值网络本身做出的估计去更新价值网络本身,因此导致了“自举”的产生. 为了克服DQN算法[20]因“自举”导致偏差传播从而引起过估计的问题,文献[21]在原算法基础上引入目标网络计算目标值,降低了“自举”造成的过估计. 本文在文献[21]的基础上设计了基于时序误差的优先经验回放机制进一步提高样本利用率;同时引入适应性探索策略,既确保了算法训练前期的探索性,又确保算法后期对已有知识的充分利用,提高了算法训练效率,算法框架如图3.

其中,评估网络用于计算当前状态的动作价值,目标网络用于计算下一状态的动作价值. 经验回放池以二叉树的形式存放经验五元组$<\boldsymbol{S},\boldsymbol{A},R,\boldsymbol{S}',p>$,$p$为经验优先级,当新经验存入或批采样数据训练网络时进行

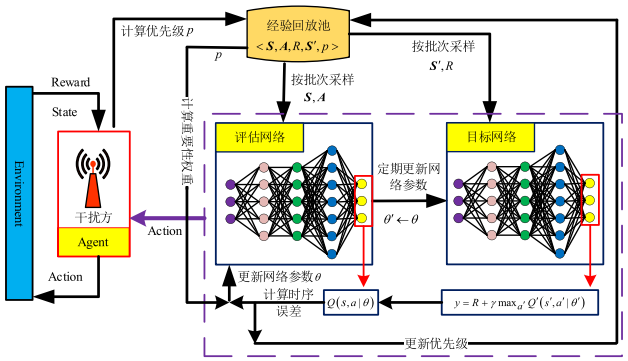


图3　算法框架

优先级更新,优先级高的采样概率大.

基于值的强化学习算法的基本思想是根据最优贝尔曼方程迭代更新动作价值函数,如式(12):

$$Q_*(s,a)=\left(R(s,a)+\gamma\cdot \mathrm{E}_{s'\sim\psi(\cdot|s,a)}\left[\max_{a'}Q_*(s',a')\right]\right) \tag{12}$$

其中,$Q_*(s,a)$为最优动作价值函数;$R(s,a)$为状态s下采取动作a后即时奖励;s'为当前状态s下采取动作a后的下一个状态,服从概率分布$\psi(\cdot|s,a)$;a'为状态s'下的最优动作. 通过充分的训练交互,可以得到任意时刻不同状态下选择最佳动作的规则,即动作价值函数收敛到最优策略ϕ^*,实现回报G_t期望的最大化.

$$\begin{aligned}\phi^*&=\arg\max_{\phi}\mathrm{E}(G_t)=\arg\max_{\phi}\mathrm{E}(R_{t+1}+\gamma R_{t+2}+\cdots)\\&=\arg\max_{\phi}\mathrm{E}\left(\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\right)\end{aligned} \tag{13}$$

经验回放池的设计打破了历史经验间的时间相关性,提高了样本利用率,加速了算法收敛. 传统的经验回放技术采用均匀采样的方法,对所有的经验数据予以相同的重视程度,这与实际中历史经验的非等价性相违背,因此传统经验回放技术重要经验的利用率较低,仍有较大提升空间. 为进一步提高算法效率,区别不同经验的重要程度,本文为每一个经验设置优先级p(通过δ的幅值来衡量[22]),同时赋予优先级高的经验更高的采样概率ρ,计算式如下:

$$\begin{cases}\rho(i)=p_i^{\alpha}\Big/\sum_{k,k\in\{1,2,\cdots,K\}}p_k^{\alpha}+\zeta\\p_i=|\delta_i|+\xi,i\in\{1,2,\cdots,K\}\\|\delta_i|=\left|R_i+\gamma\max_{a'}Q'(s',a'|\theta')-Q(s,a|\theta)\right|\end{cases} \tag{14}$$

其中,K为经验回放池容量;指数α为优先级重视程度,$\alpha=0$时等价于均匀采样;ζ为概率偏差,用于确保可以选择优先级极低的样本;δ_i即索引为i的经验的时序误差;ξ为优先级偏差,是一个很小的正常数,用于防止经验误差为零时不重新访问经验的极端情况;p_i为索引为i的经验的优先级;$\rho(i)$为索引为i的经验的采样概率,显然采样概率是关于优先级的单调函数.

另一方面,基于优先级的经验回放以一种不受控的方式改变原始经验的分布,造成了计算偏差[22]. 而统计学当中的重要性采样可以从与原先分布不同的其它分布中采样,实现对原先分布性质的估计. 因此本文引入重要性抽样权重来纠正这种偏差,定义式如下:

$$\omega_i=\left(\frac{\rho(i)\cdot\rho_{\mathrm{total}}}{\rho_{\min}}\right)^{-\beta} \tag{15}$$

ρ_{total}为优先级总和,$\rho_{\min}$为最小优先级,ω_i为索引为i的经验的重要性权重,β为修正系数.

平衡对“环境的探索”和“已有经验的利用”二者间的关系一直是强化学习方法解决问题的关键,对避免算法陷入局部最优和提高收敛速度至关重要. ε贪婪策略是最常用的探索策略,但是ε贪婪策略在训练过程中始终保持ε为固定值,不利于算法后期的收敛,因此本文在ε贪婪策略的基础上设计了自适应探索策略,如式(16):

$$\begin{cases}1-\varepsilon,a=\arg\max_{a}Q(s,a)\\\varepsilon,\text{random select }a\\\varepsilon=\text{Initial\_}\varepsilon-\mu\times\Delta\\\Delta=(\text{Initial\_}\varepsilon-\text{Final\_}\varepsilon)/T'\\1\geq\text{Initial\_}\varepsilon\geq\text{Final\_}\varepsilon\geq 0\end{cases} \tag{16}$$

其中,Initial_ε和Final_ε分别为ε的初始值和终止值;T'为总迭代次数,即总交互时间为T'个时隙;μ为当前迭代次数;Δ为ε的变化步长. 可见ε会随着迭代次数的增加而逐渐减小,既保证了训练初期的充分探索,又确保后期对经验的充分利用,加速算法收敛.

训练过程网络更新规则如下:

步骤1　计算目标函数

$$y_i=R_i+\gamma\max_{a'}Q'(s',a'|\theta') \tag{17}$$

步骤2　计算损失函数

$$\begin{aligned}L(\theta)&=\frac{1}{B}\sum_i\omega_i\left(y_i-Q(s,a|\theta)\right)^2\\&=\frac{1}{B}\sum_i\omega_i\begin{pmatrix}R_i+\gamma\max_{a'}Q'(s',a'|\theta')\\-Q(s,a|\theta)\end{pmatrix}^2\end{aligned} \tag{18}$$

其中,$\max_{a'}Q'(s',a'|\theta')$为目标网络在状态$s'$时对应的最大状态动作价值;$\theta$和$\theta'$分别为评估网络和目标网络的网络参数;$R_i$为当前状态$s$采取动作$a$的即时回报;$B$为批处理大小.

步骤 3 使用梯度下降法最小化损失函数更新网络参数 θ，$\theta \leftarrow \theta - \alpha_\theta \cdot \nabla L(\theta)$，$\alpha_\theta$ 为学习率，∇ 为梯度算子. 目标网络参数由评估网络参数定期更新，$\theta' \leftarrow \theta$.

基于 DRL 的动态自适应干扰功率分配方法如算法 1 所示.

算法 1 基于 DRL 的动态自适应干扰功率分配方法

Step1:随机初始化评估网络参数 θ，初始化目标网络参数 $\theta' \leftarrow \theta$;
Step2:初始化经验回放池容量为 K，批处理大小为 B，训练总回合为 T';
Step3:初始化参数 $\alpha, \gamma, \zeta, \xi, \beta, \alpha_\theta$, Initial_$\varepsilon$, Final_$\varepsilon$, O, C;
Step4:初始化初始状态 $s(1)$;
Step5:FOR $t=1,\cdots,T'$ do:
(1) 根据自适应探索策略选择动作 $a(t)$;
(2) 执行动作 $a(t)$，获得奖励 $R(t)$，通信方使用策略 1 或策略 2 更新 $P_{\mathrm{T}i}(t+1)$，得到新状态 $s(t+1)$;
(3) 存储 $(s(t), a(t), R(t), s(t+1))$ 到回放池中，最大化其优先级为 $p(t)=\max_{j<t} p(j), j \in \{1,2,\cdots,K\}$;
(4) IF $t \geq O$ THEN:
FOR $i=1$ to B do:
(a) 经验回放池中采样样本
$\rho(i)=p_i^\alpha \Big/ \sum_{k,k\in\{1,2,\cdots,K\}} p_k^\alpha + \zeta$;
(b) 计算重要性抽样权重 $\omega_i=\left(\rho(i)\cdot\rho_{\mathrm{total}}/\rho_{\min}\right)^{-\beta}$;
(c) 计算时序误差
$|\delta_i| = \left| R_i + \gamma \max_{a'} Q'(s', a'|\theta') - Q(s, a|\theta) \right|$;
(d) 更新经验优先级 $p_i=|\delta_i|+\zeta$;
END FOR
梯度下降法最小化损失函数更新网络参数
$\theta \leftarrow \theta - \alpha_\theta \cdot \nabla L(\theta)$
END IF
(5) 每 C 步更新目标网络参数，$\theta' \leftarrow \theta$;
(6) IF $s(t)$ 是目标状态 THEN:
重新初始化初始状态 $s(t+1)$;
END IF
END FOR

4 实验与仿真

为验证所提方法的性能，本文进行了充分的实验. 首先，利用控制变量法进行对比实验，确定一组最佳参数；其次，基于最佳参数分别分析侦察节点数目 N 和观测噪声方差 σ_n^2 对算法效果的影响以及不同目标功率控制策略下算法的适用性；最后，为了进一步评估 DAJPA-DRL 算法的性能，将其与传统干扰功率分配方法[23]和基于 DQN 的功率控制算法[19]进行对比. 由于对抗中通信方和干扰方以时隙为单位进行功率调整，双方没有严格的时间同步，因此本文引入时隙转移限度 χ 来贴合现实，如果干扰方在 χ 内成功干扰目标链路，则认为单次试验成功. 模型训练完成后，利用蒙特卡罗法进行 500 回合测试实验，单回合 100 次取平均值，以每回合的测试平均成功率和平均功率作为综合衡量算法性能好坏的关键指标.

4.1 模型及网络参数设置

干扰功率集 $P_{\mathrm{J}} \triangleq \{0.05, 0.1, 0.15, \cdots, 0.8\}$，通信功率集 $P_{\mathrm{T}} \triangleq \{0.05, 0.1, 0.15, \cdots, 0.4\}$（单位：kW）. 假设在通信方 100~300 km 的范围内布设有 N 个侦察节点. 二维坐标系下，干扰方坐标为（0，0），通信接收机坐标为（0，300）（单位：km），如图 4.

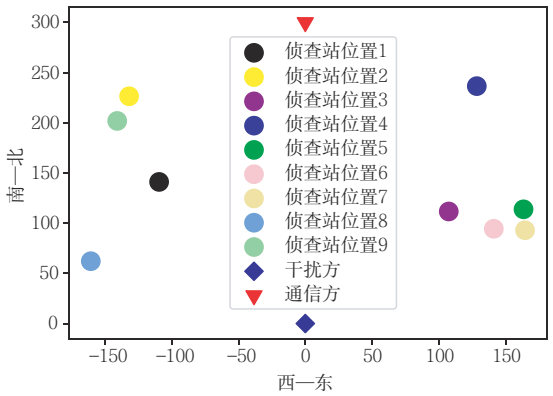


图 4 N=9 时的对抗空间分布

其他模型参数如表 1.

本文算法中评估网络和目标网络均使用全连接神经网络，具体参数设置如表 2.

表 1 模型参数

物理参数	值	物理参数	值
环境噪声 Ω/kW	0.01	α	1
干扰/通信链路增益 G	1	ζ	10^{-5}
发射机和接收机间的传播损耗 $L_{\mathrm{T}i}$	0.8	v	10^{-5}
干扰机和通信接收机间的传播损耗 L_{ji}	0.8	ξ	0.01
干扰/通信中心频率 f/MHz	300	β	0.9
方差控制因子 κ	1 或 5	Initial_ε	0.8
信干噪比阈值 K_i	0.5	Final_ε	0
经验回放池容量 K	1 024	T'	2×10^4
时隙转移限度 χ	5	O	256
学习率 α_θ	10^{-5}	C	300

表 2 DAJPA-DRL 算法网络参数

	评估网络	目标网络
输入层	N	N
隐藏层 1	(256,Relu)	(256,Relu)
隐藏层 2	(256,Relu)	(256,Relu)
隐藏层 3	(512,Tanh)	(512,Tanh)
输出层	16	16
优化器	Adam	

4.2　网络参数优化

首先,利用控制变量法进行对比实验,确定批处理大小B以及折扣因子γ. 批处理大小B很大程度上影响着算法性能,B越大单次采样越多,收敛时间越短,计算越复杂,模型泛化能力越小,反之,B越小算法收敛越慢,可能导致准确率来回震荡,但是具有更好的泛化能力[24]. 本文在N=3、κ=5、γ=0.8以及目标采取策略2的条件下,通过控制变量确定批处理大小B.

由图5和图6知,B为32时网络收敛较慢,但是测试成功率最优;图7显示其平均功率最高,这是干扰成功率提高的结果. 因此,本文后续实验选定B为32.

γ的大小决定了算法的"价值观",合适的γ同样对算法的收敛至关重要. 本文在N=3、κ=5、B=32以及目标采用策略2的前提下,采用控制变量法确定γ的大小,其中γ分别取0.7,0.8,0.9,0.99.

由图8~10可知,γ=0.9时算法收敛最快,测试平均奖励值最高,在测试成功率与γ=0.99相当的情况下,平均功率更低,因此,本文后续实验选定γ为0.9.

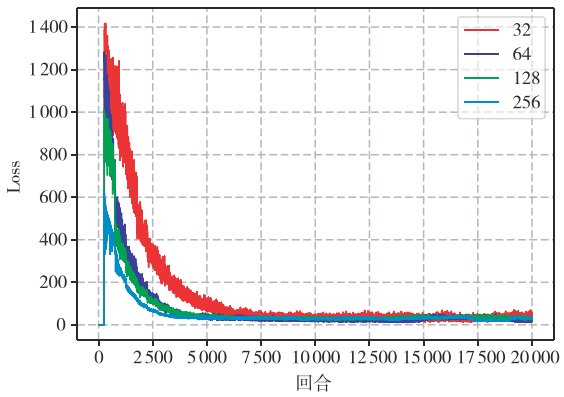


图5　目标策略2下不同B的训练损失曲线

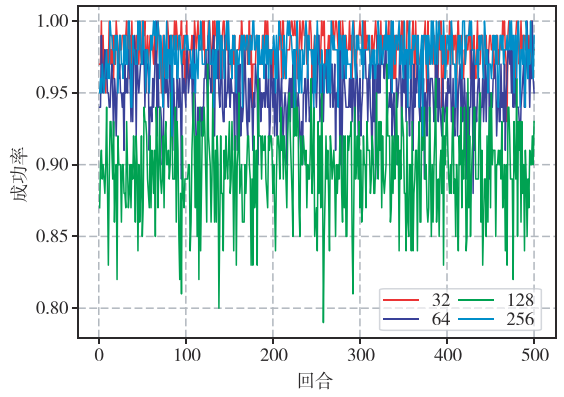


图6　目标策略2下不同B的测试成功率

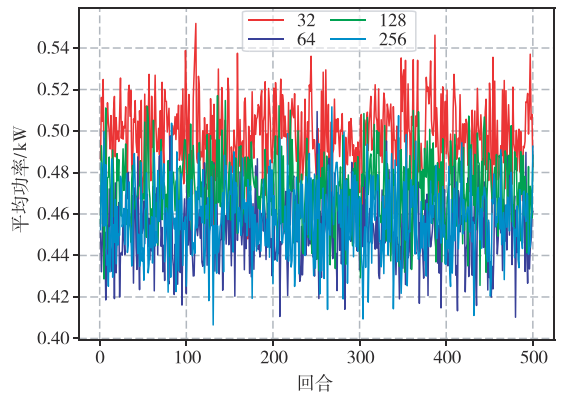


图7　目标策略2下不同B的测试平均功率

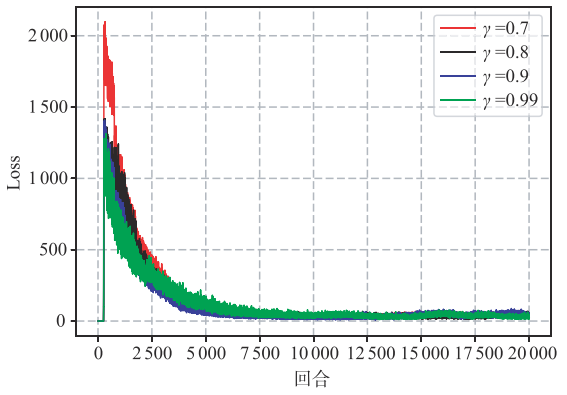


图8　目标策略2下不同γ的训练损失曲线

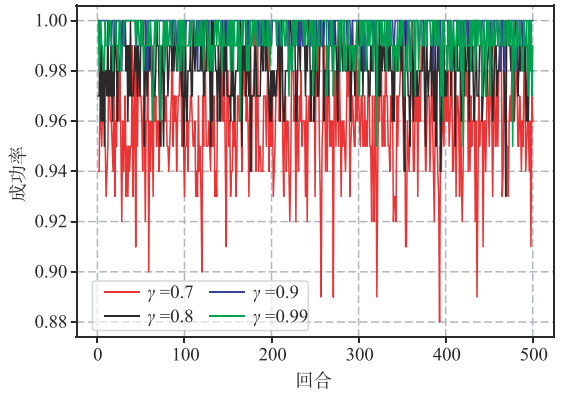


图9　目标策略2下不同γ的测试成功率

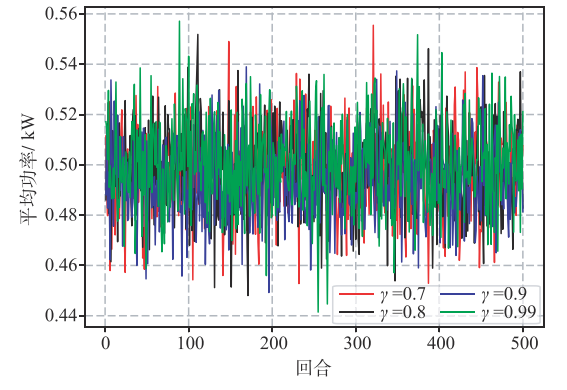


图10　目标策略2下不同γ的测试平均功率

4.3 实验结果及分析

4.3.1 侦察节点数目N和观测噪声方差σ_n^2对算法效果的影响

首先基于上文参数分别分析侦察节点数目N和观测噪声方差σ_n^2对算法效果的影响，这里引入平均时隙转移步数作为衡量算法时间效率的标准. 本文在κ=1及目标选择策略1的条件下进行实验，分析N分别为3、6、9时算法效果.

结果如图11~13，可见N=9时算法的收敛速度最慢但是干扰成功率最高，成功干扰目标所需的时隙最少. 分析可知N越大，干扰方获取的环境信息越充分，需要处理的信息越多，导致算法收敛越慢，但是训练完成后，干扰方对环境状态更为了解，因此测试的干扰成功率越高，成功干扰所需时隙越少，更能满足实时性需求，这与实际情况相符，在现实对抗中可以通过布设更多的侦察节点来获得更可靠的环境态势.

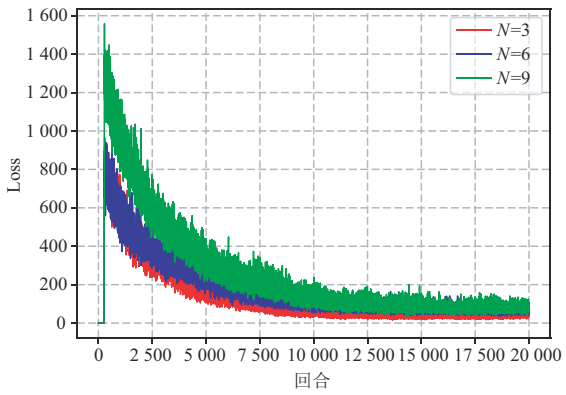


图11 目标策略1下不同N的训练损失曲线

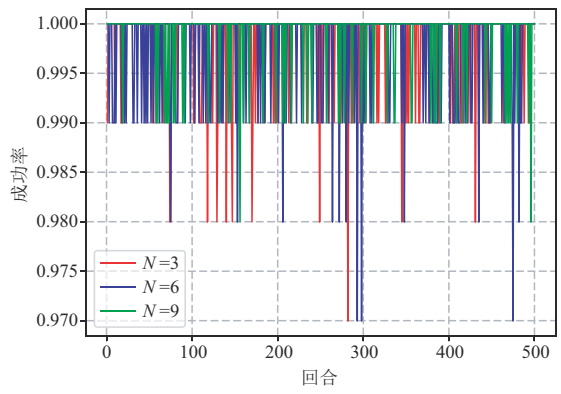


图12 目标策略1下不同N的测试成功率

接下来在N=3，目标选择策略1的条件下分析观测噪声方差σ_n^2对算法效果的影响. 由式(4)可知σ_n^2的大小取决于κ，本文κ分别取值1或5进行对比实验.

图14表明，κ越大，σ^2越小，算法收敛越快，但是训练过程中观测值的随机性越低，导致算法的泛化性下降，体现在图15中κ=5时算法测试成功率相对较低，图16中κ=5时测试平均时隙转移步数波动较大.

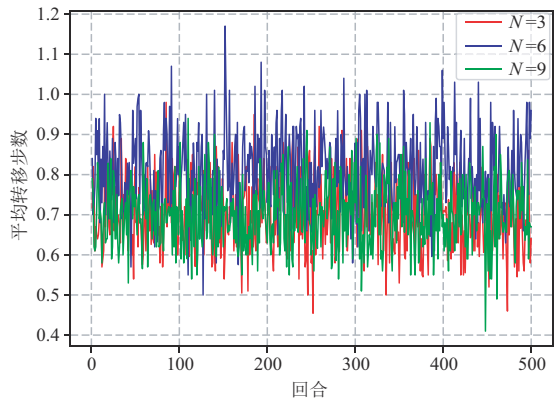


图13 目标策略1下不同N的平均转移步数

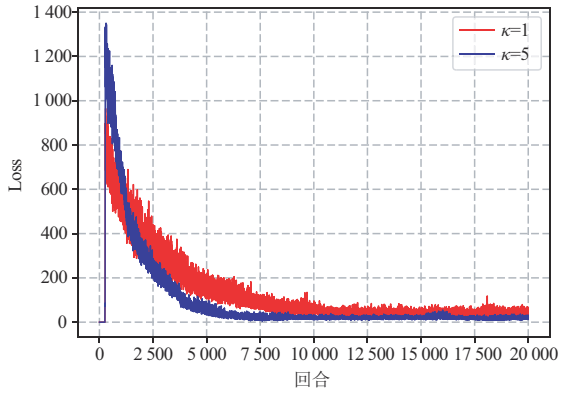


图14 目标策略1下不同κ的训练损失曲线

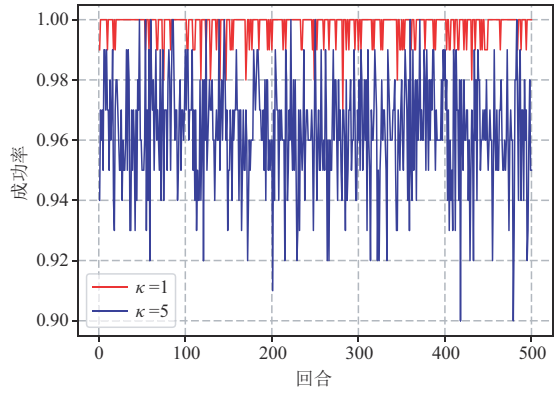


图15 目标策略1下不同κ的测试成功率

综合分析以上两组实验，可以发现观测噪声方差或侦察节点数目发生变化时，所提算法学习仍然有效，保持着相近的干扰成功率和平均时隙转移步数，κ为1时可以保证97%以上的成功率，在κ=5，N=3的情况下也能保证90%以上的成功率，体现了所提算法的稳健性.

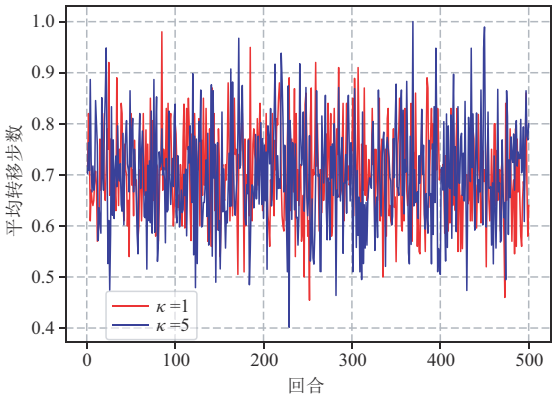
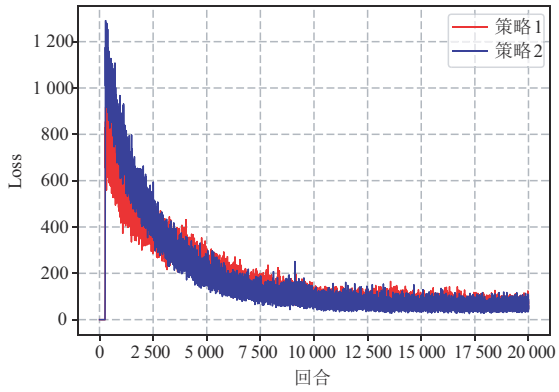


图 16　目标策略 1 下不同 κ 的平均转移步数

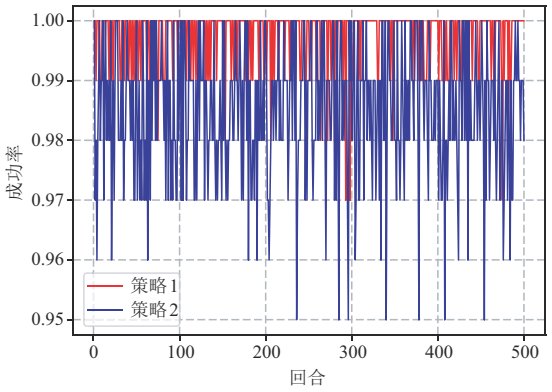
4.3.2　不同目标功率控制策略下算法效果分析

接下来分析算法在目标策略 1 和策略 2 下的性能，验证算法的适用性，取 N=6，κ 分别为 1 和 5.

分析图 17(a)和图 18(a)可知，由于策略 2 较为保守，因此算法需要更长时间学习，相比策略 1 算法在策略 2 下收敛更慢. 由图 17(b)可知，κ=1 时算法在两种目标策略下测试成功率均能保证在 95% 以上；由图 18(b)可知，在 κ=5 时算法在目标策略 1 下的测试成功率在 94% 以上，即使目标选择策略 2 时也能维持成功率在 88% 上下，最低为 74%. 由此可知，所提算法在不同目标策略下具备较强适用性，算法泛化性能较强，可靠性较高.

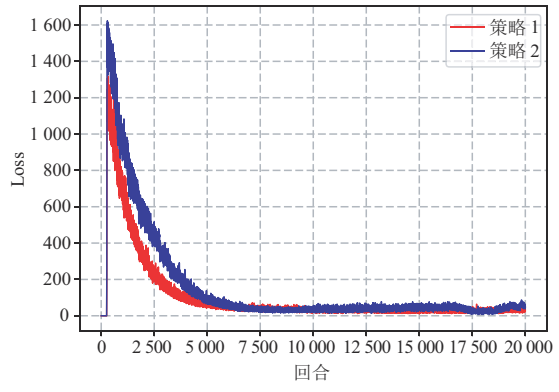


(a) 不同目标策略下的训练损失曲线

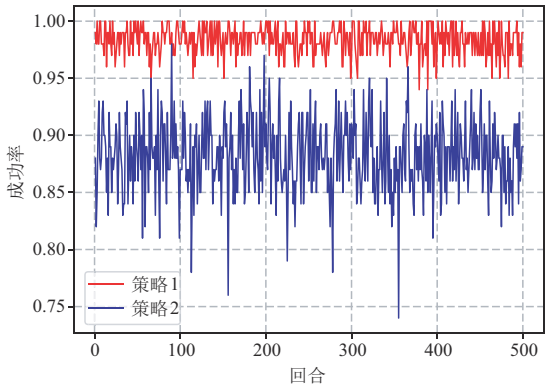


(b) 不同目标策略下的测试成功率

图 17　N=6,κ=1 时算法效果



(a) 不同目标策略下的训练损失曲线



(b) 不同目标策略下的测试成功率

图 18　N=6,κ=5 时算法效果

4.3.3　不同算法实验效果分析

最后将本文算法与传统干扰功率分配方法[23]和基于 DQN 的功率控制算法[19]进行对比. 为保证公平性，设置相同的模型参数及网络结构，取 N=6，κ=1，目标选择策略 1.

本文中传统干扰功率分配方法持续使用最大功率对目标进行干扰，因此平均成功率为 1. 图 19 表明两种智能算法网络收敛速度相近，但 DAJPA-DRL 算法前期的训练损失更大，这是因为基于时序误差的优先经验回放机制使得时序误差大的经验获得了更高的采样概率，进一步提高了重要样本的利用率. 但是优先经验回放机制的引入并不会因算法复杂度上

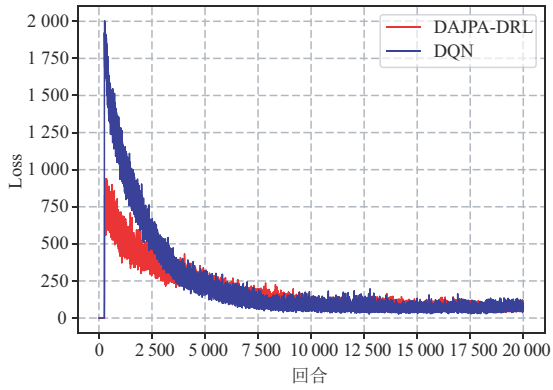


图 19 目标策略 1 下训练损失曲线对比

升而使得网络难以收敛，反而可以提升算法效果，由图 20 和图 21 可知，DAJPA-DRL 算法的干扰成功率明显优于基于 DQN 的功率控制算法，而且消耗了更少的功率资源．表 3 给出了三种方法的性能对比结果，可以明显看出使用深度强化学习方法可以大幅节约功率资源，而且本文算法在干扰成功率和功率消耗方面皆优于基于 DQN 的功率控制算法，DAJPA-DRL 算法在与传统干扰功率分配方法干扰成功率相当的情况下可节约 42.5% 的功率资源，干扰效费比有明显的提升．

表 3 性能对比

功率分配方法	平均消耗功率/kW	平均成功率	节约功率占比	效费比(干扰成功率/消耗功率)
传统干扰功率分配方法[23]	0.8	1	0	1.25
基于 DQN 的功率控制算法[19]	0.48	98.3%	40%	2.05
DAJPA-DRL 算法	0.46	99.7%	42.5%	2.17

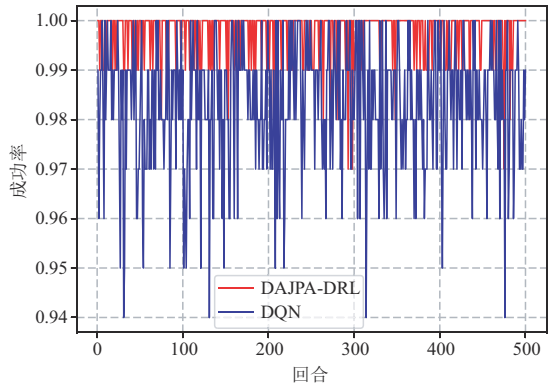


图 20 目标策略 1 下测试成功率对比

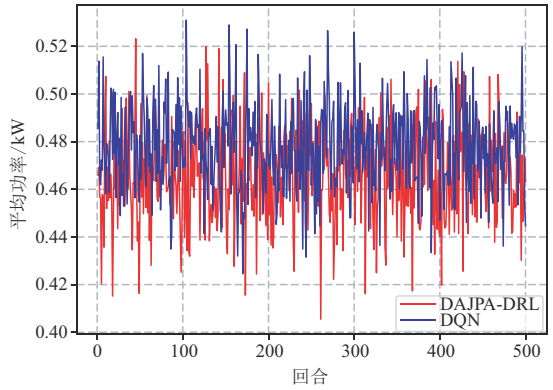


图 21 目标策略 1 下测试平均功率对比

5 结论

本文提出一种基于深度强化学习的动态自适应干扰功率分配方法，解决了传统干扰功率分配方法干扰效费比低的问题．首先将动态干扰功率分配问题建模为马尔可夫决策过程，为方法的提出搭建了基础环境；其次通过设计基于时序误差的优先经验回放机制和自适应探索策略提升了方法性能．仿真结果表明该方法可通过对目标策略的学习，实现干扰功率的自适应分配，在干扰成功率与传统干扰功率分配方法相当的情况下减少了 42.5% 的功率消耗，有效提高了干扰效费比．同时，所提算法在解决动态功率资源分配问题时，成功率和功率损耗皆优于基于 DQN 的功率控制算法．

参考文献

[1] XIONG X, ZHENG K, LEI L, et al. Resource allocation based on deep reinforcement learning in IoT edge computing[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(6): 1133-1146.

[2] SHI W S, LI J L, WU H Q, et al. Drone-cell trajectory planning and resource allocation for highly mobile networks: A hierarchical DRL approach[J]. IEEE Internet of Things Journal, 2021, 8(12): 9800-9813.

[3] ZHAO B K, LIU J H, WEI Z L, et al. A deep reinforcement learning based approach for energy-efficient channel allocation in satellite internet of things[J]. IEEE Access, 2020, 8: 62197-62206.

[4] LEI W L, YE Y, XIAO M. Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(3): 970-979.

[5] HE C F, HU Y, CHEN Y, et al. Joint power allocation and channel assignment for NOMA with deep reinforcement learning[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2200-2210.

[6] ALWARAFY A, CIFTLER B S, ABDALLAH M, et al. DeepRAT: A DRL-based framework for multi-RAT assignment and power allocation in HetNets[C]//2021 IEEE International Conference on Communications Workshops (ICC Workshops). Montreal: IEEE, 2021: 1-6.

[7] MENG F, CHEN P, WU L, et al. Power allocation in multi-user cellular networks: Deep reinforcement learning approaches[J]. IEEE Transactions on Wireless Communications, 2020, 19(10): 6255-6267.

[8] 宗思光, 刘涛, 梁善永. 基于改进遗传算法的干扰资源分配问题研究[J]. 电光与控制, 2018, 25(05): 41-45.
ZONG Si-guang, LIU Tao, LIANG Shan-yong. Research on interference resource allocation based on improved genetic algorithm[J]. Electronics Optics & Control, 2018, 25(05): 41-45. (in Chinese)

[9] WANG Q Y, JIAO D Z, SHI S, et al. Improved ant colony optimization algorithm for jamming resource allocation[J]. Journal of System Simulation, 2021, 33(12): 2967-2974.

[10] 黄星源, 李岩屹. 基于双Q学习算法的干扰资源分配策略[J]. 系统仿真学报, 2021, 33(08): 1801-1808.
HUANG Xing-yuan, LI Yan-yi. Interference resource allocation strategy based on double-Q learning algorithm [J]. Journal of System Simulation, 2021, 33(08): 1801-1808. (in Chinese)

[11] 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 Hua, SONG Bai-lin, JIANG Lei, et al. An intelligent decision-making algorithm for communication countermeasure jamming resource allocation[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3086-3095. (in Chinese)

[12] 饶宁, 许华, 齐子森, 等. 基于最大策略熵深度强化学习的通信干扰资源分配方法[J]. 西北工业大学学报, 2021, 39(05): 1077-1086.
RAO Ning, XU Hua, QI Zi-sen, et al. Allocation method of communication interference resource based on deep reinforcement learning of maximum policy entropy[J]. Journal of Northwestern Polytechnical University, 2021, 39(05): 1077-1086. (in Chinese)

[13] 粟平, 赵国庆, 杨小牛, 等. 信息对抗技术[M]. 北京: 清华大学出版社, 2008.

[14] LEI L, YUAN D, HO C K, et al. Joint optimization of power and channel allocation with non-orthogonal multiple access for 5G cellular systems[C]//2015 IEEE Global Communications Conference(GLOBECOM). San Diego: IEEE, 2015: 1-6.

[15] TAN J, LIANG Y C, ZHANG L, et al. Deep reinforcement learning for joint channel selection and power control in D2D networks[J]. IEEE Transactions on Wireless Communications, 2020, 20(2): 1363-1378.

[16] NIE H R, LI S S, LIU Y. Multi-agent deep reinforcement learning for resource allocation in the multi-objective HetNet[C]//2021 International Wireless Communications and Mobile Computing(IWCMC). Harbin: IEEE, 2021: 116-121.

[17] 邓兵, 张韫, 李炳荣. 通信对抗原理及应用[M]. 北京: 电子工业出版社, 2017: 35-156.

[18] YICK J, MUKHERJEE B, GHOSAL D. Wireless sensor network survey[J]. Computer Networks, 2008, 52(12): 2292-2330.

[19] LI X, FANG J, CHENG W, et al. Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach[J]. IEEE Access, 2018, 6: 25463-25473.

[20] VOLODYMYR M, KORAY K, DAVID S, et al. Playing atari with deep reinforcement learning[C]//2013 Conference and Workshop on Neural Information Processing Systems(NIPS). Lake Tahoe: MIT Press, 2013: 1-9.

[21] VOLODYMYR M, KORAY K, DAVID S, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[22] TOM S, JOHN Q, IOANNIS A, et al. Prioritized experience replay[C]//2016 International Conference on Learning Representations. Caribe Hilton: ICLR, 2016: 1-21.

[23] AREF M A, JAYAWEERA S K, YEPEZ E. Survey on cognitive anti-jamming communications[J]. IET Communications, 2020, 14(18): 3110-3127.

[24] KESKAR N S, NOCEDAL J, TANG P T P, et al. On large-batch training for deep learning: Generalization gap

and sharp minima[C]//2017 International Conference on Learning Representations. Toulon: ICLR, 2017: 1-16.

作者简介

彭 翔 男,1998年4月生,山西大同人.现为空军工程大学信息与导航学院博士研究生.主要研究方向为通信对抗、强化学习、智能决策.

E-mail: pengxiang0538@163.com

许 华 男,1976年4月生,湖北宜昌人.现为空军工程大学信息与导航学院教授、博士生导师.主要研究方向为通信对抗、信号盲处理、智能决策.

E-mail: 13720720010@139.com

蒋 磊 男,1974年6月生,江苏无锡人.现为空军工程大学信息与导航学院副教授、硕士生导师.主要研究方向为通信对抗、无线通信技术.

E-mail: jleimail@126.com

张 悦 男,1989年12月生,陕西西安人.现为空军工程大学信息与导航学院在站博士后、讲师、硕士生导师.主要研究方向为深度学习、强化学习、博弈论、通信资源分配.

E-mail: catchmeifyoucan@uestc.edu.cn

饶 宁 男,1997年8月生,江西上饶人.现为空军工程大学信息与导航学院博士研究生.主要研究方向为通信对抗、强化学习、智能决策.

E-mail: raoningmabma@163.com