基于多智能体深度强化学习的分布式协同干扰功率分配算法

饶 宁,许 华,蒋 磊,宋佰霖,史蕴豪
(空军工程大学信息与导航学院,陕西西安 710077)

摘 要: 针对战场通信对抗协同干扰中的干扰功率分配难题,本文基于多智能体深度强化学习设计了一种分布式协同干扰功率分配算法.具体地,将通信干扰功率分配问题构建为完全协作的多智能体任务,采用集中式训练、分布式决策的方式缓解多智能体系统环境非平稳、决策维度高的问题,减少智能体之间的通信开销,并加入最大策略熵准则控制各智能体的探索效率,以最大化累积干扰奖励和最大化干扰策略熵为优化目标,加速各智能体间协同策略的学习.仿真结果表明,所提出的分布式算法能有效解决高维协同干扰功率分配难题,相比于已有的集中式分配算法具有学习速度更快、波动性更小等优点,且相同条件下干扰效率可高出集中式算法 16.8%.

关键词: 通信对抗;协同功率分配;多智能体深度强化学习;分布式策略;最大策略熵
中图分类号: TN975 **文献标识码:** A **文章编号:** 0372-2112(2022)06-1319-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210818

Allocation Algorithm of Distributed Cooperative Jamming Power Based on Multi-Agent Deep Reinforcement Learning

RAO Ning, XU Hua, JIANG Lei, SONG Bai-lin, SHI Yun-hao
(*Information and Navigation College of Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*)

Abstract: In order to solve the problem of jamming power allocation in battlefield cooperative communication countermeasures, this paper designs a distributed cooperative jamming power allocation method based on multi-agent deep reinforcement learning. Specifically, modeling the communication jamming power allocation as a fully cooperative multi-agent task, then the framework of centralized training and distributed decision-making is adopted to alleviate the characteristic of non-stationary environment and high dimensions in multi-agent system, reducing the communication overhead between agents as well, and introducing the maximum policy entropy criterion to control the exploration efficiency of each agent. Regarding maximizing the cumulative jamming reward and maximizing the entropy of the jamming policy as the optimization goal, then accelerates the learning of cooperative strategies. Simulation results indicate the proposed distributed method can effectively solve the high-dimensional cooperative jamming power allocation problem. Compared with the existing centralized allocation method, it has faster learning speed and less volatility, and the jamming efficiency is 16.8% higher than that of the centralized method under the same conditions.

Key words: communication countermeasures; cooperative resource allocation; multi-agent deep reinforcement learning; distributed strategy; maximum policy entropy

1 引言

电磁空间是继陆、海、空、天的第五维战场.在感知、决策、行动、评估的闭环电磁频谱作战过程中,决策是确保电子对抗效能发挥的关键环节,科学决策可最优化资源的配置利用.近年,智能决策已经成为认知电子战的一个重要研究方向[1],遗传算法、博弈论、分布式优化等理论[2~4]被相继用于干扰参数优化、资源分配等领域,但这些方法都需要较多的先验参数信息.强化学习作为不需要先验信息的机器学习方法,能以与未知环境交互的方式优化策略,目前在通信干扰领域已有初步应用,如文献[5,6]通过建立多臂赌博机模型来学习最佳干扰样式,文献[7]将对无线网络的干扰建模为

收稿日期:2021-06-30;修回日期:2021-10-10;责任编辑:冯珂

增广马尔科夫决策过程，通过实验表明干扰方可通过与环境交互的方式学习到干扰成本、网络吞吐量等重要信息.

随着计算机运算和存储能力的大幅提升，深度学习在人工智能领域获得了巨大成功，其与强化学习相结合的深度强化学习技术在无人驾驶、视频游戏、云边计算服务、机器人控制等领域也展现了惊人的自主决策能力[8~13]. 同时，人工智能也不断驱动无线通信网络的智能化发展[14]. 当前利用深度强化学习解决高维空间的资源分配问题成为研究热点，主要研究成果可分为基于单智能体强化学习的方法[15~17]和基于多智能体强化学习(Multi-Agent Reinforcement Learning，MARL)的方法[18~22]. 在单智能体强化学习的方法中，智能体将所有设备或用户的状态和动作信息集中在一起，构成一个整体的状态和动作空间，通过集中式控制完成用户调度[15]、信道管理[16]和功率分配[17]等任务，但这种集中控制的调度方法不可避免地带来决策维度高、通信开销大、系统扩展性差等问题[23]，一般适用于决策维度较低的场景. 在基于MARL的方法中，每个设备或用户均是一个智能体，通过各智能体协同决策的方式完成任务，可减小神经网络的输入和输出维度[24]. 为进一步提高决策效率，文献[18]在频率切换控制和功率分配的完全协作多智能体任务中采用集中式的策略梯度方法，各设备使用全局状态信息进行训练，得到了较好的协作策略；文献[19]和文献[20]采用分布式深度Q网络，先通过中心节点集中训练，而后将模型参数分发给各基站，提高了业务需求量较大情况下的用户满意度和系统稳定性；文献[21]在分布式深度Q网络中采用竞争双Q网络结构，各用户设备依靠信息传递获得的全局状态信息进行随机博弈；文献[22]假设不同地区的通信链路属性大致相同，使每个智能体可共享一个策略网络，通过集中决策的方式提高了多用户无线蜂窝网络的总传输速率. 综上所述，现有关于通信干扰领域的决策研究相对较少且大部分在信号体制层级[5~7]，而未来电子战的体系对抗模式亟需开展协同干扰资源分配的相关研究.

本文面向对抗组网通信场景下多干扰设备协同干扰中的干扰功率分配问题，提出了一种基于多智能体深度强化学习的分布式干扰功率分配机制(Multi-Agent Distributive Jamming Power Allocation，MADJPA)，通过建立多干扰设备对多通信目标协同干扰的干扰资源分配模型，搭建多干扰设备集中训练与分布执行的决策网络训练架构，并融合强化学习方法和最大策略熵理论设计干扰功率智能分配算法，在满足不同干扰压制系数的整体干扰压制条件下，优化了干扰资源利用，提高了学习最优分配策略的收敛速度.

本文的主要贡献如下.

(1)为了适应对多通信链路的多干扰设备协同干扰任务，将协同干扰功率分配问题转化为完全协作的多智能体任务，建立了战场条件下非完全信息决策的部分马尔科夫决策过程(Partially Observable Markov Decision Processes，POMDP)，在所设计的POMDP奖励函数中，综合考虑了整体干扰压制任务的实现以及干扰功率利用的最优化问题，可以在不同干扰压制系数条件下自适应地调整合理的干扰功率分配方案.

(2)为了降低多干扰设备协同决策的维度，并缓解多干扰设备条件下决策网络训练环境的不稳定性，构建了适用于战场通信对抗场景的集中训练与分布执行的决策网络架构. 基于此架构，单个干扰设备在决策时不需要其他干扰设备的信息，只依靠本地信息即可完成干扰设备之间的协同决策，减少了干扰设备之间由信息交换带来的通信时延和通信开销，更契合战场环境对决策时效性的要求.

(3)设计基于多智能体深度强化学习的分布式干扰功率分配算法，为了加快各干扰设备对全局最优协同策略的学习，在强化学习目标函数中引入干扰策略熵项，使算法在优化过程中同时最大化累积干扰奖励和最大化干扰策略熵，并结合梯度下降自适应调整干扰策略的熵系数，适时地控制各智能体在未知环境中的探索能力，进一步提升算法收敛速度，在干扰压制系数较大的情况下可更精细地分配各干扰设备的干扰功率.

2 对抗模型

本文以战场对抗环境下干扰方对敌方前突飞机遂行压制干扰任务为例，如图1所示. 当敌方飞机发现其所使用的通信链路被干扰后，可切换至区域内其他基站的通信链路继续通信. 干扰方通过指挥控制端对目标频谱进行侦察，指控端内部的智能引擎根据侦察情报完成干扰任务分配，下发至各干扰设备. 假设干扰方有N台干扰设备，干扰设备的集合为$N_s=\{1,2,\cdots,N\}$，干扰设备均采用拦阻干扰样式. 敌方飞机可根据实际情况和自身通联状态与不同基站进行通信，受干扰后可重新选择通信链路，$M_s=\{1,2,\cdots,M\}$表示所有基站的通信链路集合，假设各链路信道为互不干扰、相互独立的等带宽正交信道.

假设干扰方通过通信侦察和情报分析综合掌握了各链路的中心频率，获得了各链路的相对重要性指数：

$$W=[\omega_1,\omega_2,\cdots,\omega_i];\ i\in M_s \tag{1}$$

为尽可能破坏敌通联情况，形成对区域内的完全压制干扰，干扰方应在当前干扰资源条件下合理分配各干扰设备的干扰任务，通过各设备协同合作的方式

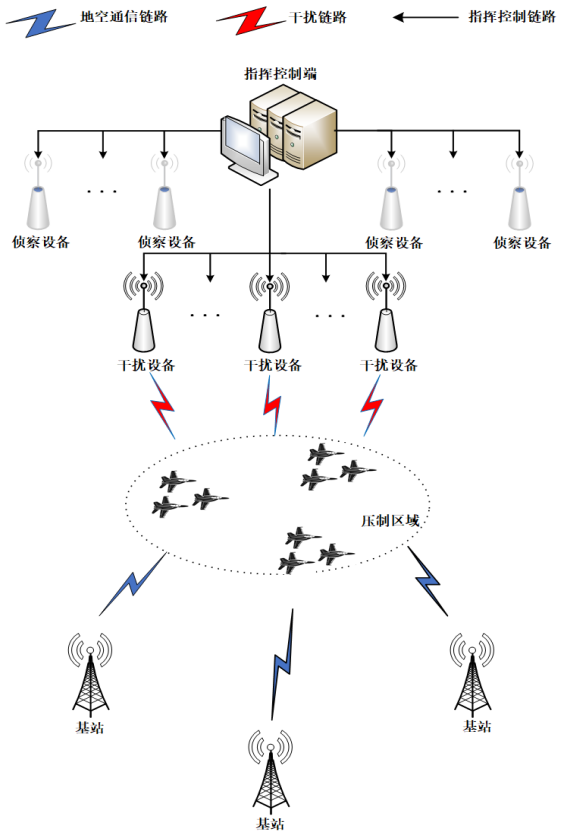


图 1　对抗模型示意图

对侦察到的所有通信链路都进行压制干扰,即达到对整体通信网的完全压制效果.

假设每部干扰设备可同时干扰 U 个目标,对不同目标的干扰功率需满足频分原则,即

$$\sum_{u=1}^{U} P_u \leqslant P_{\max} \tag{2}$$

其中,$P_{\max}$ 为干扰设备的最大辐射功率.

设 t 时刻链路 i 的基站发射信号功率为 P_i,信道增益为 G_i,用 P_i^j 和 G_i^j 分别表示干扰设备 j 分配至该链路的干扰信号功率以及相应的干扰信道增益. 考虑到一条链路可能受到不同干扰设备的干扰,链路 i 中飞机接收电台处的干信比(Jamming Signal Ratio,JSR)可表示为

$$\mathrm{JSR}_i(t)=\frac{\sum_{j=1}^{N} P_i^j(t)G_i^j(t)L_j+\sigma^2}{P_i(t)G_i(t)L_i} \tag{3}$$

其中,σ^2 表示环境噪声功率;L_i 和 L_j 分别表示地空通信链路和干扰链路的传输损耗. 为便于分析不考虑信号的带外损失,假设路径损耗为自由空间传播损耗[25],损耗可表示为

$$L(\mathrm{dB})=32.4+20\lg(f/\mathrm{MHz})+20\lg(d/\mathrm{km}) \tag{4}$$

其中,f 为链路中心频率;d 为信号传播距离.

为了定量描述通信干扰对通信接收机的影响程度,引入干扰压制系数. 当每条链路的干信比均超过干扰压制系数 K,即满足式(5)时视为实现整体完全压制干扰.

$$\mathrm{JSR}_i(t)\geqslant K_i;\ \forall i\in M_s \tag{5}$$

其中,K_i 为链路 i 所对应的干扰压制系数. 各通信链路的干扰压制系数对于干扰方而言是未知的.

结合各链路重要性指数,针对干扰压制系数未知时,实现整体完全压制干扰的干扰功率分配问题可转为求解优化问题,如下所示:

$$\max \sum_{i=1}^{M}\omega_i\cdot\frac{\sum_{j=1}^{N}P_i^jx_i^jG_i^jL_j+\sigma^2}{P_iG_iL_i};\quad \omega_i\in W \tag{6}$$

$$\text{s.t.}\begin{cases} C1:\dfrac{\sum_{j=1}^{N}P_i^jx_i^jG_i^jL_j+\sigma^2}{P_iG_iL_i}\geqslant K_i, & \forall i\in M_s \\ C2:\sum_{i=1}^{M}P_i^j\leqslant P_{\max}, & \forall j\in N_s \\ C3:\sum_{i=1}^{M}x_i^j\leqslant U, & \forall j\in N_s, U\leqslant N \\ C4: x_i^j\in\{0,1\}, & \forall i\in M_s,\forall j\in N_s \end{cases} \tag{7}$$

目标函数式(6)为在优先干扰重要链路的前提下最大化各链路总干信比. 约束条件式(7)中,$C1$ 表示各链路的干信比均需超过干扰压制系数即对整体达到完全压制,$C2$ 需要表示干扰设备对不同目标的干扰功率满足频分原则,$C3$ 和 $C4$ 表示一个干扰设备可同时干扰 U 条链路,x_i^j 是二进制指示变量,$x_i^j=1$ 表示第 j 部干扰设备对链路 i 进行干扰,且干扰功率为 P_i^j.

3　基于MARL的分布式干扰功率分配算法

3.1　POMDP

式(6)、式(7)是非凸优化的 NP-Hard 难题,尤其当同时存在离散空间和连续空间的待优化参数时应用传统数学优化方法难以求解. 本文采用深度强化学习(Deep Reinforcement Learning,DRL)方法解决该问题. 不同于一般的监督学习与无监督学习,DRL 作为不需要先验信息的机器学习方法,采用试错方式进行策略优化,即控制智能体不断与环境交互,根据环境给出反馈修正策略,目的是使得累积奖励期望最大,这种学习方法能够很好地处理本文研究的优化问题.

DRL 需要根据问题模型建立相应的马尔科夫决策过程,本文将多干扰设备的协同资源分配问题建模为完全协作的多智能体任务[26],考虑到战场条件下该任务的非完全信息决策属性,将其定义为 POMDP,可用 $\boldsymbol{\Gamma}=\langle \boldsymbol{S},\boldsymbol{A},P,\boldsymbol{Z},O,r,N,\gamma\rangle$ 表示,其中 $\boldsymbol{S}$ 为全局环境状态空

间，$\boldsymbol{A}$为动作空间，P为状态转移概率，$\boldsymbol{Z}$为局部观测空间，O为观测函数，r为奖励函数，N为智能体数量，γ为折扣因子. POMDP过程可描述如下.

在每个时间步t，每个智能体根据观测函数$O(s)$: $\boldsymbol{S}\rightarrow\boldsymbol{Z}$获得各自对外部环境的观测$z\in\boldsymbol{Z}$，每个智能体（本文为干扰设备）$j\in N_s=\{1,2,\cdots,N\}$，基于各自策略$\pi^j(a_t^j|z_t^j)$: $\boldsymbol{Z}\times\boldsymbol{A}\rightarrow[0,1]$选择动作$a_t^j\in\boldsymbol{A}$，$z_t^j$表示智能体$j$的本地观测. 所有智能体的动作可构成一个联合动作$a_t$. 当$a_t$作用于状态$s_t$下的环境后，环境会根据状态转移函数$P(s_{t+1}|s_t,a_t)$: $\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\rightarrow[0,1]$转移到下一个状态并获得奖励$r_t$，循环往复直到任务结束. 在完全协作的任务中，所有智能体共享一个奖励函数$r(s_t,a_t)$: $\boldsymbol{S}\times\boldsymbol{A}\rightarrow\mathbb{R}$. 各智能体的联合策略可用$\pi=\{\pi^1,\pi^2,\cdots,\pi^N\}$表示，协作任务最终目的是所有智能体协同地找到一个最优联合策略π^*，满足

$$\pi^*=\arg\max_{\pi}\eta(\pi) \tag{8}$$

其中，$\eta(\pi)=\mathrm{E}_{(s_t,a_t)\sim\rho_\pi}\left[\sum_{t=0}^{+\infty}r\left(s_t,a_t\right)\right]$为期望折扣奖励，能使期望折扣奖励最大的策略即为最优策略.

根据本文所研究的多智能体协作任务，将POMDP的元素具体定义如下.

（1）动作

每个干扰设备的动作包括干扰链路的选择和相应功率分配，如干扰设备j的动作可表示为$a_t^j=[p_t^1,p_t^2,\cdots,p_t^i,\cdots,p_t^M]$，其中$0\leqslant p_t^i\leqslant P_{\max},1\leqslant i\leqslant M$. 若该设备选择干扰链路$i$，则$p_t^i\in(0,P_{\max}]$，否则$p_t^i=0$且满足

$$\sum_{i=1}^{M}p_t^i\leqslant P_{\max} \tag{9}$$

$$\sum_{i=1}^{M}\mathrm{sign}(p_t^i)\leqslant U \tag{10}$$

其中，sign(·)为符号函数.

所有干扰设备的联合动作可表示为

$$a_t=(a_t^1,a_t^2,\cdots,a_t^N) \tag{11}$$

（2）全局状态和局部观测

每个干扰设备的局部观测包含该设备上一时刻的干扰方案a_{t-1}^j和对应干扰效果，可表示为$z_t^j=[a_{t-1}^j,\mathrm{JSR}^j]$，其中$\mathrm{JSR}^j=(\mathrm{JSR}_1^j,\mathrm{JSR}_2^j,\cdots,\mathrm{JSR}_M^j)$. 将所有设备的观测集合定义为全局状态，表示为

$$s_t=(z_t^1,z_t^2,\cdots,z_t^N)\in\boldsymbol{S} \tag{12}$$

其中，$\boldsymbol{S}$为全局状态空间.

（3）奖励

MARL中奖励函数可引导算法的优化方向，将奖励函数的设计与实际优化目标联系起来，算法性能可在奖励驱动下得到提高. 本文定义的奖励函数包含对整体的完全压制干扰奖励和干扰功率利用最优化奖励.

定义对整体的完全压制干扰奖励为

$$R_{\mathrm{T}}(s_t,a_t)=k_1+k_2\cdot\sum_{i=1}^{M}\omega_i\cdot\left[\mathrm{sign}\left(\mathrm{JSR}_i(t)-K_j\right)\right] \tag{13}$$

其中，k_1为负常数，k_2为正比例常数，且$\left|k_2\right|\gg\left|k_1\right|$；$\omega_i\cdot\left[\mathrm{sign}\left(\mathrm{JSR}_i(t)-K_i\right)\right]$表示只有对链路达到压制干扰后才会获得干扰收益，且收益与该链路的重要性指数成正比. 加入固定负收益k_1，当功率分配方案不合理时所有链路的总干扰收益$k_2\cdot\sum_{i=1}^{M}\omega_i\cdot\left[\mathrm{sign}\left(\mathrm{JSR}_i(t)-K_i\right)\right]$将无法抵消固定负收益$k_1$，说明该功率分配方案干扰效果不佳，负的收益将使决策网络在梯度反向传播时获得更大的梯度更新值，促进决策网络的优化. 对整体的完全压制干扰奖励函数表明，当对所有链路均可以完全压制时，将获得最大的干扰收益；当干扰资源有限，而无法压制所有链路时，将优先压制干扰相对重要性指数较大的链路.

定义干扰功率利用最优化奖励为

$$R_{\mathrm{P}}(s_t,a_t)=k_3\cdot\sum_{i=1}^{M}\omega_i\cdot\left[\mathrm{sign}\left(\mathrm{JSR}_i(t)-K_i\right)\Big/P_i(t)\right] \tag{14}$$

其中，k_3为正比例常数；$P_i(t)$为各干扰设备对链路i施加的总干扰功率；$\omega_i\cdot\left[\mathrm{sign}\left(\mathrm{JSR}_i(t)-K_i\right)\Big/P_i(t)\right]$表示在对某链路达到压制干扰的前提下，获得的干扰收益与所用的总干扰功率成反比，与链路重要性指数成正比. 此式表明在满足完成压制干扰任务的前提下，使用较小的干扰功率将带来更大的干扰收益，可避免因辐射功率过大而影响己方通信或暴露己方干扰设备位置.

因此总奖励函数为

$$R(s_t,a_t)=R_{\mathrm{T}}(s_t,a_t)+R_{\mathrm{P}}(s_t,a_t) \tag{15}$$

在多干扰设备协同任务中，所有设备共享一个奖励值，利用该公共奖励值驱动MARL算法实现整体的完全压制效果和最优干扰资源利用之间的平衡.

3.2 集中式训练、分布式执行

多智能体协同任务中，各智能体的策略与其他智能体的行为和合作关系相关联，相关的学习算法通常可分为以下几种结构.

一是集中式学习. 将所有智能体的动作和观测进行联合，得到一个扩张的动作空间和观测空间，利用神经网络将所有智能体的联合观测动作映射到一个集中策略函数和集中价值函数，然后直接使用传统的单智能体强化学习方法，如图2所示. 每个设备将自身状态上传至集中网络，由集中策略网络统一决策所有设备的策略. 此种学习方式会使得联合观测和状态空间随着智能体数量的增加而扩大，如本文假设干扰设备数量为N，通信链路数为M，由POMDP可知集中策略网络

输入维度为$2MN$,输出维度为MN. 随着干扰设备数量N增多,集中策略网络的维度增多,策略探索开销变大[23].

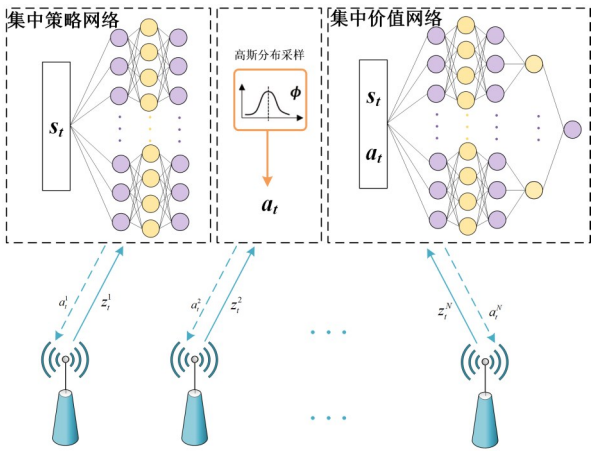


图2　集中式学习

二是独立式学习. 各智能体独立维护自身策略函数和价值函数,且各函数的输入只依赖智能体各自观测和动作,各智能体基于自身策略网络独立决策并独立训练自身网络. 此种学习方式中,策略网络的输入维度为$2M$,输出维度为M,与智能体数量无关. 但对于某个特定智能体而言,由于其他智能体学习过程中策略不断变化,容易造成环境非平稳,训练难以收敛.

三是值函数分解. 在独立式学习的基础上,将各智能体的值函数进行加和,以值函数近似的方式求解全局值函数,然后站在全局的角度去优化更新每个智能体的值函数. 此种学习方式能解决环境非平稳性,但只适用于离散动作空间,不适用于本文所研究的问题.

本文将集中式学习与独立式学习的优势相融合,采用集中式训练、分布式决策的结构,如图3所示.

"集中式训练"重点在于每个设备在训练时需将其他干扰设备的观测和动作(可视为全局状态信息)输入

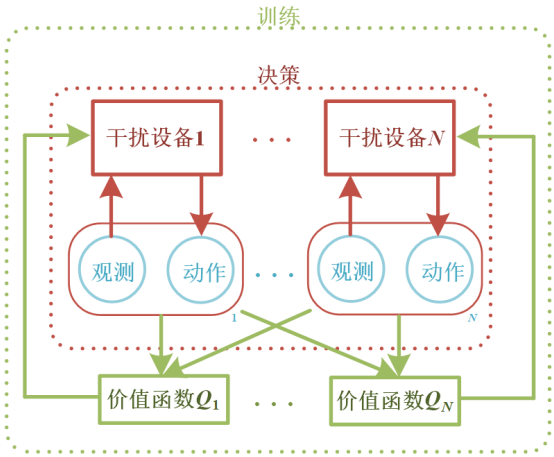


图3　集中式训练、分布式决策结构

其评估网络,通过集中式地评估联合动作来增强各干扰设备的协调配合. 集中评估的方式可使得其他干扰设备策略相对已知,克服了策略变化造成的环境不平稳[27]. "分布式决策"意为各设备在决策干扰动作时只需将各自的观测输入至各自的策略网络中即可完成协同决策,不再需要中心控制器集中处理各干扰设备的联合观测信息. 此时智能体策略网络的输入维度为$2M$,输出维度为M,相比于集中式学习,决策维度降低了$M(N-1)$,而由实际经验看,决策维度太大是导致决策失败的重要原因之一,降低决策维度可提升方法的可行性.

3.3　MADJPA算法

在集中式训练、分布式决策框架下,为了提高每个智能体在未知环境中的探索效率,本文采用同时最大累积奖励和策略熵[28]的优化路线,在式(8)中加入策略熵项,即

$$\pi^*=\arg\max_{\pi}\mathrm{E}_{(s_t,a_t)\sim\rho_\pi}\left[\sum_{t=0}^{+\infty}R\left(s_t,a_t\right)+\alpha H(\pi(\cdot|s_t))\right] \tag{16}$$

其中,$H(\pi(\cdot|s_t))=-\log\left(\pi\left(a_t|s_t\right)\right)$,$\alpha$为熵系数.

策略熵即策略分布熵,当策略熵较大时意味着策略的随机性较强,在未知环境中的探索能力较强,而适度的探索可实现对环境模型的充分学习,避免陷入局部最优.

为平衡智能体在未知环境中的探索和对现有知识的利用,设置熵系数α的优化目标函数[24],以梯度下降方式更新其值,熵系数优化目标函数为

$$J(\alpha)=\mathrm{E}_{a_t\sim\pi}\left[-\alpha\log\pi\left(a_t|\ s_t\right)-\alpha\bar{\mathcal{H}}\right] \tag{17}$$

其中,$\bar{\mathcal{H}}$为智能体的动作维度.

初始阶段,α值较大策略随机性也较大,探索效率较高;随着智能体对环境模型的不断学习,α自适应减小;当α下降至0时,式(16)中无熵项,此时智能体的优化目标就变为传统的最大化累积奖励.

在递归求解最佳策略π^*时采用的Q函数迭代式可表示为

$$Q\left(s_t,a_t\right)=r_t+\gamma\mathrm{E}\left[Q\left(s_{t+1},a_{t+1}\right)-\alpha\log\left(\pi_\phi\left(a_{t+1}|s_{t+1}\right)\right)\right] \tag{18}$$

本文用神经网络拟合Q函数和策略函数,并采用Kullback-Leibler(KL)散度约束来更新策略[29],即

$$\pi_{\mathrm{new}}=\arg\min D_{\mathrm{KL}}\left(\pi_\phi\left(\cdot|s_t\right)\middle\|\frac{\exp\left(\frac{1}{\alpha}Q^{\pi_{\mathrm{old}}}\left(s_t,\cdot\right)\right)}{Z^{\pi_{\mathrm{old}}}\left(s_t\right)}\right) \tag{19}$$

其中,$D_{\mathrm{KL}}(\cdot)$表示KL散度约束;$Q^{\pi_{\mathrm{old}}}\left(s_i,\cdot\right)$表示原策略下

的Q函数;$Z^{\pi_{\text{old}}}(s_i)$表示原策略的对数配分函数.

在各个智能体内部进行基于最大策略熵深度强化学习的策略优化,据此本文提出了基于MARL的分布式干扰功率分配算法(MADJPA),图4所示是MADJPA算法示意图.

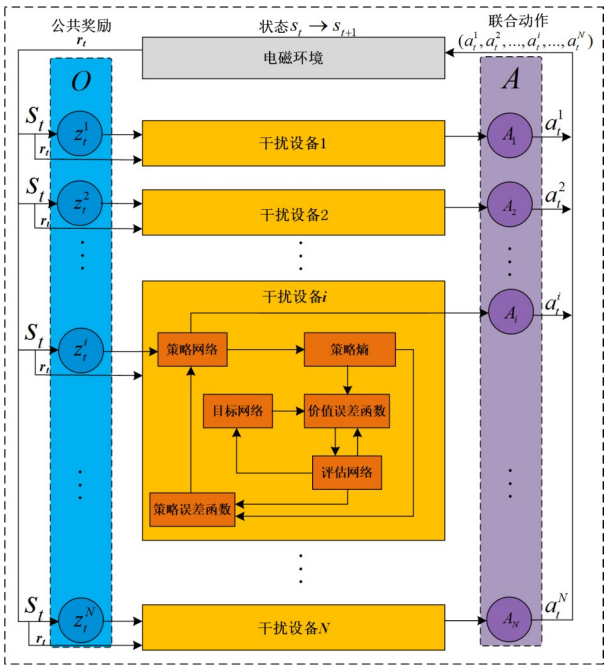


图4 MADJPA算法示意图

每一时间步t,各个干扰设备根据自身策略网络同时决策,得到当前局部观测下执行的干扰动作$a_t^i=\pi_i(z_t^i)$,执行联合动作$a_t=(a_t^1,a_t^2,\cdots,a_t^N)$后得到一个公共奖励,将与环境交互得到的经验$(s_t,a_t^1,a_t^2,\cdots,a_t^i,\cdots,a_t^N,s_{t+1},r_t)$存入公共的经验回放池(Common Replay Buffer, CRB),其中$s_t=(z_t^1,z_t^2,\cdots,z_t^i,\cdots,z_t^N)$. 当经验回放池内样本积累到一定阶段时,从经验池中随机抽取一批样本,供每个设备训练各自的神经网络. 在训练过程中,为提升网络训练稳定性,引入与评估网络结构完全相同的目标网络[30],用目标网络的输出Q值与公共奖励之和作为评估网络训练的标签.

此外,为了避免评估网络对动作Q值的过高估计,本文在评估网络中采用孪生网络结构[31],即评估网络内部有两个结构相同的神经网络,每次取两者输出较小的结果计算干扰动作的目标价值,故式(18)可改写为

$$\begin{aligned}Q_i(s_t,a_t)=r_t+\gamma\mathrm{E}\Big[&\min_{i=1,2}Q_i(s_{t+1},a_{t+1})\\&-\alpha\log\left(\pi_\phi(a_{t+1}|s_{t+1})\right)\Big]\end{aligned}\quad(20)$$

for $i=1,2$

利用孪生网络的输出计算联合干扰动作的目标价值为

$$\begin{aligned}&y\left(r(a_k^1,a_k^2,\cdots,a_k^N,s_k),s_{k+1}\right)\\&=r(a_k^1,a_k^2,\cdots,a_k^N,s_k)\\&+\gamma\left[\min_{j=1,2}\bar{Q}_{ij}(s_k,\bar{a}')-\alpha\log\pi_\phi(\bar{a}'|s_{k+1})\right]\\&\bar{a}'\sim\pi_{i=1,2,\cdots,N}(\cdot|s_{k+1},\phi_i)\end{aligned}\quad(21)$$

并更新孪生评估网络参数θ_{i1}和θ_{i2},即

$$\nabla_{\theta_{ij}}\frac{1}{|B|}\sum_{(s_k,a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N,s_{k+1},r_k)\in B}\left(Q_{\theta_{ij}}(s_k,a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N)-y\left(r(a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N,s_k),s_{k+1}\right)\right)^2\quad(22)$$

for $j=1,2$

$$\theta_{ij}\leftarrow\theta_{ij}-\nabla_{\theta_{ij}}J_Q(\theta_{ij}),\quad\text{for}\,j=1,2\quad(23)$$

紧接着更新策略网络的参数ϕ_i,即

$$\nabla_{\phi_i}\frac{1}{|B|}\sum_{s_k\in B}\left(\min Q_{\theta_{ij}}(s_k,\bar{a}_{\phi_i}(s_k))-\alpha\log\pi_{\phi_i}(\bar{a}_{\phi_i}(s_k)|s_k)\right)^2$$

$$\bar{a}_{\phi_i}(s_k)\sim\pi_{i=1,2,\cdots,N}(\cdot|s_k,\phi_i)\quad(24)$$

$$\phi_i\leftarrow\phi_i-\nabla_{\phi_i}J_\pi(\phi_i)\quad(25)$$

最后以柔性赋值的方式更新孪生目标网络参数$\bar{\theta}_{i1}$和$\bar{\theta}_{i2}$,即

$$\bar{\theta}_{ij}\leftarrow\tau\cdot\bar{\theta}_{ij}+(1-\tau)\cdot\theta_{ij},\quad\text{for}\,j=1,2\quad(26)$$

当训练完成后各干扰设备获得分布式干扰策略,执行任务分配时每个干扰设备仅依靠本地观测可完成决策. 分布式策略如图5所示.

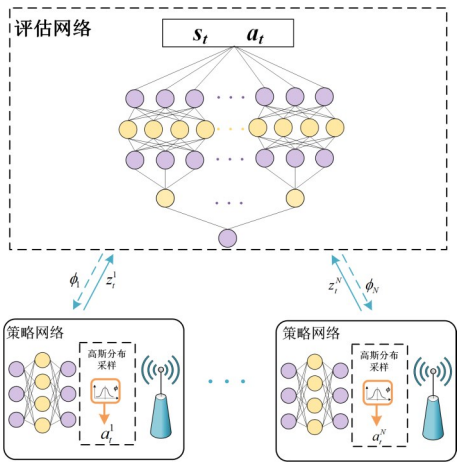


图5 分布式策略

算法伪代码如算法1所示.

算法1　MADJPA伪代码

输入:干扰设备数量N,设备编号$i\in\{1,2,\cdots,N\}$;通信链路数M;干扰压制系数K_j

输出:各设备干扰策略$\pi_i^*, i\in\{1,2,\cdots,N\}$

开始:

步骤1:初始化每个干扰设备的策略网络$\pi_i(z_t,\phi_i)$,以及孪生评估网络$Q_{i1}(s,a_t^1,a_t^2,\cdots,a_t^N,\theta_{i1})$,$Q_{i2}(s,a_t^1,a_t^2,\cdots,a_t^N,\theta_{i2})$,网络参数分别为$\phi_i$,$\theta_{i1}$,$\theta_{i2}$;

步骤2:初始化每个干扰设备的目标孪生网络$\bar{Q}_{i1}(s,a_t^1,a_t^2,\cdots,a_t^N,\bar{\theta}_{i1})$,$\bar{Q}_{i2}(s,a_t^1,a_t^2,\cdots,a_t^N,\bar{\theta}_{i2})$,网络参数分别为$\bar{\theta}_{i1}$,$\bar{\theta}_{i2}$;

步骤3:初始化共享经验回放池CRB;

步骤4:

FOR each episode:

　初始化环境和状态;

　FOR each step t:

　　对每个干扰设备i:

　　根据当前的观测z_t^i,利用策略网络选择干扰方案$a_t^i\sim\pi_i(a_t^i|z_t^i,\phi_i)$;

　　得到各设备的联合干扰方案,执行当前干扰方案a_t^i得到下一观测z_{t+1}^i和公共奖励r_t;

　　将所有干扰设备的经验$(s_t,a_t^1,a_t^2,\cdots,a_t^i,\cdots,a_t^N,s_{t+1},r_t)$存入公共经验回放池CRB:

$$D\leftarrow D\cup\left\{(s_t,a_t^1,a_t^2,\cdots,a_t^i,\cdots,a_t^N,s_{t+1},r_t)\right\}$$

　　当CRB内样本数量大于τ时,训练网络:

　　从CRB中采样小批次样本

$$B=\left\{\cdots,(s_k,a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N,s_{k+1},r_k),\cdots\right\}_{\text{Length}=\text{batch\_size}}$$

　　对每个干扰设备i:

　　计算干扰方案目标价值

$$y\left(r(a_k^1,a_k^2,\cdots,a_k^N,s_k),s_{k+1}\right)$$

$$=r(a_k^1,a_k^2,\cdots,a_k^N,s_k)+\gamma\left[\min_{j=1,2}\bar{Q}_{ij}(s_k,\bar{a}')-\alpha\log\pi_\phi(\bar{a}'|s_{k+1})\right]$$

$$\bar{a}'\sim\pi_{i=1,2,\cdots,N}(\cdot|s_{k+1},\phi_i)$$

　　利用梯度下降更新孪生评估网络参数θ_{i1}和θ_{i2},即

$$\nabla_{\theta_{ij}}\frac{1}{|B|}\sum_{(s_k,a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N,s_{k+1},r_k)\in B}\begin{pmatrix}Q_{\theta_{ij}}(s_k,a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N)\\-y\left(r(a_k^1,a_k^2,\cdots,a_k^i,\cdots,a_k^N,s_k),s_{k+1}\right)\end{pmatrix}^2$$

$\text{for}\ j=1,2$

$$\theta_{ij}\leftarrow\theta_{ij}-\nabla_{\theta_{ij}}J_Q(\theta_{ij}),\ \text{for}\ j=1,2$$

　　利用梯度下降更新策略网络参数ϕ_i,即

$$\nabla_{\phi_i}\frac{1}{|B|}\sum_{s_k\in B}\left(\min Q_{\theta_{ij}}(s_k,\bar{a}_{\phi_i}(s_k))-\alpha\log\pi_{\phi_i}(\bar{a}_{\phi_i}(s_k)|s_k)\right)^2$$

$$\bar{a}_{\phi_i}(s_k)\sim\pi_{i=1,2,\cdots,N}(\cdot|s_k,\phi_i)$$

$$\phi_i\leftarrow\phi_i-\nabla_{\phi_i}J_\pi(\phi_i)$$

　　柔性更新孪生目标网络参数$\bar{\theta}_{i1}$和$\bar{\theta}_{i2}$,即

$$\bar{\theta}_{ij}\leftarrow\tau\cdot\bar{\theta}_{ij}+(1-\tau)\cdot\theta_{ij},\quad\text{for}\ j=1,2$$

　END FOR

END For

结束 得到各设备策略网络$\pi_i^*, i\in\{1,2,\cdots,N\}$

4　算法计算复杂度分析和收敛性证明

4.1　计算复杂度

本文分布式算法计算复杂度主要由评估网络和策略网络的网络结构决定. 假设评估网络是隐藏层为H_c层的全连接网络,第h层$(1<h<H_c)$隐藏层含n_h^c个神经元;输入层由联合状态和联合动作的维度决定,为$3MN$;输出层神经元个数为1;因此评估网络总神经元数为$3MNn_1^c+\sum_{h=2}^{H_c}n_{h-1}^c n_h^c+n_{H_c}^c$. 同样地,设策略网络是隐藏层为$H_a$层的全连接网络,第$h$层$(1<h<H_a)$隐藏层含$n_h^a$个神经元;输入层由局部状态维度决定,为$2M$;输出层神经元个数为$M$;因此策略网络总神经元数为$2Mn_1^a+\sum_{h=2}^{H_a}n_{h-1}^a n_{H_a}^a+n_{H_a}^a M$. 若训练一个神经元权重的计算复杂度为$W_i$,则本文分布式算法的计算复杂度为 $O\left(W_i\left[3MNn_1^c+\sum_{h=2}^{H_c}n_{h-1}^c n_h^c+n_{H_c}^c+2Mn_1^a+\sum_{h=2}^{H_a}n_{h-1}^a n_{H_a}^a+n_{H_a}^a M\right]\right)$,可知算法计算复杂度与干扰设备数量$N$、通信链路数$M$成正相关.

集中式学习算法中评估网络复杂度与分布式算法相同,区别在于其策略网络需要输入所有设备的观测并输出所有设备的方案,故输入神经元个数为$2MN$,输出层神经元个数为MN. 可得集中式算法计算复杂度为 $O\left(W_i\left[3MNn_1^c+\sum_{h=2}^{H_c}n_{h-1}^c n_h^c+n_{H_c}^c+2MNn_1^a+\sum_{h=2}^{H_a}n_{h-1}^a n_{H_a}^a+n_{H_a}^a MN\right]\right)$. 对比可知,集中式算法的复杂度高出分布式算法$O\left(W_i\left[2M(N-1)n_1^a+n_{H_a}^a M(N-1)\right]\right)$. 此外由3.2节可知集中式学习的决策维度高出分布式算法$2M(N-1)$.

4.2　算法收敛性证明

对于本文算法的收敛性分析,给出下述定理.

定理　在联合策略集合$\boldsymbol{\Pi}$中,当动作空间维度有限即$|\boldsymbol{A}|<\infty$,存在策略$\pi\in\boldsymbol{\Pi}$,可收敛至最佳联合策略π^*,且有$Q^{\pi^*}(s_t,a_t)\geqslant Q^\pi(s_t,a_t),\forall\pi\in\boldsymbol{\Pi}$.

证明　将策略迭代优化分为策略评估和策略改进两个阶段,在策略评估中,定义带熵奖励为

$$r_\pi\left(s_t,a_t\right)\triangleq r\left(s_t,a_t\right)+\mathrm{E}_{s_{t+1}\sim p}\left[\mathcal{H}\left(\pi\left(\cdot|s_{t+1}\right)\right)\right]\tag{27}$$

将式(18)重写为

$$Q\left(s_t,a_t\right)=r_\pi\left(s_t,a_t\right)+\gamma\mathrm{E}_{s_{t+1}\sim p,a_{t+1}\sim\pi}\left[Q\left(s_{t+1},a_{t+1}\right)\right]\tag{28}$$

根据贝尔曼迭代公式有

$$\begin{aligned} Q^{\pi_{\text{old}}}(s_t,a_t) &= r(s_t,a_t)+\gamma \mathrm{E}_{s_{t+1}\sim p}\left[V^{\pi_{\text{old}}}(s_{t+1})\right] \\ &\leqslant r(s_t,a_t)+\gamma \mathrm{E}_{s_{t+1}\sim p}\left[\mathrm{E}_{a_{t+1}\sim \pi_{\text{new}}}\left[Q^{\pi_{\text{old}}}(s_{t+1},a_{t+1})-\log \pi_{\text{new}}(a_{t+1}|s_{t+1})\right]\right] \\ &\vdots \\ &\leqslant Q^{\pi_{\text{new}}}(s_t,a_t) \end{aligned} \tag{29}$$

令 π^i 表示第 i 次迭代时的策略，可知序列 $\{Q^{\pi^1},Q^{\pi^2},\cdots,Q^{\pi^i}\}$ 是单调递增的，由于奖励和熵有界，故该序列可收敛于某个最佳策略 π^*.

在策略改进中，令式(19)中 $\pi_{\text{new}}(\cdot|s_t)$ 为

$$\begin{aligned} \pi_{\text{new}}(\cdot|s_t) &= \arg\min_{\pi'\in\Pi} D_{\text{KL}}\left(\pi'(\cdot|s_t)\,/\!/\,\exp\left(Q^{\pi_{\text{old}}}(s_t,\cdot)-\log Z^{\pi_{\text{old}}}(s_t)\right)\right) \\ &= \arg\min_{\pi'\in\Pi} J_{\pi_{\text{old}}}(\pi'(\cdot|s_t)) \end{aligned} \tag{30}$$

对所有的 $\pi\in\boldsymbol{\Pi},\pi\neq\pi^*$ 易知 $J_{\pi_{\text{old}}}(\pi_{\text{new}}(\cdot|s_t))\leqslant J_{\pi_{\text{old}}}(\pi_{\text{old}}(\cdot|s))$，同样利用策略评估中的迭代证明，可得对所有的 (s_t,a_t) 均有 $Q^{\pi^*}(s_t,a_t)\geqslant Q^{\pi}(s_t,a_t)$. 可知 $\boldsymbol{\Pi}$ 中其他策略的 Q^{π} 低于收敛后的策略，因此 π^* 为 $\boldsymbol{\Pi}$ 中最优.

5 实验仿真与分析

5.1 仿真参数设定

仿真场景中，设干扰方指挥控制端下属3个干扰设备，每个干扰设备可同时干扰2个目标. 在敌任务区域内有若干地面通信基站，为敌机提供5条可用的通信链路，假设每条通信链路有相同的干扰压制系数. 表1是经过通信侦察情报分析处理后获得的各链路综合情报信息. 为实现对敌机群任务空域内的整体完全压制干扰，在计算信号传播路径损耗时，均以各基站中与飞机电台最近的距离为通信信号的传播距离，以干扰设备中与飞机电台的最远距离作为干扰信号的传播距离.

表1 各通信链路信息

编号	类型	中心频率/MHz	跳频点数	重要性指数
链路1	定频	230.25	—	0.2
链路2	定频	275.50	—	0.3
链路3	跳频	—	160	0.6
链路4	定频	366.00	—	0.5
链路5	跳频	—	200	0.8

实验及网络模型参数如表2.

智能体的策略网络、评估网络、目标网络的隐藏层均为3层全连接网络，每层神经元分别为256，128，64，网络优化器均采用Adam，且策略网络的学习速率为0.000 1，评估网络和目标网络的学习速率为0.000 3，激活函数为Relu，策略网络输出层为tanh.

5.2 实验结果分析

在相同实验条件下，将本文所提的分布式算法MADJPA与文献[22]中的算法进行比较，文献[22]采用的是基于深度确定性策略梯度的集中式分配策略(Deep Deterministic Policy Gradient Centralized Allocation，DDPGCA). 此外为定性分析最大策略熵对算法性能的影响，增加本文算法优化函数中不含熵项的MADJPA-No Entropy算法(记为MADJPA-NE)的消融对比.

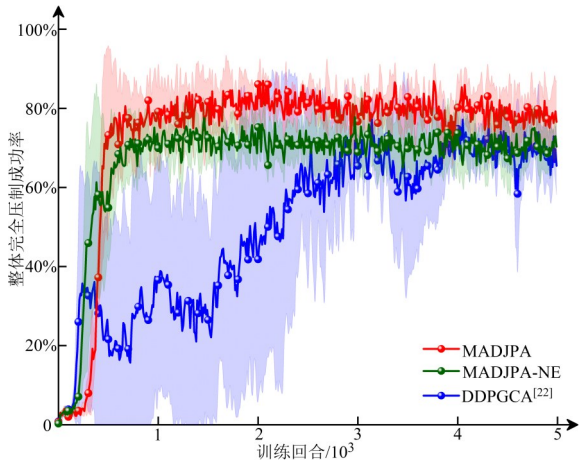
表2 实验及网络模型参数

参数	取值
干扰设备数量 N	3
通信链路数量 M	5
单设备最多可同时干扰目标数 U	2
总干扰带宽 B_j	2 MHz
定频链路带宽 B_d	50 kHz
跳频链路频率间隔 f_i	25 kHz·n(n=1,2,3,4)
干扰设备最大辐射功率 $P_{\max}$	77 dBm(约50 kW)
通信电台辐射功率 P_c	55 dBm(约300 W)
噪声功率 σ^2	-85 dBm
通信链路增益 G_c	8 dB
干扰链路增益 G_j	3 dB
基站与电台最近距离 R_c	110 km
干扰设备与电台最远距离 R_j	300 km
压制系数 K_i	2
柔性更新系数 τ	0.01
训练回合数 E	5 000
每回合交互次数 T	500
经验回放池容量CRB	2^{17}
批次样本大小 B	256
折扣因子 γ	0.98
熵系数初始值 α	1

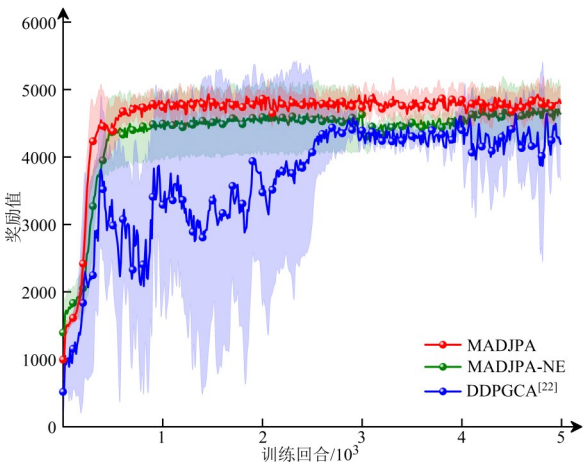
首先分析3种算法对所有通信链路的整体完全压制效果，考察各算法对协同策略的学习能力.

图6(a)和(b)是压制系数为2时3种算法对所有通信链路的整体完全压制效果. 从图6(a)的学习曲线可以明显看出分布式的MADJPA和MADJPA-NE算法学习速度均相对更快，在300~500回合左右整体完全压制

成功率有较大提升,其中MADJPA最高整体压制成功率可达85%以上,MADJPA-NE由于只追求最大化累积奖励而未同时最大化策略熵,探索略有不足,容易陷入局部最优,最高整体完全压制成功率不如MADJPA. 而集中式的DDPGCA初始阶段由于各设备的联合干扰动作空间较大,探索的时间较长,学习过程波动性较大,加之DDPGCA采用的深度确定性策略本身对未知干扰动作探索效率不够,最终整体完全压制成功率在70%左右. 可见将各干扰设备的干扰动作空间联合起来集中决策,会增加决策的复杂度,无论是从收敛速度还是收敛后的效果看,都不如分布式策略,并且加入最大策略熵准则后,分布式策略的整体性能可得到一定提升.



(a)　整体完全压制成功率

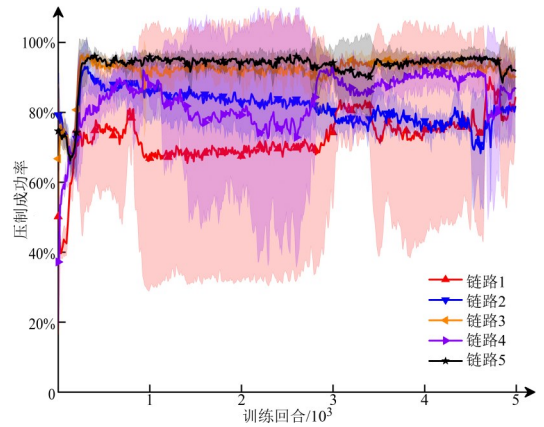


(b)　奖励值

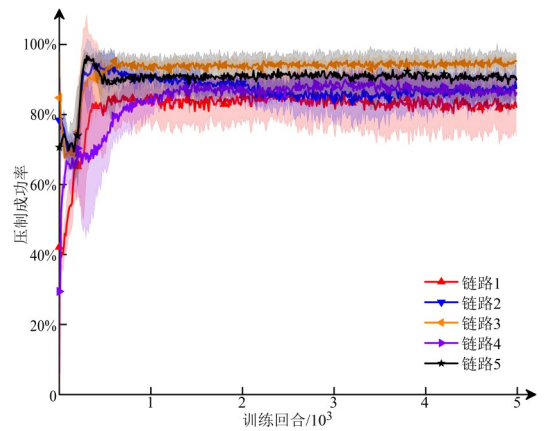
图6　整体完全压制效果

此外,图6中阴影部分表示根据500次重复实验结果计算的波动范围,图6(a)和(b)均表现出DDPGCA整体振荡幅度较大,而MADJPA和MADJPA-NE学习过程相对稳定,波动性较小,其原因在于DDPGCA作为集中式策略,是在更高维的扩张动作空间进行策略探索和优化的,高维空间会增加决策困难度,而MADJPA是分布式策略,其决策维度取决于单个设备的动作空间,维度相对更小,学习效率更高.

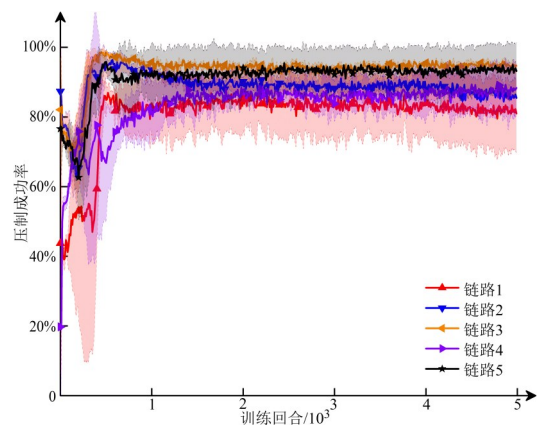
下面对比算法对各链路的压制效果. 图7是3种算法对各链路的压制成功率曲线. 从图7可知,3种算法均能优先干扰重要性指数相对较高的链路5和链路3,然而对于重要性指数较低的链路1和链路2,DDPGCA



(a)　DDPGCA



(b)　MADJPA-NE



(c)　MADJPA

图7　3种算法对各链路的压制成功率曲线

的压制率不高，相较而言，其他2种算法能更好地协调各设备的干扰功率分配，各链路的压制成功率均较DDPGCA有所提升，表明分布式的MADJPA算法更有利于协同策略的学习．此外，同样由阴影部分可知分布式算法的学习过程更平稳．

为考察各算法在实现整体完全压制的同时能否尽量减少资源利用，对比了3种算法对所有链路分配的总干扰功率，结果如图8所示．当3个干扰设备额定最大功率和为81.2 dBm，3种算法均能一定程度地减少资源利用，MADJPA算法最终分配给各链路的干扰功率为80 dBm左右，相比于全功率干扰节省了一定的干扰资源，同样也比DDPGCA算法更节省干扰功率．战场环境下，在压制敌方的前提下减小自身辐射功率可减轻对己方通信的影响，也可避免功率过大暴露自身位置．

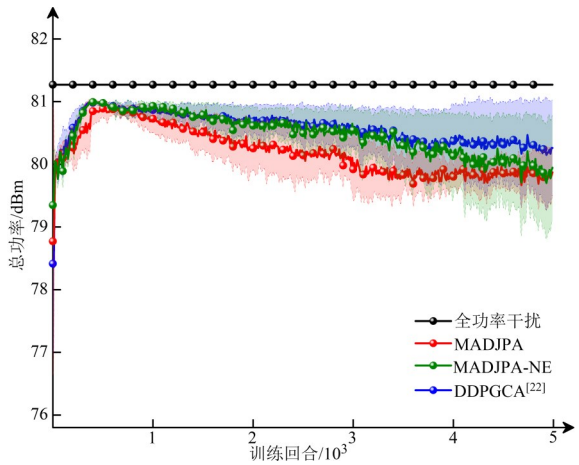


图8 分配的总干扰功率对比

最后对比了不同干扰压制系数条件下各算法能达到的最高整体完全压制成功率，如图9所示．当压制系数变大时，对相同目标压制干扰所需的资源更多，在有限资源条件下需要更合理更精细地协调各干扰设备的干扰功率分配．图9中随着压制系数上升，3种算法的整体完全压制成功率都呈下降趋势．其中压制系数为2时，MADJPA整体完全压制成功率比DDPGCA高出12.5%；当压制系数为4时，相对地MADJPA可高出16.8%．在压制系数较大的条件下，集中式算法分配各设备干扰资源的效率较低，原因在于以单智能体集中决策的形式造成了维度更高的干扰动作空间，难以协调各干扰设备的任务调配，而分布式算法通过多智能体协作的方式降低了各设备决策维度，通过全局信息训练各设备策略网络的方式可更好地调度各设备的干扰功率，分布式算法的协同资源分配能力相比于集中式算法表现较优．

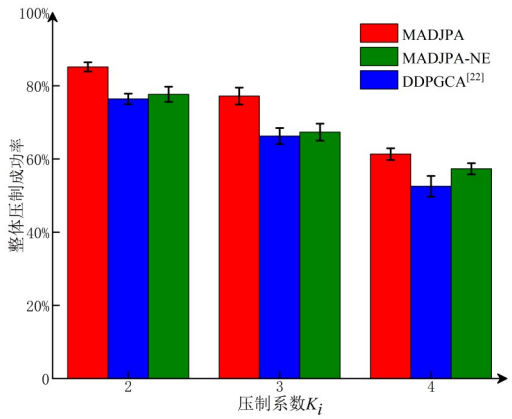


图9 不同压制系数的整体完全压制成功率

6 结论

针对通信组网对抗中的协同干扰功率分配问题，本文基于多智能体深度强化学习提出了一种新的分布式干扰功率分配算法．算法通过构建完全协作的多智能体任务，在集中训练、分布决策的框架中将各干扰设备分别作为一个智能体，在训练时共享全局信息，并利用最大策略熵准则加速智能体间协同策略的学习．相较于集中式的分配算法，本文提出的分布式算法收敛速度更快，学习过程更稳定，且干扰效率高于集中式算法．

参考文献

[1] 王沙飞, 鲍雁飞, 李岩. 认知电子战体系结构与技术[J]. 中国科学: 信息科学, 2018, 48(12): 1603-1613, 1709.
WANG S F, BAO Y F, LI Y. The architecture and technology of cognitive electronic warfare[J]. Scientia Sinica(Informationis), 2018, 48(12): 1603-1613, 1709. (in Chinese)

[2] BAYRAM S, VANLI N D, DULEK B, et al. Optimum power allocation for average power constrained jammers in the presence of non-Gaussian noise[J]. IEEE Communications Letters, 2012, 16(8): 1153-1156.

[3] XU C, SHENG M, WANG X J, et al. Distributed subchannel allocation for interference mitigation in OFDMA femtocells: A utility-based learning approach[J]. IEEE Transactions on Vehicular Technology, 2015, 64(6): 2463-2475.

[4] GOMADAM K, CADAMBE V R, JAFAR S A. Approaching the capacity of wireless networks through distributed interference alignment[C]//2008 IEEE Global Telecommunications Conference. New Orleans: IEEE, 2008: 1-6.

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

[6] 颛孙少帅, 杨俊安, 刘辉, 等. 基于正强化学习和正交分解的干扰策略选择算法[J]. 系统工程与电子技术, 2018, 40(3): 518-525.
ZHUANSUN S S, YANG J N, LIU H, et al. Jamming strategy learning based on positive reinforcement learning and orthogonal decomposition[J]. Systems Engineering and Electronics, 2018, 40(3): 518-525. (in Chinese)

[7] AMURU S, BUEHRER R M. Optimal jamming using delayed learning[C]//2014 IEEE Military Communications Conference. Baltimore: IEEE, 2014: 1528-1533.

[8] 黄志清, 曲志伟, 张吉, 等. 基于深度强化学习的端到端无人驾驶决策[J]. 电子学报, 2020, 48(9): 1711-1719.
HUANG Z Q, QU Z W, ZHANG J, et al. End-to-end autonomous driving decision based on deep reinforcement learning[J]. Acta Electronica Sinica, 2020, 48(9): 1711-1719. (in Chinese)

[9] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search [J]. Nature, 2016, 529(7587): 484-489.

[10] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575(7782): 350-354.

[11] 陈思光, 陈佳民, 赵传信. 基于深度强化学习的云边协同计算迁移研究[J]. 电子学报, 2021, 49(1): 157-166.
CHEN S G, CHEN J M, ZHAO C X. Deep reinforcement learning based cloud-edge collaborative computation offloading mechanism[J]. Acta Electronica Sinica, 2021, 49(1): 157-166. (in Chinese)

[12] LI S, YAN Y H, REN J, et al. A sample-efficient actor-critic algorithm for recommendation diversification[J]. Chinese Journal of Electronics, 2020, 29(1): 89-96.

[13] 杨启萌, 禹龙, 田生伟, 等. 基于深度强化学习的维吾尔语人称代词指代消解[J]. 电子学报, 2020, 48(6): 1077-1083.
YANG Q M, YU L, TIAN S W, et al. Anaphora resolution of uyghur personal pronouns based on deep reinforcement learning[J]. Acta Electronica Sinica, 2020, 48(6): 1077-1083. (in Chinese)

[14] LUONG N C, HOANG D T, GONG S M, et al. Applications of deep reinforcement learning in communications and networking: A survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.

[15] ZHAO D, QIN H, SONG B, et al. A graph convolutional network-based deep reinforcement learning approach for resource allocation in a cognitive radio network[J]. Sensors(Basel, Switzerland), 2020, 20(18): 5216-5239.

[16] WANG S X, LIU H P, GOMES P H, et al. Deep reinforcement learning for dynamic multichannel access in wireless networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(2): 257-265.

[17] XU Z Y, WANG Y Z, TANG J, et al. A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs[C]//2017 IEEE International Conference on Communications. Paris: IEEE, 2017: 1-6.

[18] GUO D L, TANG L, ZHANG X G, et al. Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 13124-13138.

[19] 刘婷婷, 罗义南, 杨晨阳. 基于多智能体深度强化学习的分布式干扰协调[J]. 通信学报, 2020, 41(7): 38-48.
LIU T T, LUO Y N, YANG C Y. Distributed interference coordination based on multi-agent deep reinforcement learning[J]. Journal on Communications, 2020, 41(7): 38-48. (in Chinese)

[20] NASIR Y S, GUO D N. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(10): 2239-2250.

[21] ZHAO N, LIANG Y C, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.

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

[23] ZHANG K Q, YANG Z R, BAŞAR T. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms[M/OL]. [2021]. https://link.springer.com/chapter/10.1007/978-3-030-60990-0_12.

[24] NGUYEN T T, NGUYEN N D, NAHAVANDI S. Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 3826-3839.

[25] 冯小平, 李鹏, 杨绍全. 通信对抗原理[M]. 西安: 西安电子科技大学出版社, 2009.

[26] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[C]//Proceedings of the 32nd AAAI Conference on Artificial Intelli-

gence(AAAI). New Orleans:ACM, 2018: 2974-2983.
[27] LOWE R, WU Y, TAMAR A, et al. Multiagent actor-critic for mixed cooperative-competitive environments[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS). Long Beach: MIT Press, 2017: 6379-6390.
[28] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning(ICML). Stockholm: IMLS, 2018: 1861-1870.
[29] HAARNOJA T, TANG H, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]//Proceedings of the 34th International Conference on Machine Learning(ICML). Sydney: IMLS, 2017: 1352-1361.
[30] MNIHL V, KAVUKCUOGLU K, SLIVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.
[31] FUJIMOTO S, HOOF H, MEGER M. Addressing function approximation error in actor-critic methods[C]//Proceedings of the 35th International Conference on Machine Learning(ICML). Stockholm: IMLS, 2018: 1587-1596.

作者简介

饶 宁 男，1997 年 8 月出生，江西上饶人. 现为空军工程大学信息与导航学院硕士研究生. 主要研究方向为通信对抗、强化学习.
E-mail: raoningmabma@163.com

许 华 男，1976 年 4 月出生，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.
E-mail: 13720720010@139.com

蒋 磊 男，1974 年 6 月出生，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

宋佰霖 男，1997 年 11 月出生，辽宁沈阳人. 现为空军工程大学信息与导航学院硕士研究生. 主要研究方向为通信对抗、强化学习.
E-mail: songbail@126.com

史蕴豪 男，1996 年 7 月出生，陕西咸阳人. 现为空军工程大学信息与导航学院博士研究生. 主要研究方向为信号识别、深度学习.
E-mail: shiyunhaoai@163.com