基于 ET-PPO 的双变跳频图案智能决策

陈一波，赵知劲

（杭州电子科技大学通信工程学院，浙江 杭州 310018）

摘　要：为进一步提高双变跳频系统在复杂电磁环境中的抗干扰能力，提出了一种基于资格迹的近端策略优化（proximal policy optimization with eligibility traces，ET-PPO）算法。在传统跳频图案的基础上，引入时变参数，通过状态－动作－奖励三元组的构造将“双变”跳频图案决策问题建模为马尔可夫决策问题。针对 PPO 算法“行动器”网络样本更新方式的高方差问题，引入加权重要性采样减小方差；采用 Beta 分布的动作选择策略，增强学习阶段的稳定性。针对“评判器”网络收敛速度慢的问题，引入资格迹方法，较好地平衡了收敛速度和全局最优解求解。在不同电磁干扰环境下的算法对比仿真结果表明，ET-PPO 有更好的适应性和稳定性，对抗阻塞干扰和扫频干扰表现较好。

关键词：复杂电磁环境；双变跳频图案；近端策略优化；资格迹

中图分类号：TN914；TP181

文献标志码：A

doi: 10.11959/j.issn.1000–0801.2022264

Intelligent anti-jamming decision algorithm of bivariate frequency hopping pattern based on ET-PPO

CHEN Yibo, ZHAO Zhijin

School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In order to further improve its anti-interference ability in complex electromagnetic environment, a PPO algorithm based on weighted importance sampling and eligibility traces (ET-PPO) was proposed. On the basis of the traditional frequency hopping pattern, time-varying parameters were introduced, and the bivariate frequency hopping pattern decision problem was modeled as a Markov decision problem through the construction of the state-action-reward triple. Aiming at the high variance problem of the sample update method of an actor network of the PPO algorithm, weighted importance sampling was introduced to reduce the variance, and the action selection strategy of Beta distribution was used to enhance the stability of the learning stage. Aiming at the problem of slow convergence speed of the evaluator network, the eligibility trace method was introduced, which better balanced the convergence speed and the global optimal solution. The algorithm comparison simulation results in different electromagnetic interference environments show that ET-PPO has better adaptability and stability, and has better performance against obstruction interference and sweep frequency interference.

Key words: complex electromagnetic environment, bivariate frequency hopping pattern, proximal policy optimization, eligibility trace

收稿日期：2022−06−02；**修回日期**：2022−09−29

基金项目：国家自然科学基金资助项目（No.U19B2016）

Foundation Item: The National Natural Science Foundation of China (No.U19B2016)

0　引言

相比于常规通信系统，跳频通信系统有跳频速度、信道划分间隔和跳频序列等参数，根据不同通信环境合理配置这些参数可以提升跳频通信系统的抗干扰性能[1]。传统跳频通信系统预先设定跳频速度和信道划分间隔，在电子战中容易被敌方侦破并窃取信息。变间隔信道划分与变速率信号驻留时间的“双变”思想的引入给予跳频系统二维伪随机变化特性，相比于传统跳频通信系统的一维随机变化特性，提升了跳频系统频带选择的灵活性[2-3]，大大增加了敌方破译跳频参数的难度，因此引起了学者们的关注。文献[4-6]从增加复杂性角度设计、分析了可变跳频速度和信道划分间隔的跳频图案。文献[4]根据带均匀性补偿的频点分布函数，将随机跳频序列映射到指定零散频带上，由此得到的通信系统跳频图案有较好的频谱均匀性；文献[5]提出非线性模 d 法，在 m 序列的基础上，对生成的跳频序列宽间隔化，构造了性能优良的跳频图案；文献[6]采用跳频驻留时间的伪随机变化，即通信系统占用时隙的非线性变化，有效增加敌方的干扰追踪时间。然而文献[4-6]没有根据干扰的变化设计跳频图案。文献[7]采用自适应跳频技术，自动选择无干扰信道以建立高质量稳定通信，然而与人工智能结合的人为干扰的出现让实时性较差的自适应跳频难以应对。

深度强化学习（deep reinforcement learning，DRL）结合深度学习的表示能力和强化学习的推理能力，可以适应时变的电磁环境。DRL 根据学习目标的不同可以分为基于价值和基于策略的两类学习算法，基于价值的 DRL 有深度 Q 网络（deep Q-network，DQN），基于策略的 DRL 有深度确定性策略梯度（deep deterministic policy gradient，DDPG）和近端策略优化（proximal policy optimization，PPO）。文献[8]应用 DQN 进行“双变”跳频图案决策，对经验回放 DQN 引入帕雷托（Pareto）样本的概念，提出了基于帕雷托样本的优先经验回放深度 Q 网络（deep Q-network with priority experience replay based on Pareto samples，PPER-DQN）算法，提升了 DQN 的经验池筛选性能，更好地适应变化的电磁环境。但是其采用传统优先级定义和 Pareto 样本，导致复杂度增大。文献[9]将分类经验回放引入 DDPG，提出了采用分类经验回放的深度确定性策略梯度（deep deterministic policy gradient with classified experience replay，CER-DDPG）方法，降低了传统优先级经验回放机制的复杂度。但是 DDPG 缺少对状态的“评判”，在不同环境中表现差异较大，稳定性较差。PPO 是 DRL 中基于“行动器－评判器”的算法，并且适用于连续状态－动作空间，它具有信赖域策略优化（trust region policy optimization，TRPO）的优点，同时更易于实现，相比其他在线策略梯度方法，PPO 还具有更好的稳定性和可靠性[10]，已在机器人控制[11]、模块化生产控制[12]和城市道路交通控制[13]等领域得到应用。本文通过状态－动作－奖励三元组的构造将“双变”跳频图案决策问题转化为一个序列优化问题，并且“双变”跳频系统时域和频域的连续性契合 PPO 算法中的连续状态－动作空间，因此本文研究复杂电磁干扰中应用 PPO 的“双变”跳频图案智能设计。

PPO 算法性能和效率与其“行动器”和“评判器”的策略有关。不同于传统 DRL，PPO 产生的样本数据与“行动器”的策略有关，所以不能简单地通过经验池复用更新，而需要通过另外的途径提高样本数据利用率。文献[14]利用以前学到的模型对“行动器”网络初始化，提出模型加速近端策略优化（proximal policy optimization with model accelerate，MA-PPO），加速学习过程并提高运算效率。文献[15]将策略引入“评判器”价值函数的更新过程，提出具有策略反馈的 PPO（proximal policy optimization with policy feed-

back，PF-PPO）算法，与 PPO 相比，PF-PPO 具有更快的收敛速度、更高的奖励和更小的奖励方差。DRL 的网络更新梯度包含了偏差和方差信息，偏差反映了期望预测与真实结果的偏离程度，方差反映了样本数据的变动所导致的学习性能的变化，较低的偏差和方差可以保证算法的准确性和稳定性。文献[16]将 PPO“评判器”网络的 n 步回报估计改为广义优势估计（generalized advantage estimation，GAE），较好地平衡了算法的偏差和方差；文献[17]通过蒙特卡洛估计，在“行动器”中为不同特征配置不同基线，降低了算法的方差，从而加快了学习速度；文献[18]设计稀疏奖励，取代“执行差额”奖励，降低了算法的偏差。

但文献[14-18]方法仍存在方差较高和样本利用较低的问题，对此本文提出了 ET-PPO 算法。针对 PPO“行动器”网络样本更新方式的高方差问题，本文引入加权重要度采样（weighted importance sampling，WIS）减小方差，提高学习阶段的稳定性；针对“评判器”网络收敛速度慢的问题，本文引入资格迹（eligibility trace，ET），在不陷入局部最优解的前提下加速收敛。为了智能决策跳频图案，本文设计“行动器”的动作选择策略为 Beta 分布策略，且在“行动器”的目标函数中添加策略的熵项以避免落入局部最优解。仿真结果表明，相比于传统 PPO、PPER-DQN 和 CER-DDPG，ET-PPO 具有更快的学习速度、更高的奖励和更好的平稳性。

1 问题建模

本文主要研究“双变”跳频系统在面对阻塞干扰和多频连续波干扰时的跳频图案智能决策，希望在干扰较强的频段，“双变”跳频可以自适应地提高跳速、增大信道划分间隔以尽快跳出该频段；在干扰较弱频段，“双变”跳频可以自适应地放慢跳速、减小信道划分间隔以保持长时间低误码率的高质量通信。

“双变”跳频系统可以在非连续频带灵活分配频率，例如在常规跳频系统中，本来不适用当前非连续频带的跳频序列，在“双变”跳频系统中通过可变的跳频速度和信道划分间隔则可将原本出现在不可用频带的跳频信号转移至可用频带上，提高了频谱资源的利用率。由此降低跳频序列设计难度，增加跳频序列集合的跳频序列数量，使以码分多址为基础的跳频系统支持更大的用户容量。

多级频移键控（multiple-frequency-shift keying，MFSK）下的“双变”跳频信号可以表示为：

$$x(t)=A\sum_{k=1}^{K}\cos\left[2\pi f_{\mathrm{c}}t+a_k 2\pi f(a_k)t\right]g_{T(a_k)}\left[t-kT(a_k)\right] \tag{1}$$

其中，A 为振幅；K 为总跳变次数；k 为第 k 跳；f_{c} 为最小跳频频率；a_k 为跳频序列发生器生成的频率控制字，对应不同跳频频率；$f(a_k)$ 和 $T(a_k)$ 为时变的跳频频率间隔和跳频速率，由 a_k 控制；$g_T(t)$ 为持续时间 T 的单位门函数。

在加性白高斯噪声背景下，二进制频移键控（frequency-shift keying，2FSK）相干解调总误码率为：

$$P=\frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{r}{2}}\right) \tag{2}$$

其中，erfc(·)表示补余误差函数，r 为信噪比。由此可得，通过跳频图案智能决策增大解调器输入端的平均信干噪比可以降低跳频通信系统误码率，考虑到实际场景中电波传输的传播损耗和瑞利衰落，本文将决策目标函数设计为最大化平均信干噪比（signal to interference plus noise ratio，SINR），即：

$$\begin{aligned}\max \mathrm{SINR}_k &= \max\int_{t_k-\tau_k/2}^{t_k+\tau_k/2}\frac{h(t)P(t)L_J}{P_{J_k}(t)L_x+P_n(t)}\mathrm{d}t/\tau_k\\ &= \max\int_{t_k-\tau_k/2}^{t_k+\tau_k/2}\frac{h(t)P(t)L_J}{\int_{f_k-\varphi_k/2}^{f_k+\varphi_k/2}J_k(f,t)\mathrm{d}fL_x+\sigma_n^2(t)}\mathrm{d}t/\tau_k\\ &\text{s.t.}\quad L=32.5+20\lg d+20\lg f\end{aligned} \tag{3}$$

其中，$P(t)$为信号功率，$P_{J_k}(t)$为干扰功率，$P_n(t)$是均值为 0、方差为$\sigma_n^2(t)$的高斯白噪声功率，$h(t)$是均值为 0、方差为$\sigma_h^2(t)$的瑞利衰落信道响应，L_J和L_x分别为跳频信号和干扰信号的传播损耗，d为信号传播距离，t_k为第k跳信号的时隙中点，τ_k为第k跳信号的时隙长度，$J_k(f,t)$为第k跳干扰功率谱密度，f_k为第k跳信号的中心频率，φ_k为第k跳信号的频带宽度。

2　算法设计

2.1　改进的近端策略优化算法

PPO 是 DRL 中一种基于行动器－评判器（actor-critic，AC）的算法，结合了策略迭代法和价值迭代法。

PPO“行动器”网络的更新方式采用策略梯度（policy gradient，PG）法进行策略迭代。在 PG 算法中，智能体（行动器）的参数化策略为π_θ，其中θ为策略参数。从初始状态s_0出发，智能体在每一步动作选择a_t后，获得一个奖励r_{t+1}，进入一个新的状态s_{t+1}，这样通过一系列的动作选择一直到终止状态s_T，形成一个如图 1 所示序列$\tau(s_0,a_0,r_1,s_1,a_1,r_2,s_2\cdots)$。智能体从一个序列中获得的总奖励为$R(\tau)$。PG 算法通过更新参数$\theta$以最大化目标函数$J(\theta)$，即：

$$\begin{cases} J(\theta)=E_{\tau\sim\pi_\theta(\tau)}[R(\tau)] \\ \theta\leftarrow\theta+\alpha^\theta\nabla_\theta J(\theta) \end{cases} \tag{4}$$

其中，α^θ为参数θ的学习率，$E\{\cdot\}$是数学期望运算，$\nabla_\theta J(\theta)$是$J(\theta)$对θ的梯度。

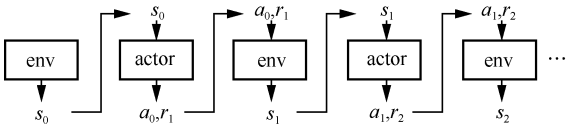


图 1　序列τ

在不改变更新的期望值的前提下，为了降低 PG 算法的方差，引入状态价值函数$v_\omega(\tau)$作为基线，$\boldsymbol{\omega}$是状态价值参数向量。为了提高训练速度和数据利用率，PG 算法的更新从同轨策略（on-policy）更新转换为离轨策略（off-policy）更新。设对应参数θ的策略为目标策略，对应参数θ'的策略为行动策略，定义目标策略和行动策略的重要度采样比（importance sampling ratio，ISR）为：

$$\rho=\frac{\prod_{t=0}^{T-1}\pi_\theta(a_t\mid s_t)p(s_{t+1}\mid s_t,a_t)}{\prod_{t=0}^{T-1}\pi_{\theta'}(a_t\mid s_t)p(s_{t+1}\mid s_t,a_t)}=\prod_{t=0}^{T-1}\frac{\pi_\theta(a_t\mid s_t)}{\pi_{\theta'}(a_t\mid s_t)} \tag{5}$$

其中，p为环境的状态转移概率，与策略参数无关，所以可被约分。

结合基线和重要性采样，PG 算法的离轨策略更新为：

$$\begin{cases} J(\theta)=E_{\tau\sim\pi_\theta(\tau)}\{[R(\tau)-v_\omega(\tau)]\rho\} \\ \theta\leftarrow\theta+\alpha^\theta\nabla_\theta J(\theta) \end{cases} \tag{6}$$

通过 Clip 技巧将重要度采样比ρ限制在一定范围内，可以防止参数θ和θ'分布差异过大而导致策略更新的幅度异常。为了避免在学习初始阶段落入局部最优解，本文在目标函数中引入策略的熵项（entropy），即：

$$\mathrm{entropy}(\pi_\theta)=-\int_R p_{\pi_\theta}(x)\mathrm{lb}p_{\pi_\theta}(x)\mathrm{d}x \tag{7}$$

其中，$p_{\pi_\theta}(x)$为策略π_θ决定的“行动器”输出分布。在学习初始阶段，策略随机选择动作，该熵具有最大值；从学习中期起，策略仅固定选择一个动作，该熵具有最小值。综上可得改进的随机梯度的 PPO“行动器”网络的更新为：

$$\begin{cases} J(\theta)=\min\{[R(\tau)-v_\omega(\tau)]\rho, \\ [R(\tau)-v_\omega(\tau)]\mathrm{clip}(\rho,1-\varepsilon,1+\varepsilon)\}+\eta\cdot\mathrm{entropy}(\pi_\theta) \\ \theta\leftarrow\theta+\alpha^\theta\nabla_\theta J(\theta) \end{cases} \tag{8}$$

其中，ε为 Clip 参数，η为熵项系数，对目标函数取 min 的目的是避免参数θ更新太快。

PPO“评判器”网络通过价值迭代法更新参数$\boldsymbol{\omega}$以最小化目标函数$J(\omega)$，即：

$$\begin{cases} J(\omega)=E_{\tau\sim\pi_\theta(\tau)}\{[R(\tau)-v_\omega(\tau)]^2\} \\ \boldsymbol{\omega}\leftarrow\boldsymbol{\omega}+\alpha^\omega\nabla_\omega J(\omega) \end{cases} \tag{9}$$

其中，α^{ω} 为参数 $\boldsymbol{\omega}$ 的学习率。

2.2 加权重要度采样和资格迹

PPO 算法“行动器”网络进行期望更新时，其重要度采样通过简单平均实现，被称为普通重要度采样（ordinary importance sampling，OIS），如式（10）所示。

$$\frac{\sum_{i=1}^{i'}\rho_i[R(\tau_i)-v_{\omega}(\tau_i)]}{i'}\leftarrow R(\tau)-v_{\omega}(\tau) \tag{10}$$

其中，ρ_i 是第 i 个采样数据的重要度采样比，i' 为采样数据个数。假设当 $i'=1$、$\rho_1=10$ 时，估计值是观测到的回报值的 10 倍，是有偏差的，且方差较大。因此本文引入加权重要度采样（weighted importance sampling，WIS），它采用加权平均的方法，如式（11）所示。加权重要度采样的估计是无偏的，且其方差较小，可以加快学习速度。

$$\frac{\sum_{i=1}^{i'}\rho_i[R(\tau_i)-v_{\omega}(\tau_i)]}{\sum_{i=1}^{i'}\rho_i}\leftarrow R(\tau)-v_{\omega}(\tau) \tag{11}$$

PPO 算法“评判器”网络更新采用 n 步时序差分算法，其需要存储最近的 n 个特征向量，且延迟 n 步后才执行，从而影响算法的收敛速度。本文引入资格迹方法，其只需要追踪一个与参数向量 $\boldsymbol{\omega}$ 同维的迹，同时，在状态改变后就马上学习并影响后续决策。用 $\boldsymbol{z}$ 表示资格迹，其更新如下：

$$\boldsymbol{z}\leftarrow\gamma\lambda z+\nabla_{\omega}v_{\omega}(\tau) \tag{12}$$

其中，$\gamma\in[0,1]$ 为奖励折扣系数，$\lambda\in[0,1]$ 为迹衰减率参数。结合资格迹，式（9）参数向量 $\boldsymbol{\omega}$ 的更新改写为：

$$\boldsymbol{\omega}\leftarrow\boldsymbol{\omega}-\alpha^{\omega}E_{\tau\sim\pi_{\theta}(\tau)}\left\{\left[R(\tau)-v_{\omega}(\tau)\right]\boldsymbol{z}\right\} \tag{13}$$

由式（13）可知，资格迹追踪参数向量 $\boldsymbol{\omega}$ 的梯度，决定 $\boldsymbol{\omega}$ 分量的更新。

由式（12）可知，资格迹向量在每一步以 $\gamma\lambda$ 的速率衰减，半衰期为：

$$\tau_{\gamma\lambda}=\frac{-1}{\mathrm{lb}(\gamma\lambda)},(0<\gamma\lambda<1) \tag{14}$$

其中，这里的半衰期指 t 时刻得到的价值梯度将于 $t+\tau_{\gamma\lambda}$ 时刻衰减一半，价值梯度权重以速率 $\gamma\lambda$ 衰减示意图如图 2 所示。λ 越小表示资格迹向量衰减得越快，算法越看重浅层回报，当 $\lambda=0$ 时，算法退化为单步时序差分算法，当前时刻的更新仅被前一个时刻的价值梯度分配。λ 越大表示资格迹向量衰减得越慢，算法赋予深层回报更大的权重，当 $\lambda=1$ 时，价值梯度每一时间步仅衰减 γ，与蒙特卡洛算法一致。

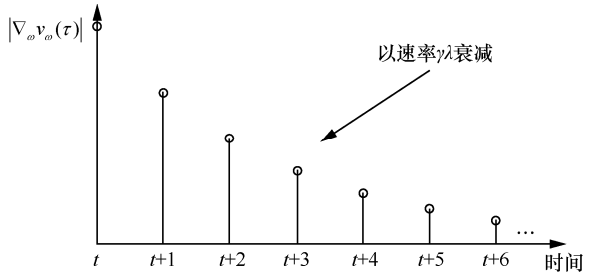


图 2 价值梯度权重以速率 $\gamma\lambda$ 衰减示意图

2.3 状态动作空间、奖励函数及动作选择策略设计

假设系统在时间段 $T\in[T_{\text{lower}},T_{\text{upper}}]$，频段 $W\in[W_{\text{lower}},W_{\text{upper}}]$ 内应用“双变”跳频技术通信，信源功率为 P，跳频速度 $V\in[V_{\text{lower}},V_{\text{upper}}]$，信道划分间隔 $D\in[D_{\text{lower}},D_{\text{upper}}]$。为将“双变”跳频图案决策转化为一个序列优化问题，假设系统在信道估计技术的支持下预测了时间段 T 内的信道干扰环境。定义状态 $\boldsymbol{s}$ 为二维向量 (t,f)，其中，$t\in[T_{\text{lower}},T_{\text{upper}}]$，$f\in[W_{\text{lower}},W_{\text{upper}}]$，设定终止状态的价值为 0，即 $v_{\boldsymbol{\omega}}(s_{\mathrm{T}})=0$；定义动作 $\boldsymbol{a}$ 为二维向量 (v,d)，其中 $v\in[V_{\text{lower}},V_{\text{upper}}]$，$d\in[D_{\text{lower}},D_{\text{upper}}]$。

在每一步中，智能体从环境得到一个奖励，强化学习算法的唯一目标就是最大化长期总奖励。为了实现高质量通信，通信系统需要一个高信干噪比的跳频图案，所以本文使用式（3）作为奖励函数。

针对本文所定义的有限范围内的动作空间，

本文采用如式（15）所示的 Beta 分布概率密度函数作为“行动器”网络输出的动作选择策略。

$$f(x;\alpha,\beta)=\frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\mathrm{d}u} \tag{15}$$

其中，α 和 β 为 Beta 分布参数。

2.4　算法步骤

综上可得，本文基于改进 PPO 的“双变”跳频图案智能决策算法具体步骤如下。

步骤 1　预测时间段 T 内的干扰环境，初始化信号功率 P、跳频序列 F。

步骤 2　初始化“行动器”网络的参数 θ、学习率 α^{θ}、经验利用次数 N^{θ}、“评判器”网络的参数 $\boldsymbol{\omega}$、学习率 α^{ω} 和经验利用次数 N^{ω}，初始化“旧行动器”网络 θ'。

步骤 3　初始化奖励折扣系数 γ、资格迹衰减率 λ、Clip 参数 ε、熵项系数 η、Beta 分布参数 α 和 β、数据采样轮数 N，初始化当前数据采样轮次 $n\leftarrow 0$。

步骤 4　初始化时间步 $t\leftarrow 0$。

步骤 5　根据当前状态 s_t，通过“行动器”网络采样生成的 Beta 分布获得动作 a_t，执行并得到下一状态 s_{t+1}，由式（3）计算奖励 r_{t+1}。

步骤 6　将“行动器”网络参数向量赋值给“旧行动器”网络，固定行动策略 $\pi_{\theta'}$，根据式（8）结合加权重要度采样更新“行动器”网络 N^{θ} 次，根据式（12）更新资格迹向量，根据式（9）结合资格迹向量更新“评判器”网络 N^{ω} 次。

步骤 7　如果 s_{t+1} 是终止状态，则结束本轮数据采样，令 $n\leftarrow n+1$，跳转至步骤（8）；如果 s_{t+1} 为非终止状态，则令 $s_t\leftarrow s_{t+1}$ 并跳转至步骤（5）。

步骤 8　如果 n=N，结束算法；否则跳转至步骤 4。

3　实验结果及性能分析

仿真中参数设置如下：跳频系统的信源功率 $P=100$ mW，时间段 $T\in[0,100]$ ms，频段 $W\in[0,200]$ MHz，跳频速度 $V\in[125,2\,000]$ hop / s，信道划分间隔 $D\in[1,4]$ MHz，高斯白噪声功率 $P_{\mathrm{n}}=10^{-7}$ mW，瑞利衰落方差 $\sigma_{\mathrm{h}}^2(t)=1$；改进 PPO 的数据采样轮数 $N=200$，行动器网络学习速率 $\alpha^{\theta}=0.000\,01$，“评判器”网络学习速率 $\alpha^{\omega}=0.000\,1$，“行动器”网络经验利用次数 $N^{\theta}=10$，“评判器”网络经验利用次数 $N^{\omega}=5$，奖励折扣系数 $\gamma=0.9$，Clip 参数 $\varepsilon=0.2$。

3.1　实验 1　熵项系数 η 对算法性能的影响

干扰环境用频谱瀑布图表示，包含高斯白噪声、阻塞干扰和扫频干扰的干扰环境如图 3 所示，颜色越深，干扰功率越大。

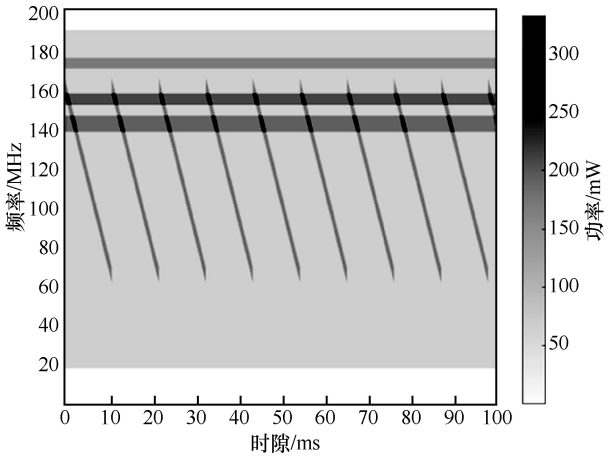


图 3　干扰环境

在图 3 所示干扰环境中，资格迹衰减率 $\lambda=0.8$，熵项系数 η 对算法性能的影响如图 4 所示，给出了熵项系数 η 分别取 0、0.01、0.1 时，本文算法生成的总收益 $R(\tau)$ 与数据采样轮次 N 关系曲线，为便于分析比较，仅展示前 100 轮迭代数据。由图 4 可知，$\eta=0$ 时，算法稳态收益最少；$\eta=0.1$ 时，算法稳态收益较高，但是在第 50 轮采样前曲线振荡幅度较大，第 50 轮之后曲线渐渐平稳；$\eta=0.01$ 时，算法在学习过程中较为稳定，且稳态收益与 $\eta=0.1$ 时基本一样。综合考虑稳态收益和算法的平稳性，下文取 $\eta=0.01$。

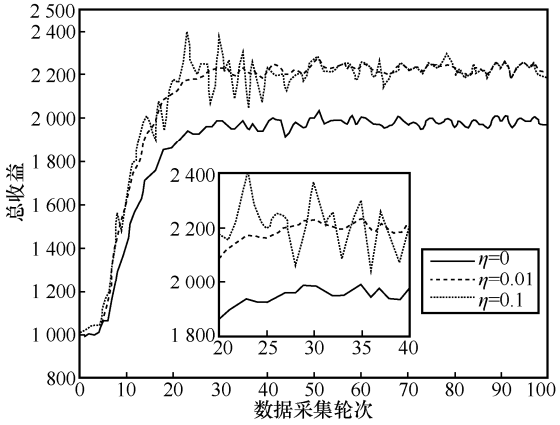


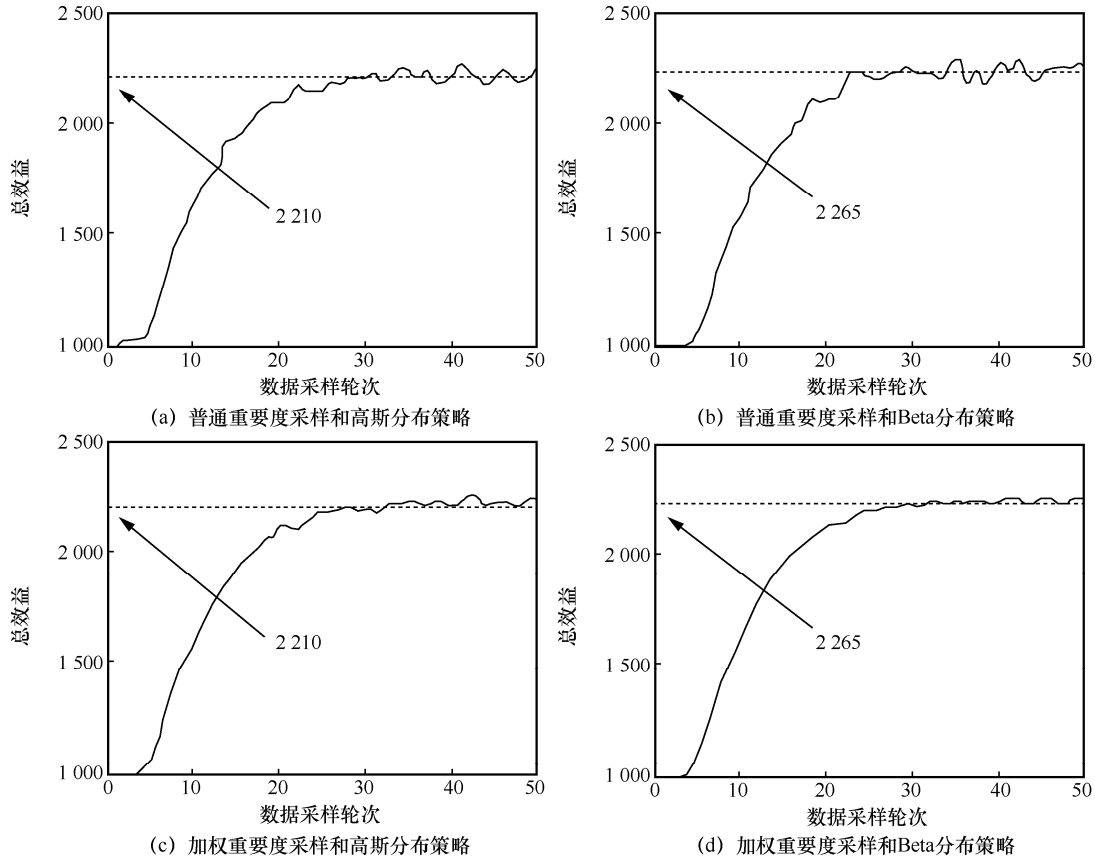
图 4　熵项系数 η 对算法性能的影响

3.2　实验 2　加权重要度采样和 Beta 分布策略的应用

干扰环境同实验 1，“行动器”不同更新方式对应性能曲线如图 5 所示，是“行动器”网络分别采用高斯分布和 Beta 分布的动作选择策略及普通重要度采样和加权重要度采样的更新方法得到的总收益 $R(\tau)$ 与数据采样轮次 N 的关系曲线。由图 5 可知，图 5（a）、图 5（c）收敛在 2 210 附近，图 5（b）、图 5（d）收敛在 2 265 附近，这说明 Beta 分布策略有效提升了系统的总收益。图 5(a)、图 5（b）的振荡幅度分别大于图 5（c）、图 5（d），这说明加权重要度采样使“行动器”网络的学习过程具有更好的稳定性。

3.3　实验 3　资格迹衰减率 λ 对 ET-PPO 算法的影响

干扰环境同实验 1，固定 $\gamma = 0.9$，不同资格迹衰减率和不同 n 取值对应的性能曲线如图 6 所示。迹衰减率 λ 分别取 0.2、0.4、0.6、0.8 和 1 时，ET-PPO 算法得到的总收益 $R(\tau)$ 与数据采样轮次 N 关系曲线如图 6（a）所示。由图 6（a）可见，相比于 n 步时序差分算法，不同 λ 取值的 ET-PPO 算法收敛速度远高于 PPO 算法；λ 值对算法的总收益影响较小，都稳定在 2 265 左右。λ=0.2 时收敛最慢，在第 100 轮采样完成收敛；λ=0.8 时收敛最快，在第

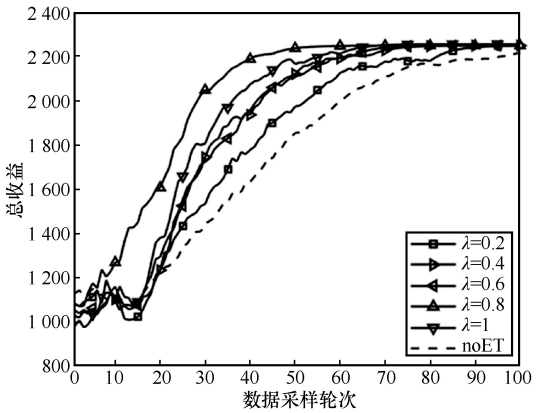


(a) 普通重要度采样和高斯分布策略　(b) 普通重要度采样和Beta分布策略

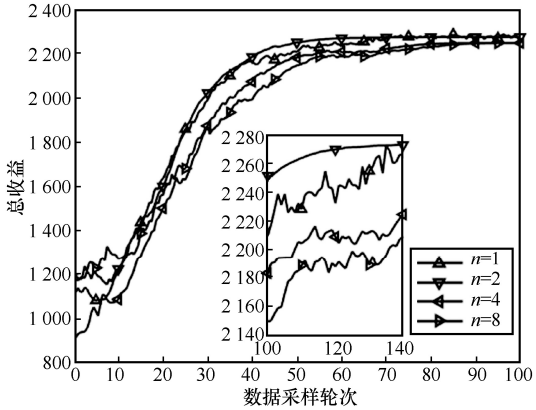
(c) 加权重要度采样和高斯分布策略　(d) 加权重要度采样和Beta分布策略

图 5　“行动器”不同更新方式对应性能曲线

40 轮采样完成收敛，下文取 λ=0.8。n 分别取 1、2、4、8 时，PPO 算法得到的总收益 $R(\tau)$与数据采样轮次 N 关系曲线如图 6（b）所示，不同 n 取值对总收益 $R(\tau)$的影响较小，对网络收敛速度的影响较大；当 $n=2$ 时，PPO 算法收敛最快且稳态收益较高。



(a) 不同迹衰减率对应的ET-PPO算法性能曲线



(b) 不同n取值对应的PPO算法性能曲线

图 6　不同资格迹衰减率和不同 n 取值对应的性能曲线

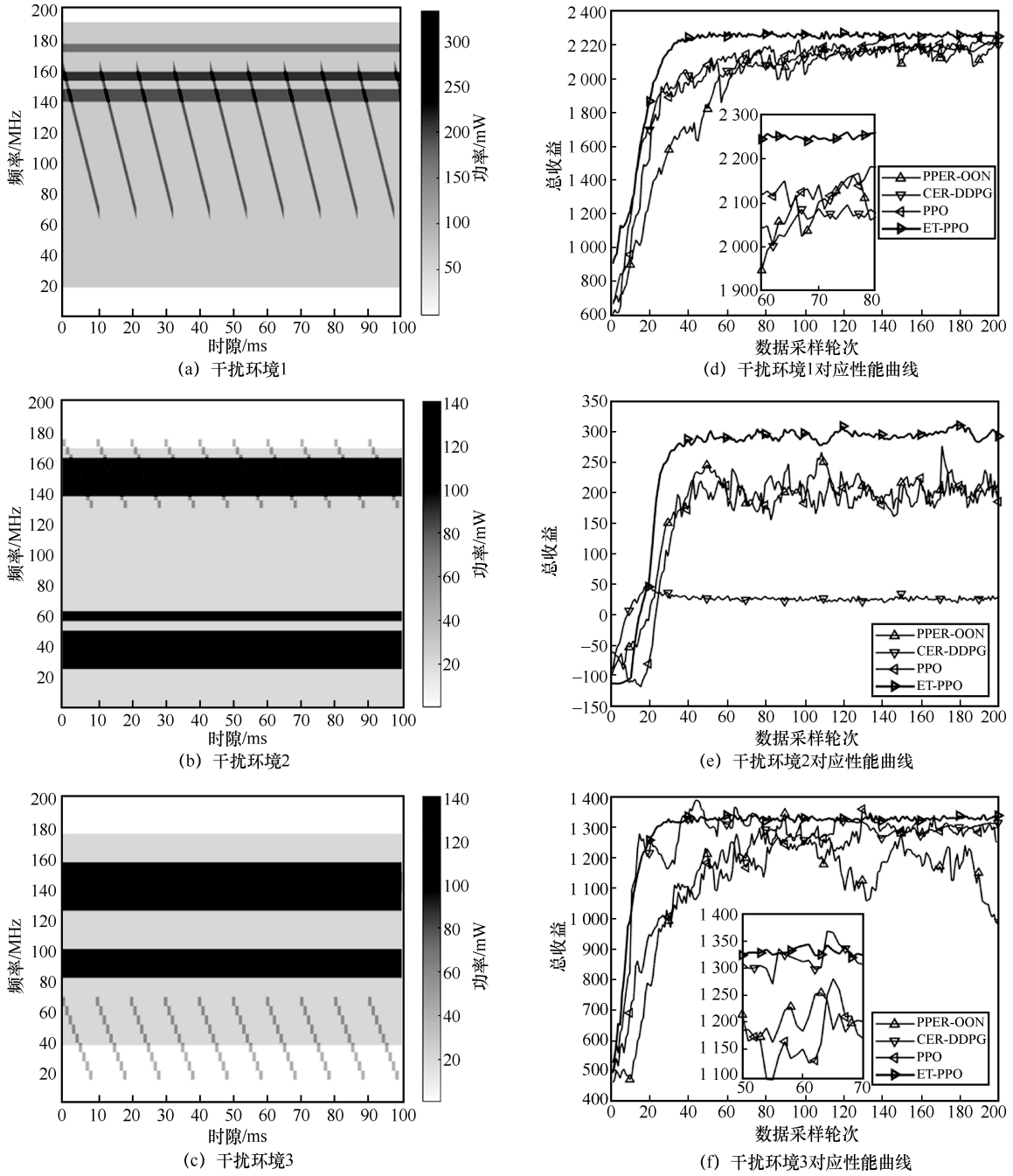
3.4　实验 4　不同干扰环境下算法对比分析

本节分析比较应用 PPER-DQN[8]、CER-DDPG[9]、传统 PPO[10]和本文 ET-PPO 的“双变”跳频图案决策性能。为保证公平性，令所有算法具有相同的学习率、奖励设定和奖励折扣系数。本文“双变”跳频图案决策属于连续状态–连续动作问题，而 PPER-DQN 适用于连续状态–离散动作问题，所以 PPER-DQN 的仿真实验中，对动作空间进行离散化处理，设置跳频速度集合为[125, 250, 500, 1 000, 2 000] hop/s，信道划分间隔集合为[1, 2, 3, 4] MHz。另外 PPER-DQN 以时序差分误差作为优先经验回放依据，CER- DDPG 以立即奖励作为分类经验回放依据。3 种干扰环境及相应性能曲线如图 7 所示。

由图 7 可知，ET-PPO 在不同干扰环境中具有更快的收敛速度和更稳定的性能，这说明 ET-PPO 对动态变化干扰环境适应性较强。ET-PPO 通过加权重要性采样提高了采样数据利用率，降低了学习的方差，在不同干扰环境中算法都在第 40 轮采样之前完成收敛；通过 Beta 分布的动作选择策略平衡了强化学习探索与利用的矛盾，在学习前期保证较高的探索度，在学习后期以利用为主，所以曲线振荡幅度较小。PPER-DQN 性能收敛在比较低的水平，这是因为离散化动作空间寻找的最优动作不精细，而更精细的动作空间离散化将增加训练成本。PPO 和 CER-DDPG 性能受环境影响较大，CER-DDPG 在环境 2 中陷入局部最优解仅有 50 的总收益表现，PPO 曲线振荡幅度大，需要手动设置参数以平衡算法的探索与利用。

4　结束语

针对 PPO 算法“行动器”网络更新方差大、“评判器”网络更新收敛速度慢的问题，本文将加权重要性采样和资格迹方法引入 PPO 算法；将 Beta 分布作为“行动器”网络输出的动作选择策略，并将该策略的熵项添加到“行动器”网络的目标函数上，使算法在学习初始阶段充分学习参数以避免落入局部最优解；将“双变”跳频图案决策建模为序列优化问题，设计了合适的奖励函数和策略函数。在不同电磁干扰环境中应用本文算法的“双变”跳频图案决策结果表明，相比于 PPER-DQN、CER-DDPG 和传统 PPO，本文所提出算法具有更快的收敛速度且不易落入局部最优解，对环境适应性强。本文针对的是连续状态–动作空间问题，相比于离散问题具有更复杂的随机性，但是状态–动作维数不高，未来将研究强化学习在高维特征空间中的应用。



(a) 干扰环境1　(d) 干扰环境1对应性能曲线

(b) 干扰环境2　(e) 干扰环境2对应性能曲线

(c) 干扰环境3　(f) 干扰环境3对应性能曲线

图 7　3 种干扰环境及相应性能曲线

参考文献：

[1] 任兴旌. 跳频通信关键技术研究及系统设计[D]. 兰州: 兰州交通大学, 2018.
REN X J. Key technology research and system design of frequency hopping communication[D]. Lanzhou: Lanzhou Jiatong University, 2018.

[2] 柳永祥, 姚富强, 梁涛. 变间隔、变跳速跳频通信技术[C]// 军事电子信息学术会议. 2006:518-521.
LIU Y X, YAO F Q, LIANG T. Bivariate frequency hopping communication technology[C]//Academic Conference on Military Electronic Information. 2006: 518-521.

[3] 严季, 梁涛, 祈竹. 变跳速、变间隔跳频通信技术研究[J]. 无线通信技术, 2012, 21(4): 25-29.
YAN J, LIANG T, QI Z. Research on thefrequenct hopping communication technology of variable hopping rate and variable interval[J]. Wireless Communication Technology, 2012,

21(4): 25-29.

[4] 汪小林, 黎亮, 张抒. 基于均匀性补偿的跳频图案生成方法[J]. 兵工自动化, 2018, 37(9): 12-14.
WANG X L, LI L, ZHANG S. Frequency hopping based on uniformity compensation[J]. Ordnance Industry Automation, 2018, 37(9): 12-14.

[5] 李金涛. 宽间隔跳频序列设计与性能研究[D]. 成都: 西南交通大学, 2007.
LI J T. Study on frequency hopping sequences with givenminimumgap[D]. Chengdu: Southwest Jiaotong University, 2007.

[6] 陈刚, 黎福海. 变速跳频通信抗跟踪干扰性能的研究[J]. 火力与指挥控制, 2016, 41(7): 107-109.
CHEN G, LI F H. Research on anti-follower jamming performance of variable rate frequency hopping communications[J]. Fire Control & Command Control, 2016, 41(7): 107-109.

[7] 王越超. 自适应跳频通信系统关键技术研究[D]. 南京: 东南大学, 2018.
WANG Y C. Research on key technology of adaptive frequency hopping communication system[D]. Nanjing: Southeast University, 2018.

[8] ZHU J S, ZHAO Z J, ZHENG S L. Intelligent anti-jamming decision algorithm of bivariate frequency hopping pattern based on DQN with PER and Pareto[J]. International Journal of Information Technology and Web Engineering, 2022, 17(1): 1-23.

[9] 时圣苗, 刘全. 采用分类经验回放的深度确定性策略梯度方法[J]. 自动化学报, 2022, 48(7): 1816-1823.
SHI S M, LIU Q. Deep deterministic policy gradient with classified experience replay[J]. Acta Automatica Sinica, 2022, 48(7): 1816-1823.

[10] CANO L G, FERREIRA M, DA S S A, et al. Intelligent control of a quadrotor with proximal policy optimization reinforcement learning[C]//Proceedings of 2018 Latin American Robotic Symposium, 2018 Brazilian Symposium on Robotics (SBR) and 2018 Workshop on Robotics in Education (WRE). Piscataway: IEEE Press, 2018: 503-508.

[11] 张浩昱, 熊凯. 基于近端策略优化算法的四足机器人步态控制研究[J]. 空间控制技术与应用, 2019, 45(3): 53-58.
ZHANG H Y, XIONG K. On gait control of quadruped robot based on proximal policy optimization algorithm[J]. Aerospace Control and Application, 2019, 45(3): 53-58.

[12] MAYER S, CLASSEN T, ENDISCH C. Modular production control using deep reinforcement learning: proximal policy optimization[J]. Journal of Intelligent Manufacturing, 2021, 32(8): 2335-2351.

[13] 舒凌洲. 基于深度强化学习的城市道路交通控制算法研究[D]. 成都: 电子科技大学, 2020.
SHU L Z. Research on urban traffic control algorithm based on deep reinforcement learning[D]. Chengdu: University of Electronic Science and Technology of China, 2020.

[14] GUAN Y, REN Y G, LI S E, et al. Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization[J]. IEEE Transactions on Vehicular Technology, 2020, 69(11): 12597-12608.

[15] GU Y, CHENG Y H, CHEN C L P, et al. Proximal policy optimization with policy feedback[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(7):4600-4610.

[16] 王鸿涛. 基于强化学习的机械臂自学习控制[D]. 哈尔滨: 哈尔滨工业大学, 2019.
WANG H T. Self learning control of mechanical arm based on reinforcement learning[D]. Harbin: Harbin Institute of Technology, 2019.

[17] ZHANG L, ZHANG Y S, ZHAO X, et al. Image captioning via proximal policy optimization[J]. Image and Vision Computing, 2021, 108: 104126.

[18] LIN S Y, BELING P A. An end-to-end optimal trade execution framework based on proximal policy optimization[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. California: International Joint Conferences on Artificial Intelligence Organization, 2020: 4548-4554.

[作者简介]

陈一波（1998− ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为认知无线电。

赵知劲（1959− ），女，博士，杭州电子科技大学通信工程学院教授、博士生导师，主要研究方向为信号处理、认知无线电技术。