基于近端策略优化的智能抗干扰决策算法

马松[1,2]，李黎[3]，黎伟[2]，黄巍[2]，王军[2]
（1.中国西南电子技术研究所，四川 成都 610036；2.电子科技大学通信抗干扰全国重点实验室，四川 成都 611731；
3.中国西南电子设备研究所，四川 成都 610036）

摘　要：针对现有基于深度强化学习的智能抗干扰方法应用于天地测控通信链路时，用于决策的深度神经网络结构复杂，卫星等飞行器资源受限，难以在有限的复杂度约束下独立完成复杂神经网络的及时训练，抗干扰决策无法收敛的问题，提出了一种基于近端策略优化的智能抗干扰决策算法。分别在飞行器和地面站部署决策神经网络和训练神经网络，地面站根据飞行器反馈的经验信息进行最优化离线训练，辅助决策神经网络进行参数更新，在满足飞行器资源约束的同时实现有效的抗干扰策略选择。仿真结果表明，与基于策略梯度和基于深度Q学习的决策算法相比，所提算法收敛速度提升37%，收敛后的系统容量提升25%。

关键词：近端策略优化；深度强化学习；智能抗干扰；抗干扰决策
中图分类号：TN92
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024137

Intelligent anti-jamming decision algorithm based on proximal policy optimization

MA Song[1,2], LI Li[3], LI Wei[2], HUANG Wei[2], WANG Jun[2]
1. Southwest China Institute of Electronic Technology, Chengdu 610036, China
2. National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu 611731, China
3. Southwest China Research Institute of Electronic Equipment, Chengdu 610036, China

Abstract: The existing intelligent anti-jamming methods based on deep reinforcement learning are applied to space-ground TT&C and communication links, in which the deep neural network used for decision-making has a complex structure, and the resources of satellites and other vehicles are limited, making it difficult to independently complete the timely training of complex neural network under the constraints of limited complexity, and the decision-making of anti-jamming cannot converge. Aiming at the above problems, an intelligent anti-jamming decision algorithm based on proximal policy optimization was proposed, which deployed the decision-making neural network and the training neural network in the vehicles and the ground station, respectively. The ground station conducted the optimal offline training based on the empirical information feedback from the vehicles, and assisted the decision-making neural network in parameter updating, thereby achieving the effective selection of anti-jamming strategies while satisfying the resource constraints of the vehicles. The simulation results demonstrate that the convergence speed of the proposed algorithm is increased by 37%, and the system capacity after convergence is increased by 25%, compared with the decision algorithms of policy gradient and deep Q-learning.

Keywords: proximal policy optimization, deep reinforcement learning, intelligent anti-jamming, anti-jamming decision

收稿日期：2023–12–26；修回日期：2024–04–10
通信作者：马松，ma_song@139.com
基金项目：国家自然科学基金资助项目（No.62131005, No.62071096）
Foundation Items：The National Natural Science Foundation of China (No.62131005, No.62071096)

0　引言

随着全球“经济一体化”的迅速发展，人们对通信的传输速率和地域要求越来越高。在全球范围内实现高传输速率、全天候无缝通信已经成为未来移动通信的发展目标。天地一体化信息网络以地面网络为基础，以空间网络为延伸，为天、空、地、海等自然空间内的各类用户提供无缝覆盖与信息互通，具有重要的民用、商用以及军事价值[1-2]。随着通信和航天技术的发展，天地一体化信息网络迎来了重大发展机遇，同时也面临着巨大挑战[3-4]。

飞行器与地面站之间的天地测控通信链路是天地一体化信息网络的重要组成部分，天地测控通信链路正常工作是天地一体化信息网络能够正常运行并发挥效能的基本保障。天地测控通信链路高度暴露于远距离开放空间中，极易受到各种有意、无意干扰[5-6]。现有体制本身虽能依靠扩频增益和跳频增益获得一定的抗干扰能力[7]，但工作模式相对固定，在复杂多变的干扰环境中仍然面临着性能降低甚至功能丧失的潜在风险，因此有必要加强对天地测控通信链路抗干扰技术的应用研究，提高天地一体化信息网络系统的安全性和可靠性[8-9]。

近年来，随着人工智能的快速发展，以深度强化学习（DRL, deep reinforcement learning）为代表的机器学习技术在通信抗干扰领域得到了广泛应用[10]。在人工智能与认知无线电相结合的智能化抗干扰系统中，智能抗干扰决策是核心技术，可根据环境信息和信道质量在一个巨大的抗干扰策略空间中自适应地寻找接近最优的抗干扰策略，提升系统对干扰环境的适应能力[11]。文献[12]建立了Stackelberg博弈模型，引入强化学习算法提升了系统的抗干扰决策性能。文献[13]将干扰机与电台之间的对抗建模为马尔可夫决策过程，采用双深度Q网络（DQN, deep Q-network）算法进行抗干扰决策，改善了跟踪干扰下的抗干扰决策性能。文献[14-15]利用深度Q学习算法进行抗干扰决策，能够在不需要先验干扰信息和信道环境信息的条件下通过迭代做出最优抗干扰决策。文献[16]利用确定性策略梯度（PG, policy gradient）算法实现连续抗干扰决策，能够有效降低决策空间离散误差，进一步提高了抗干扰决策最优性能。

从已有研究来看，基于深度强化学习的抗干扰决策方法能够通过“试错”的方式在决策交互过程中学习信道信息和干扰源信息，并基于训练优化决策策略做出正确的抗干扰决策，取得良好的抗干扰决策性能。然而，深度强化学习的收敛复杂度取决于迭代次数和所使用的网络复杂度。在深度强化学习机制下，用于决策的深度神经网络结构相对复杂，且飞行器受运算能力、功率资源等限制，无法独立通过大量交互经验信息完成对复杂神经网络的训练，导致抗干扰决策无法收敛[17-18]。

针对以上问题，本文提出了一种基于近端策略优化（PPO, proximal policy optimization）的智能抗干扰决策算法。该算法分别在飞行器和地面站部署决策神经网络和训练神经网络，将飞行器的决策结果与其所处电磁环境中的交互信息结合构成经验信息，由下行测控通信链路反馈给地面站；地面站根据飞行器反馈的经验信息进行最优化离线训练；将训练结果关键参数通过高可靠控制链路传回飞行器实现决策神经网络参数更新，从而在显著降低飞行器处理资源和功耗开销的同时，提升了决策神经网络的收敛速度和决策性能。

1　系统模型与问题分析

1.1　系统模型

天地测控通信系统模型如图1所示。飞行器和地面站之间建立天地测控通信链路，周围分布着各类干扰源，释放各种干扰信号，形成复杂多变的干扰环境，影响飞行器对上行测控通信链路的正常接收。

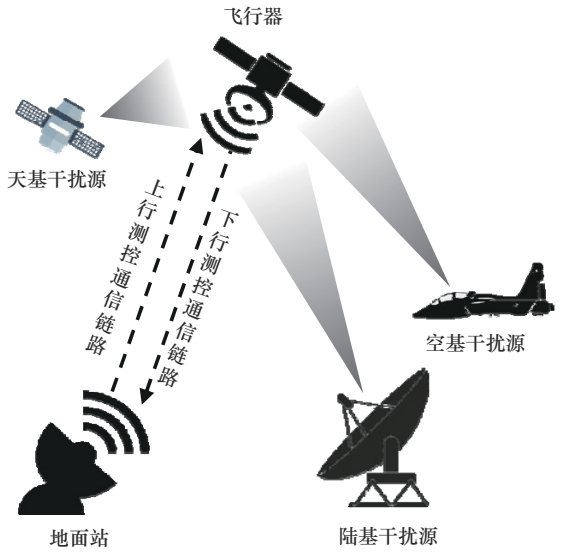


图1　天地测控通信系统模型

本文将各类干扰源构成的复杂多变的干扰环境建模为随机干扰、扫频干扰和反应式干扰结合的混合干扰。随机干扰以固定的发射功率在每个时隙随机选择子信道。扫频干扰以固定的发射功率循环扫过若干个子信道。反应式干扰按照一定规律对目标信号可能使用的子信道进行监听，并以固定的发射功率对发现的目标信号信道进行干扰[19-20]。

如文献[21]所述，干扰源与天地测控通信链路的持续干扰对抗可以建模为“零和”博弈过程。干扰源通过选择合适的子信道和功率释放干扰信号。上行测控通信链路以扩频和跳频方式实现基础抗干扰能力，采用基于深度强化学习的抗干扰决策方法优选工作频率、调制编码方式、扩频倍数和功率，提升复杂多变的干扰环境下的传输性能。

飞行器观察环境状态，根据抗干扰决策得到行为，然后将该行为作用于环境获得相应反馈。根据获得的反馈计算当前行为的奖励和转移状态，将该过程获得的环境状态−行为−奖励−转移状态作为经验信息存储，该经验信息作为决策神经网络的训练样本用于策略更新。该过程可以被描述为马尔可夫决策过程，数学表达式为$(\mathcal{S},\mathcal{A},P,R,\gamma)$，其中，$\mathcal{S}$表示环境状态空间，$\mathcal{A}$表示决策行为空间，$P$表示状态转移概率，$R(s,a)\in\mathbb{R}$表示回报函数，$\gamma\in(0,1)$表示累积回报中的折扣因子。

1.2　问题分析

本节以传统的基于策略梯度的深度强化学习为例，对深度强化学习抗干扰决策方法的更新过程进行分析[16]。

在时刻t，智能体通过观察环境获得环境状态$s_t\in\mathcal{S}$，然后在环境状态s_t下根据当前策略$\pi(\cdot)$进行行为选择$a_t\leftarrow\pi(s_t)$。在统计策略和确定策略不同情况下，策略选择输出分别对应每个行为的概率$\pi(s,a)\in(0,1)$或者某个确定的行为$a=\pi(s)$。$R(s_t,a_t)$度量在环境状态s_t下采取行为a_t所能获得的立即回报值r_t。智能体选择行为a_t后作用于环境，环境状态以概率$P(a_t)$转移到状态$s_{t+1}\in\mathcal{S}$。根据该过程可以总结出一个回合内T个时隙的学习轨迹，表示为

$$\tau=\{s_1,a_1,r_1,s_2,a_2,r_2,\cdots,s_T,a_T,r_T,s_{T+1}\} \tag{1}$$

在该学习轨迹下，累积回报函数可表示为

$$R(\tau)=\sum_{t=1}^{T}r_t \tag{2}$$

由于决策轨迹取决于行为选择策略，用深度神经网络把决策策略表示为依赖于参数θ的非线性函数，则平均累积回报函数可以表示为

$$\bar{R}_\theta=\sum_{\tau}R(\tau)(\tau|\theta)\approx\frac{1}{N}\sum_{n=1}^{N}R(\tau^n) \tag{3}$$

其中，τ^n为第n个回合的学习轨迹，时隙为t的学习轨迹表示为$\tau_t^n=(s_t^n,a_t^n,r_t^n,s_{t+1}^n)$，第$n$个回合共包含$T_n$个时隙。

参数更新可基于求解以下优化问题实现。

$$\theta^*=\arg\max_{\theta}\bar{R}_\theta \tag{4}$$

通常，可以采用梯度上升的方法通过迭代实现参数更新。

$$\theta^i\leftarrow\theta^{i-1}+\eta\nabla\bar{R}_{\theta^{i-1}} \tag{5}$$

其中，η为学习率。

行为选择策略同样通过合理构造的深度神经网络实现，此时深度神经网络参数θ对应表示为

$$\theta=\{w_1,w_2,\cdots;b_1,b_2,\cdots\} \tag{6}$$

其中，w_l表示深度神经网络的第l层权重，b_l表示深度神经网络的第l层偏置值。根据式(1)中的学习轨迹表达式，将$P(\tau|\theta)$根据决策时序逐级展开，可表示为

$$P(\tau|\theta)=p(s_1)\prod_{n=1}^{N}p(a_n|s_n,\theta)\,p(r_1,s_{n+1}|s_n,a_n) \tag{7}$$

根据式(3)和式(7)可以推导出平均累积回报梯度，表示为

$$\nabla\bar{R}_\theta=\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n}R(\tau^n)\nabla\log p(a_t^n|s_t^n,\theta) \tag{8}$$

因此，将式(8)代入式(5)可以实现深度强化学习决策模型中深度神经网络训练和网络参数更新。

综上所述，基于策略梯度的深度强化学习抗干扰决策过程如算法 1 所示[16]。

算法 1　基于策略梯度的深度强化学习抗干扰决策过程

1) 初始化策略网络$\pi_\theta(\cdot)$、容量为M的经验池、迭代回合数E、每回合迭代步数T、批处理参数B；
2) 当回合数满足$e=1,\cdots,E$时，执行迭代；
3) 随机初始化经验池中的经验组；
4) 当迭代步数满足$t=1,\cdots,T$时，执行迭代；
5) 观察状态s_t，根据概率$\arg\max_a\pi_\theta(s_t,a)$选择当前抗干扰行为$a_t$，否则随机进行抗干扰行

为选择；

6) 执行a_t，观察立即回报值r_t和转移状态s_{t+1}；

7) 构造经验组$<s_t,a_t,r_t,s_{t+1}>$并存储在经验池中；

8) 随机从经验池中抽取B数量的经验组，根据式(3)和式(5)完成参数θ的训练；

9) 停止迭代；

10) 停止迭代。

算法1所示决策过程能够通过“观测−判断−决策−行动”的迭代方式学习环境中的干扰信息，并且通过优化策略网络做出最优抗干扰决策，对复杂多变的干扰环境有很好的适应性。文献[22]表明，当迭代学习次数满足要求后，行为能够无限逼近最优行为，即$a_t \to a^*$。

在智能抗干扰决策算法的实际应用中，算法的复杂度取决于收敛迭代次数和每次迭代训练时的运算复杂度[23]。为了适应动态变化的干扰环境，基于深度强化学习的智能抗干扰决策算法需要采用复杂的神经网络结构，并通过不断迭代优化网络性能，带来了较高的迭代训练复杂度代价和迭代次数代价。然而，由于体积、功耗和处理资源受限的飞行器很难满足深度强化学习的实时性要求，智能抗干扰决策算法难以得到广泛应用。

2　算法设计

2.1　地面站辅助训练的抗干扰决策模型

为了降低对飞行器的资源需求，本文设计了地面站辅助训练的抗干扰决策模型，如图2所示。在飞行器中合理构建一个深度神经网络来执行决策，称之为决策神经网络（网络参数表示为θ），仅进行抗干扰决策执行，不进行网络训练。在地面站构建一个与决策神经网络结构相同的深度神经网络用来执行最优化离线训练，称之为训练神经网络（网络参数表示为θ'）。通过地面站辅助飞行器进行训练，天地联合实现上行测控通信链路智能抗干扰决策。

在一个学习轨迹中，飞行器通过感知得到环境状态s_t，决策神经网络决策出抗干扰行为a_t，并作用于环境，然后计算立即回报值r_t，并和转移后的环境状态s_{t+1}组成经验组$<s_t,a_t,r_t,s_{t+1}>$，由下行测控通信链路反馈给地面站。地面站的训练神经网络基于飞行器反馈的经验信息进行最优化离线训练，并把训练得到的网络参数通过高可靠控制链路上注给飞行器实现决策神经网络参数更新。环境状态、行为空间和回报函数的设计如下。

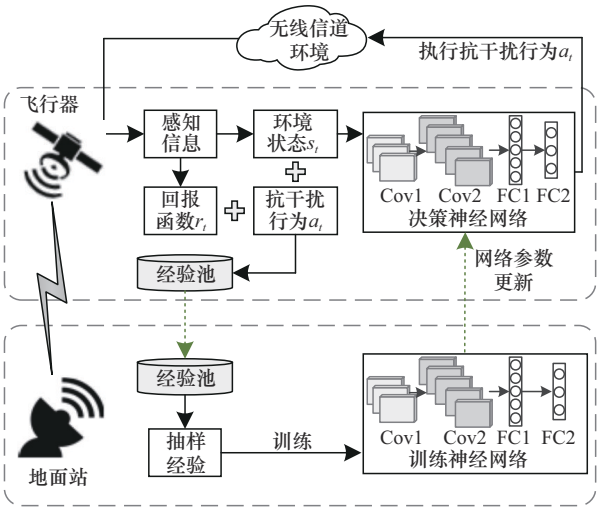


图2　地面站辅助训练的抗干扰决策模型

1) 环境状态

时刻t的环境状态s_t由全频谱功率$P^{(t)}$、干扰类型$J_k^{(t)}$、干扰带宽$J_b^{(t)}$、链路可达传输速率$R^{(t)}$、信干噪比（SINR, signal to interference plus noise ratio）$\mathrm{SINR}^{(t)}$、链路锁定状态$C^{(t)}$和链路误码率（BER, bit error ratio）$\mathrm{BER}^{(t)}$共同构造，表示为

$$s_t = \left\{P^{(t)},J_k^{(t)},J_b^{(t)},R^{(t)},\mathrm{SINR}^{(t)},C^{(t)},\mathrm{BER}^{(t)}\right\} \tag{9}$$

2) 行为空间

上行测控通信链路通过选择合适的子信道$f^{(t)}$、调制编码方式$m^{(t)}$、扩频倍数$b_{\mathrm{ds}}{}^{(t)}$和发射功率$p_f{}^{(t)}$实现抗干扰决策，时刻t的行为空间a_t表示为

$$a_t = \{f^{(t)},m^{(t)},b_{\mathrm{ds}}{}^{(t)},p_f{}^{(t)}\} \tag{10}$$

其中，$1 \leqslant f^{(t)} \leqslant N_{\mathrm{ch}}$，$1 \leqslant m^{(t)} \leqslant M_{\mathrm{mcs}}$，$1 \leqslant b_{\mathrm{ds}}{}^{(t)} \leqslant K_{\mathrm{ds}}$，$0 \leqslant p_f{}^{(t)} \leqslant P_{\max}$，$N_{\mathrm{ch}}$表示链路支持的子信道数量，$M_{\mathrm{mcs}}$表示链路支持的调制编码方式组合数量，$K_{\mathrm{ds}}$表示最大扩频倍数，$P_{\max}$表示最大发射功率。

3) 回报函数

在环境状态s_t下采取行为a_t所能获得的立即回报值r_t表示为

$$\begin{aligned} r_t = {} & B_0\mathrm{lb}\left(1+\mathrm{SINR}^{(t)}\right) - C_p p_f^{(t)} - \\ & C_h\mathcal{F}\left(\left(f^{(t)} \neq f^{(t-1)}\right)\middle|\left(m^{(t)} \neq m^{(t-1)}\right)\middle|\left(b_{\mathrm{ds}}{}^{(t)} \neq b_{\mathrm{ds}}{}^{(t-1)}\right)\right) \end{aligned} \tag{11}$$

其中，等式右边第一项表示链路容量，第二项表示链路的功率代价，第三项表示链路为了对抗干扰而

改变频率、调制编码方式和扩频倍数引入的链路重建代价，B_0表示子信道带宽，$\mathrm{SINR}^{(t)}$表示时刻t的链路信干噪比，C_p表示单位功率发射代价系数，C_h表示链路重建代价系数，$\mathcal{F}(\phi)$表示指示函数，当ϕ成立时函数值为1，否则为0，“|”表示“或”函数。

假设链路信道增益和干扰信道增益分别为$h_f^{(t)}$和$h_{j,f}^{(t)}$，干扰功率为P_j，噪声功率为σ^2，干扰信道为$f_j^{(t)}$，则链路信干噪比$\mathrm{SINR}^{(t)}$可以表示为

$$\mathrm{SINR}^{(t)}=\frac{p_f{}^{(t)}h_f^{(t)}}{\sigma^2+\sum\limits_{j=1}^{J}P_jh_{j,f}^{(t)}\mathcal{F}\left(f^{(t)}=f_j^{(t)}\right)} \tag{12}$$

2.2　近端策略优化算法

根据2.1节中所给出的推导，深度神经网络训练所需要的平均累积回报梯度可表示为

$$\begin{aligned}&\nabla\bar{R}_\theta=\mathbb{E}_{\tau;\,p_\theta(\cdot)}\left[R(\tau)\nabla\log p_\theta(\tau)\right]\approx\\&\frac{1}{N}\sum_{n=1}^{N}R\left(\tau^n\right)\nabla\log\left(p_\theta\left(\tau^n\right)\right)=\\&\mathbb{E}_{\left(s_t,a_t\right)\sim p_\theta(\cdot)}\left[A^\theta\left(s_t,a_t\right)\nabla\log\left(p_\theta\left(a_t^n|s_t^n\right)\right)\right]\end{aligned} \tag{13}$$

其中，$p_\theta(\cdot)$表示在当前神经网络参数θ影响下的决策轨迹过程分布。A^θ用来评估当前时刻状态动作对$\left(s_t,a_t\right)$的优劣，可表示为

$$A^\theta\left(s_t,a_t\right)=\sum_{t'=t}^{T_n}\lambda^{t'-t}r_{t'}^n-b_{A(\cdot)} \tag{14}$$

其中，$b_{A(\cdot)}$为预设门限，λ为折扣因子。

在离线训练情况下，地面站将训练神经网络的训练结果通过上行测控通信链路传输给飞行器，才能用于决策神经网络参数更新。训练神经网络与决策神经网络不同步，导致训练神经网络的轨迹分布$q_{\theta'}(\cdot)$与决策神经网络的轨迹分布$p_\theta(\cdot)$之间存在偏差。为了描述这种偏差，引入如下重要性采样指标。

$$\begin{aligned}&\mathbb{E}_{x;\,p_\theta(\cdot)}\left[f(x)\right]=\int f(x)\,p(x)\mathrm{d}x=\\&\int f(x)\frac{p(x)}{q(x)}q(x)\mathrm{d}x=\mathbb{E}_{x;\,q_\theta(\cdot)}\left[f(x)\,\frac{p_\theta(x)}{q_{\theta'}(x)}\right]\end{aligned} \tag{15}$$

将式(15)代入式(13)可以得到

$$\begin{aligned}&\nabla\bar{R}_\theta=\mathbb{E}_{\left(s_t,a_t\right)\sim q_\theta(\cdot)}\\&\left[\frac{p_\theta\left(a_t|s_t\right)}{q_{\theta'}\left(a_t|s_t\right)}\frac{p_\theta\left(s_t\right)}{q_\theta\left(s_t\right)}\cdot A^\theta\left(s_t,a_t\right)\nabla\log\left(p_\theta\left(a_t^n|s_t^n\right)\right)\right]\end{aligned} \tag{16}$$

由于$p_\theta\left(s_t\right)$和$q_\theta\left(s_t\right)$未知，真实环境一般不会随着决策的改变而改变。因此，通常假设$p_\theta\left(s_t\right)$和$q_\theta\left(s_t\right)$接近，即$\dfrac{p_\theta\left(s_t\right)}{q_\theta\left(s_t\right)}=1$。利用复合函数求导和对数函数求导，引入等式

$$\nabla f(x)=f(x)\nabla\log\left(f(x)\right) \tag{17}$$

式(16)可以简化为

$$\nabla\bar{R}_\theta\leftarrow J^\theta(\theta)=\mathbb{E}_{\left(s_t,a_t\right)\sim q_\theta(\cdot)}\left[\frac{p_\theta\left(a_t|s_t\right)}{q_{\theta'}\left(a_t|s_t\right)}A^\theta\left(s_t,a_t\right)\right] \tag{18}$$

然而，以上表达式对原$\nabla\bar{R}_\theta$的估计是有偏差的，即当$p_\theta\left(a_t|s_t\right)$与$q_\theta\left(a_t|s_t\right)$偏差很大时，$\nabla\bar{R}_\theta$与$J^\theta(\theta)$的方差并不相等，式(18)的赋值操作$\nabla\bar{R}_\theta\leftarrow J^\theta(\theta)$会引入误差。为了克服这种误差影响，可以引入散度KL克服，具体表达式如式(19)所示。

$$\mathrm{KL}\left(p,q\right)=\sum_x p(x)\log\left[\frac{p(x)}{q(x)}\right] \tag{19}$$

因此，式(18)可表示为

$$\begin{aligned}&\nabla\bar{R}_\theta\leftarrow J_{\mathrm{PPO}}^\theta(\theta)=\\&J^\theta(\theta)-\beta\mathrm{KL}\left(p_\theta,q_{\theta'}\right)=\\&\mathbb{E}_{\left(s_t,a_t\right)\sim q_\theta(\cdot)}\left[\frac{p_\theta\left(a_t|s_t\right)}{q_{\theta'}\left(a_t|s_t\right)}A^\theta\left(s_t,a_t\right)\right]-\\&\beta\sum_s\sum_a p_\theta\left(a|s\right)\log\left[\frac{p_\theta(a|s)}{q_{\theta'}(a|s)}\right]\end{aligned} \tag{20}$$

其中，β是散度因子，$\beta\in(0,1)$。

2.3　智能抗干扰决策算法

通过以上分析，基于近端策略优化的智能抗干扰决策算法由地面站和飞行器共同完成。飞行器主要执行在线决策，地面站主要执行离线训练，执行过程分别如算法2和算法3所示。

算法2　抗干扰决策飞行器执行算法

1) 初始化决策神经网络$\pi_\theta(\cdot)$、容量为M的经验池、每回合迭代步数T；
2) 当步数满足$t=1,\cdots,T$时，执行迭代；
3) 观察环境状态s_t，根据决策神经网络$\pi_\theta(\cdot)$选择当前抗干扰行为$a_t\leftarrow\pi_\theta(s_t)$；
4) 执行a_t，观察立即回报值r_t和转移状态s_{t+1}；
5) 构造经验组$\tau_t^e=\left\{s_t,a_t,r_t,s_{t+1}\right\}$，并存储在飞行器经验池中；

6) 将更新的经验组τ_t^e通过下行测控通信链路传给地面站；

7) 停止迭代；

8) 飞行器接收地面站上注的网络参数，执行决策神经网络参数更新：$\theta \leftarrow \theta'$。

算法 3 抗干扰决策地面站执行算法

1) 初始化训练神经网络$\pi_\theta(\cdot)$、容量为M的经验池、批处理参数B、散度因子β；

2) 地面站接收飞行器传来的经验组τ_t^e，存入经验池中；

3) 当批处理参数$b \geqslant B$时，执行迭代；

4) 根据式(14)计算增量值A^θ；

5) 从经验池中随机抽取B数量的经验组，根据式(20)计算$J_{\mathrm{PPO}}^{\theta}(\theta)$，并执行$\nabla \bar{R}_\theta \leftarrow J_{\mathrm{PPO}}^{\theta}(\theta)$；

6) 根据式(5)进行训练神经网络参数更新；

7) 停止迭代；

8) 地面站将训练神经网络参数更新的参数上注给飞行器。

飞行器和地面站联合实现基于近端策略优化的智能抗干扰决策算法，如算法 4 所示。

算法 4 基于近端策略优化的智能抗干扰决策算法

1) 初始化决策神经网络$\pi_\theta(\cdot)$、训练神经网络$\pi_{\theta'}(\cdot)$、容量为M的经验池、迭代回合数E、每回合迭代步数T、批处理参数B、散度因子β；

2) 当回合数满足$e = 1,\cdots,E$时，执行迭代；

3) 当步数满足$t = 1,\cdots,T$时，执行迭代；

4) 观察状态s_t，根据决策神经网络$\pi_\theta(\cdot)$选择当前抗干扰行为$a_t \leftarrow \pi_\theta(s_t)$；

5) 执行a_t，观察立即回报值r_t和转移状态s_{t+1}；

6) 构造经验组$\tau_t^e = \{s_t, a_t, r_t, s_{t+1}\}$，并存储在飞行器经验池中；

7) 将更新的经验组τ_t^e通过下行测控通信链路传到地面站，存入经验池中；

8) 停止迭代；

9) 当批处理参数$b \geqslant B$时，执行迭代；

10) 根据式(14)计算每步增量值A^θ；

11) 从经验池中随机抽取B数量的经验组，根据式(20)计算$J_{\mathrm{PPO}}^{\theta}(\theta)$，并执行$\nabla \bar{R}_\theta \leftarrow J_{\mathrm{PPO}}^{\theta}(\theta)$；

12) 根据式(5)进行训练神经网络参数更新；

13) 停止迭代；

14) 地面站将训练神经网络更新的参数上注给飞行器；

15) 飞行器执行决策神经网络参数更新：$\theta \leftarrow \theta'$；

16) 停止迭代。

3 仿真验证

为了验证本文所提基于近端策略优化的智能抗干扰决策算法，本节对其性能进行仿真验证。

仿真中将本文算法分别与算法 1 所示的基于策略梯度的深度强化学习抗干扰决策算法[16]（简称为基于 PG 的决策算法）、基于深度 Q 学习的抗干扰决策算法[14-15]（简称为基于 DQN 的决策算法）、随机决策算法以及传统跳频模式进行性能对比。其中，基于 PG 和 DQN 的决策算法采用地面站辅助的离线训练模式，地面站的训练神经网络采用飞行器下传的延迟经验组进行训练。而传统跳频模式采用固定的跳频图案，选择固定的调制编码方式，发射功率恒定为最大发射功率$P_{\max}$。

考虑到公平性，仿真实验中基于学习的算法采用的网络结构如表 1 所示，且经验池大小$M = 1\,024$，批处理参数$B = 32$均保持一致。

表 1 深度神经网络结构

网络层数	卷积层 1	卷积层 2	池化层	全连接层
输入维度	11×11	8×7×7	10×7×7	160
滤波结构	3×3	3×3	2×2	—
步长	2	1	2	—
补零个数	2	1	1	—
滤波器数	8	10	10	—
激活函数	ReLU	ReLU	—	Softmax
输出维度	8×7×7	10×7×7	10×4×4	6

以某地面站与某同步轨道卫星之间的上行测控通信链路为例，地面站最大发射功率$P_{\max} = 60$ dBm，天线增益为 30 dBi，辐射出去的上行信号最大功率可达 90 dBm。同步轨道卫星距离地面站约 35 786 km，工作频段为 2.4~2.5 GHz，子信道数$N_{\mathrm{ch}} = 10$，子信道带宽$B_0 = 10$ MHz。

干扰源部署在某高轨卫星平台上，可接近目标，释放随机干扰、扫频干扰和反应式干扰，同时对目标进行干扰。随机干扰在每个时隙随机选择信

道和干扰瞬时带宽。扫频干扰采用与子信道带宽 B_0 相等的瞬时带宽，循环扫过系统所有子信道。反应式干扰对系统所有信道进行监听，当在监听范围内发现目标信号信道时，以固定功率对其进行干扰[19-20]。干扰源最大发射功率为 50 dBm，天线增益为 5 dBi，干扰距离最小为 20 km[24]。

仿真实验中其他参数设置为：单位功率发射代价系数 $C_p = 1$，链路重建代价系数 $C_h = 10$，噪声功率 $\sigma^2 = -120$ dBm，散度因子 $\beta = 0.2$。

图 3 给出了本文算法、基于 PG 的决策算法和基于 DQN 的决策算法的训练曲线，纵轴以式(11)回报值作为归一化系统效能。从图 3 可以看到，本文算法在约 100 个回合时逐渐收敛到最优归一化系统效能，收敛性能良好。而其他 2 种决策算法在 160 个回合之后才开始收敛，且性能波动较大。本文算法收敛速度较这 2 种对比算法提升了 37%，这 2 种对比算法收敛性能不理想，原因在于训练过程中使用的训练样本是延迟经验组，无法准确地反映当前的环境变化，也无法有效地应对动态干扰环境。

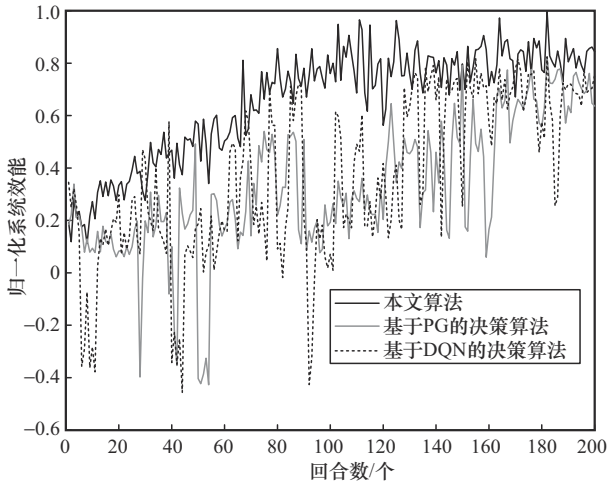


图 3　不同智能抗干扰决策算法的训练曲线

图 4 给出了不同算法的平均系统容量随回合数变化的关系曲线。从图 4 可以看到，本文算法收敛后的平均系统容量比基于 PG 的决策算法和基于 DQN 的决策算法提升约 25%，说明本文算法能够更好地应对动态的干扰环境。传统跳频模式和随机决策算法的性能明显劣于上述 3 种智能决策算法，这是因为这 2 种算法没有对干扰情况做针对性决策，无法有效对抗干扰。

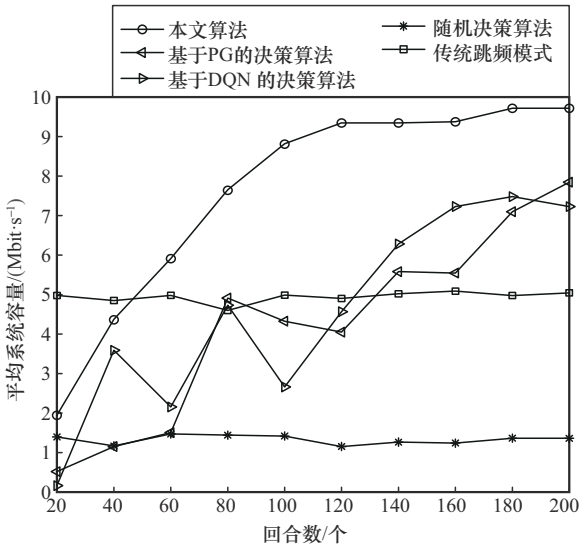


图 4　不同算法的平均系统容量随回合数变化的关系曲线

图 5 给出了不同算法的平均发射功率随回合数变化的关系曲线。从图 5 可以看到，本文算法在收敛之后的平均发射功率虽然略高于其他 2 种决策算法，但相较于系统容量的提升，增加的发射功率几乎可以忽略不计。同时，本文算法的平均发射功率距离最大发射功率仍有一定差距，表明本文算法并非简单地通过增大发射功率来对抗干扰，而是充分利用可用的维度有针对性地对抗干扰。

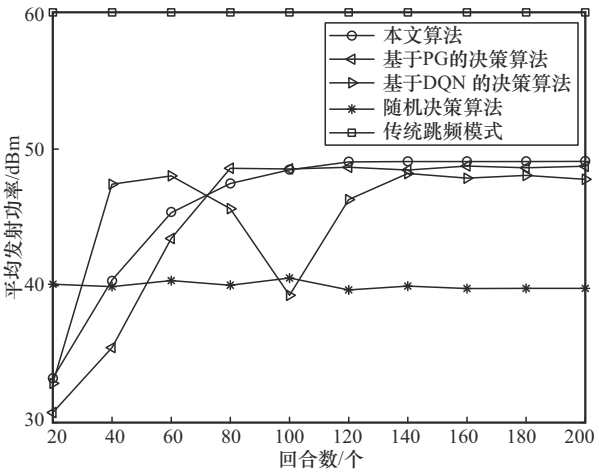


图 5　不同算法的平均发射功率随回合数变化的关系曲线

图 6 给出了不同算法的归一化跳频次数（单位时间内的跳频次数，以传统跳频模式对应数值作为 1）随回合数变化的关系曲线。从图 6 可以明显地看到，随机决策算法和传统跳频模式由于没有关于干扰的信息，只能盲目地通过跳频来规避干扰，其归一化跳频次数约为 0.9 和 1.0。当回合大于 120 个

时，本文算法的归一化跳频次数远低于其他几种对比算法，其原因是本文算法通过与环境的交互学习，通过长期回报最大化，选择当前一段时间内最优的频点，能够在未来一段时间内规避干扰或者维持通信质量。

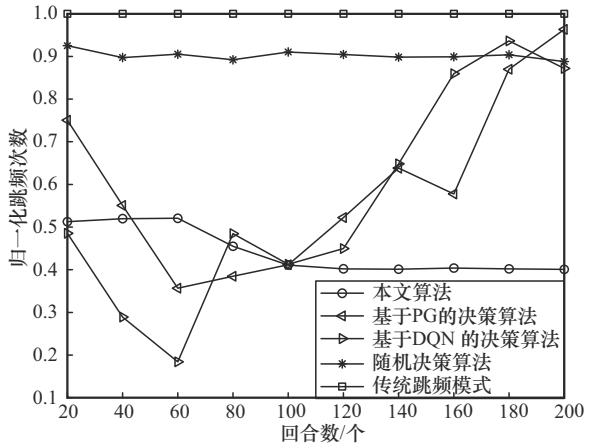


图6　不同算法的归一化跳频次数随回合数变化的关系曲线

4　结束语

在天地一体化信息网络的动态干扰场景中，飞行器运算与功率资源有限，无法独立完成复杂神经网络训练，影响深度强化学习智能抗干扰决策性能。本文提出了基于近端策略优化的智能抗干扰决策算法，分别在飞行器和地面站部署决策神经网络和训练神经网络，地面站根据飞行器反馈的经验信息进行最优化离线训练，辅助决策神经网络进行参数更新，在满足飞行器资源约束的同时实现有效的抗干扰决策。仿真结果表明，与基于策略梯度和深度Q学习的决策算法相比，本文算法收敛速度提升37%，收敛后的平均系统容量提升25%。

参考文献：

[1] 黄韬, 刘江, 汪硕, 等. 未来网络技术与发展趋势综述[J]. 通信学报, 2021, 42(1): 130-150.
HUANG T, LIU J, WANG S, et al. Survey of the future network technology and trend[J]. Journal on Communications, 2021, 42(1): 130-150.

[2] 刘天华, 王洪全. 天地一体化信息网络在我国民航领域的应用设想[J]. 电讯技术, 2018, 58(6): 738-744.
LIU T H, WANG H Q. Thought on application of space-ground integrated information network in domestic civil aviation[J]. Telecommunication Engineering, 2018, 58(6): 738-744.

[3] NIEPHAUS C, KRETSCHMER M, GHINEA G. QoS provisioning in converged satellite and terrestrial networks: a survey of the state-of-the-art[J]. IEEE Communications Surveys & Tutorials, 2016, 18(4): 2415-2441.

[4] 张海君, 陈安琪, 李亚博, 等. 6G移动网络关键技术[J]. 通信学报, 2022, 43(7): 189-202.
ZHANG H J, CHEN A Q, LI Y B, et al. Key technologies of 6G mobile network[J]. Journal on Communications, 2022, 43(7): 189-202.

[5] GUIDOTTI A, VANELLI-CORALLI A, CONTI M, et al. Architectures and key technical challenges for 5G systems incorporating satellites[J]. IEEE Transactions on Vehicular Technology, 2019, 68(3): 2624-2639.

[6] 张玲翠, 许瑶冰, 李凤华, 等. 天地一体化信息网络安全动态赋能架构[J]. 通信学报, 2021, 42(9): 87-95.
ZHANG L C, XU Y B, LI F H, et al. Dynamic security-empowering architecture for space-ground integration information network[J]. Journal on Communications, 2021, 42(9): 87-95.

[7] 朱勇刚, 孙艺夫, 姚富强, 等. 基于多智能超表面的信道空间内生抗干扰方法[J]. 通信学报, 2023, 44(10): 13-22.
ZHU Y G, SUN Y F, YAO F Q, et al. Channel-space endogenous anti-jamming method based on multi-reconfigurable intelligent surface[J]. Journal on Communications, 2023, 44(10): 13-22.

[8] BRYAN C, MARK G, JESSE S. Winning in the gray zone: using electromagnetic warfare to regain escalation dominance[R]. 2017.

[9] YAO H P, WANG L Y, WANG X D, et al. The space-terrestrial integrated network: an overview[J]. IEEE Communications Magazine, 2018, 56(9): 178-185.

[10] 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.

[11] 李少谦, 程郁凡, 董彬虹, 等. 智能抗干扰通信技术研究[J]. 无线电通信技术, 2012, 38(1): 1-4.
LI S Q, CHENG Y F, DONG B H, et al. Research on intelligent anti-jam communication techniques[J]. Radio Communications Technology, 2012, 38(1): 1-4.

[12] 张孟杰, 赵睿, 王培臣, 等. 基于强化学习的无人机辅助物联网抗敌意干扰算法[J]. 信号处理, 2021, 37(1): 11-18.
ZHANG M J, ZHAO R, WANG P C, et al. Anti-jamming algorithm with reinforcement learning in UAV-aided Internet of things[J]. Journal of Signal Processing, 2021, 37(1): 11-18.

[13] 王瑞东, 张彦龙, 魏鹏, 等. 战术跳频系统智能抗干扰决策[J]. 信号处理, 2023, 39(1): 84-95.
WANG R D, ZHANG Y L, WEI P, et al. Intelligent anti-jamming strategy for tactical frequency-hopping system[J]. Journal of Signal Processing, 2023, 39(1): 84-95.

[14] HAN G A, XIAO L, POOR H V. Two-dimensional anti-jamming communication based on deep reinforcement learning[C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 2087-2091.

[15] 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.

[16] LI W, WANG J, LI L, et al. Intelligent anti-jamming communication with continuous action decision for ultra-dense network[C]//Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2019: 1-7.

[17] 张梦钰, 豆亚杰, 陈子夷, 等. 深度强化学习及其在军事领域中的应用综述[J]. 系统工程与电子技术, 2024, 46(4): 1297-1308.
ZHANG M Y, DOU Y J, CHEN Z Y, et al. Review of deep reinforcement learning and its applications in military field[J]. Systems Engineering and Electronics, 2024, 46(4): 1297-1308.

[18] 唐斯琪, 潘志松, 胡谷雨, 等. 深度强化学习在天基信息网络中的应用: 现状与前景[J]. 系统工程与电子技术, 2023, 45(3): 886-901.
TANG S Q, PAN Z S, HU G Y, et al. Application of deep reinforcement learning in space information network—status quo and prospects[J]. Systems Engineering and Electronics, 2023, 45(3): 886-901.

[19] STRASSER M, PÖPPER C, ČAPKUN S. Efficient uncoordinated FHSS anti-jamming communication[C]//Proceedings of the Tenth ACM International Symposium on Mobile Ad Hoc Networking and Computing. New York: ACM Press, 2009: 207-218.

[20] WILHELM M, MARTINOVIC I, SCHMITT J B, et al. Short paper: reactive jamming in wireless networks: how realistic is the threat? [C]// Proceedings of the Fourth ACM Conference on Wireless Network Security. New York: ACM Press, 2011: 47-52.

[21] HE X F, DAI H Y, NING P. Faster learning and adaptation in security games by exploiting information asymmetry[J]. IEEE Transactions on Signal Processing, 2016, 64(13): 3429-3443.

[22] SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. Cambridge: MIT Press, 2018.

[23] HE K M, SUN J. Convolutional neural networks at constrained time cost[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 5353-5360.

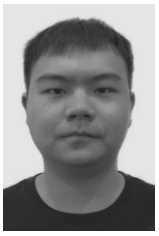
[24] 邹润, 刘阳, 臧晴, 等. 国外天基空间目标监视系统发展综述[J]. 航天器工程, 2023, 32(5): 110-118.
ZOU R, LIU Y, ZANG Q, et al. Overview of development of foreign space-based space target surveillance system[J]. Spacecraft Engineering, 2023, 32(5): 110-118.

[作者简介]

马松（1990– ），男，四川巴中人，中国西南电子技术研究所高级工程师，主要研究方向为飞行器测控通信、人工智能。

李黎（1994– ），男，重庆人，博士，中国西南电子设备研究所工程师，主要研究方向为无线与移动通信、人工智能。

黎伟（1988– ），男，四川广安人，博士，电子科技大学在站博士后，主要研究方向为无线与移动通信、人工智能。



黄巍（1995– ），男，四川达州人，电子科技大学博士生，主要研究方向为无线与移动通信、机器学习。

王军（1974– ），男，四川蓬溪人，博士，电子科技大学教授，主要研究方向为无线与移动通信。