基于多智能体强化学习的混合博弈模式下多无人机辅助通信系统设计

吴官翰[①②] 贾维敏[①] 赵建伟[*①] 高飞飞[③] 姚敏立[①]

[①](火箭军工程大学 西安 710038)

[②](中国酒泉卫星发射中心 酒泉 735000)

[③](清华大学 北京 100084)

摘 要：空天地一体化通信作为未来6G的发展方向，很好地弥补了当前无线通信覆盖不足的弊端。该文提出一种基于多智能体强化学习(MARL)的多无人机(Multi-UAV)辅助通信算法，在用户与无人机(UAVs)构成的混合博弈模式下求解纳什均衡近似解，解决了动态环境下UAVs轨迹设计、多维资源调度以及用户接入策略联合优化问题。结合马尔可夫博弈概念建模该连续决策过程，以集中训练分布式执行(CTDE)机制，将近端策略优化(PPO)算法扩展到多智能体领域。针对离散与连续共存的动作空间设计了两种策略输出模式，并且结合Beta策略改进实现，最后通过仿真实验验证了算法的有效性。

关键词：多无人机辅助通信；多智能体强化学习；混合博弈；纳什均衡

中图分类号：TN911 **文献标识码**：A **文章编号**：1009-5896(2022)03-0940-11

DOI: 10.11999/JEIT210662

MARL-based Design of Multi-Unmanned Aerial Vehicle Assisted Communication System with Hybrid Gaming Mode

WU Guanhan[①②] JIA Weimin[①] ZHAO Jianwei[①] GAO Feifei[③] YAO Minli[①]

[①](*Rocket Force University of Engineering*, *Xi'an* 710038, *China*)

[②](*Jiuquan Satellite Launch Center*, *Jiuquan* 735000, *China*)

[③](*Tsinghua University*, *Beijing* 100084, *China*)

Abstract: As the future development direction of 6G, integrated space-air-ground communication well compensates for the drawback of insufficient current wireless communication coverage. In this paper, a Multi-Unmanned Aerial Vehicle (Multi-UAV) assisted communication algorithm with Multi-Agent Reinforcement Learning (MARL) is proposed to solve the Nash equilibrium approximate solution in a hybrid game model composed of users and UAVs and solve the joint optimization problem of UAV trajectory design, multidimensional resource scheduling and user access strategy in dynamic environment. The Markov game concept is exploited to model this continuous decision process with a Centralized Training Distributed Execution (CTDE) mechanism, and the Proximal Policy Optimization (PPO) algorithm is extended to the multi-agent domain. Two policy output modes are designed for the action space, where both the discrete and continuous actions coexist. Then, the implementation is improved by combining Beta policy. Finally, the effectiveness of the algorithm is verified by simulation experiments.

Key words: Multi-Unmanned Aerial Vehicle (Multi-UAV) assisted communications; Multi-Agent Reinforcement Learning (MARL); Hybrid game; Nash equilibrium

1 引言

在当前5G移动通信中，随着各种新兴产业迅猛发展地面骨干网承受着巨大的数据传输压力。同时受限于地理条件的影响，许多偏远地区仍处于无线覆盖欠缺的状态[1]。这些前所未有的对高质量无线通信服务的需求，对当前传统地面通信网络提出了严峻挑战。为此，在未来6G及以后的无线通信中，无人机(Unmanned Aerial Vehicle, UAV)作为空中接入节点辅助地面通信成为一种有前途的解决方案[2–8]。

收稿日期：2021-07-02；改回日期：2021-09-06；网络出版：2021-09-15

*通信作者： 赵建伟 zhaojianweiep@163.com

基金项目：国家自然科学基金(62001500)

Foundation Item: The National Natural Science Foundation of China (62001500)

然而在多无人机辅助地面通信的系统设计中，由于UAV与地面用户(Ground Users, GU)位置的动态性，网络拓扑结构动态变化。同时在有限通信资源条件下，如何合理分配资源能在保证小区内用户公平通信情况下最大化系统吞吐量，是一个典型的NP-hard (Non-deterministic Polynomial hard)问题，该问题的非凸性导致传统优化方法难以应用。在现有的一些工作中，将与这类似的具有非凸性的优化问题简化为多个凸的子问题进行求解，并且通过迭代收敛到次优解[9–11]。这些方法能够在较短时间内收敛，但却是以损失精度为代价。同时，在一些基于启发式算法求解的研究中[12,13]，利用多次迭代在解空间中搜寻近似最优解，但这些方法在动态环境中的效率却大为降低。

现有的大部分工作主要针对无人机在固定的资源分配方案，或对单一通信资源调度的前提下对无人机进行轨迹优化[14–16]。优化目标仅局限于无人机或地面接入控制[17]，并未从整个通信系统博弈层面去进行分析设计[8–17]。本文针对多无人机辅助地面通信系统在混合博弈模式下进行研究，将近端策略优化(Proximal Policy Optimization, PPO)算法扩展到多智能体强化学习(Multi-Agent Reinforcement Learning, MARL)领域，提出一种结合Beta策略的混合动作空间多智能体近端策略优化(Multi-Agent PPO, MAPPO)算法，用以联合优化GU接入策略，UAV飞行轨迹，功率及带宽分配方案，在高维决策动作下最大化系统吞吐量并满足资源分配的公平性。

2　系统模型与问题表述

本文在$\mathcal{D}$区域考虑一个多UAV辅助GU通信的系统，$\mathcal{D}\subset\mathbb{R}^3$，其中有$M$架UAV作为空中基站来为$N$个GU提供无线通信服务，定义无人机集合为$\mathcal{M}=\{1,2,\cdots,M\}$，GU集合为$\mathcal{N}=\{1,2,\cdots,N\}$，如图1所示。在此区域中，所有GU在2维平面中随机运动，以$\boldsymbol{u}_n^{\mathrm{GU}}(t),n\in\mathcal{N}$表示$t$时刻第$n$个GU的地理位置。而UAV部署在高度为$h$的空中，为每个服务的GU以频分复用(Frequency Division Multiplexing, FDM)的方式提供无线服务，并且根据服务用户的位置分布式决策自己的飞行方向和资源分配策略，用$\boldsymbol{u}_m^{\mathrm{UAV}}(t),m\in\mathcal{M}$表示$t$时刻第$m$架UAV的地理位置。$P_{\mathrm{total}}$定义为每架UAV的总发射功率，$B_{\mathrm{total}}$为所有UAV共享的总带宽资源。每架UAV在每个时隙t给服务的GU采用不同的功率及带宽资源分配策略，在连续控制条件下最大化系统吞吐量。

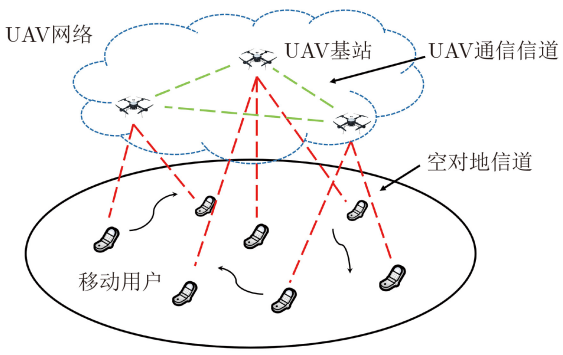


图 1 多无人机协作辅助地面通信场景

2.1　UAV空地信道模型

基于统计概率的UAV空地信道模型在许多工作中已经应用。具体来说，可以将空对地信道划分为视距(Line of Sight, LoS)和非视距(Non-Line of Sight, NLoS)出现的概率来考虑，因此t时刻第m架UAV到第n个GU信号的路径损耗$\mathrm{PL}_{m,n}(t)$可以表示为

$$\mathrm{PL}_{m,n}(t)=\mathrm{pr}_{m,n}^{\mathrm{LoS}}(t)l_{m,n}^{\mathrm{LoS}}(t)+\mathrm{pr}_{m,n}^{\mathrm{NLoS}}(t)l_{m,n}^{\mathrm{NLoS}}(t)\tag{1}$$

其中，$\mathrm{pr}_{m,n}^{\mathrm{LoS}}(t)$和$\mathrm{pr}_{m,n}^{\mathrm{NLoS}}(t)$分别为UAV m到GU n在LoS链路和NLoS链路出现的概率，以修正的Sigmoid函数表示为

$$\mathrm{pr}_{m,n}^{\mathrm{LoS}}(t)=\frac{1}{1+a\exp[-b(\varphi_{m,n}(t)-a)]}\tag{2}$$

其中，$\varphi_{m,n}(t)$为GU n与UAV m的通信仰角，a和b为与部署环境相关的参数，同时有$\mathrm{pr}_{m,n}^{\mathrm{NLoS}}(t)=1-\mathrm{pr}_{m,n}^{\mathrm{LoS}}(t)$。$l_{m,n}^{\mathrm{LoS}}(t)$和$l_{m,n}^{\mathrm{NLoS}}(t)$分别为UAV m到GU n在LoS链路和NLoS链路的路径损耗。LoS链路和NLoS链路的路径损耗是在自由空间传输模型下与附加损耗ς_{LoS}和$\varsigma_{\mathrm{NLoS}}$的组合

$$l_{m,n}^{\mathrm{LoS}}(t)=20\lg\left(\frac{4\pi f_c d_{m,n}(t)}{c}\right)+\varsigma_{\mathrm{LoS}}\tag{3}$$

$$l_{m,n}^{\mathrm{NLoS}}(t)=20\lg\left(\frac{4\pi f_c d_{m,n}(t)}{c}\right)+\varsigma_{\mathrm{NLoS}}\tag{4}$$

其中，f_c是载频，$d_{m,n}(t)=h/\sin(\varphi_{m,n}(t))$是UAV m到GU n的直线距离，c为光速。

用$p_{m,n}(t)$表示t时刻UAV m给GU n分配的发射功率，因此在GU n处可获得的接收信号功率为$p_{m,n}^{\mathrm{rec}}(t)=p_{m,n}(t)-\mathrm{PL}_{m,n}(t)$。以二进制变量$\sigma_{m,n}(t)\in\{0,1\}$表示UAV m与GU n的连接关系，当GU n选择UAV m作为接入基站时令$\sigma_{m,n}(t)=1$，反之则为0

$$\sigma_{m,n}(t)=\begin{cases}1,\text{如果GU } n\text{选择了UAV } m\\0,\text{未选择}\end{cases}\tag{5}$$

将n_0定义为高斯白噪声的功率谱密度，$b_{m,n}(t)$为t时刻UAV m给GU n分配的带宽资源，根据香农信道容量公式可以推算UAV m提供给GU n的理论通信速率为

$$c_{m,n}(t)=\sigma_{m,n}(t)b_{m,n}(t)\log_2\left(1+\frac{p^{\mathrm{rec}}_{m,n}(t)}{n_0 b_{m,n}(t)}\right) \tag{6}$$

2.2 优化问题表述

在本文中，目标是优化在$T_{\max}$的服务时间段内，所有UAV通过调整飞行轨迹和资源分配策略，最大化所有GU在该时段内的吞吐量及资源分配的公平性。以t作为一次决策的时间间隔，$t\in T=\{1,2,\cdots,T_{\max}\}$，即在每个时隙$t$的开始所有GU分别选择一架UAV作为接入点，而所有UAV在[t, t+1]时段内依据选择接入自己的GU位置，选择最合适的飞行方向以减少路径损耗，同时通过多波束成型分配GU功率和带宽资源。

对于每个GU来说，合理选择UAV接入，最大化自身可获得的总吞吐量是它们考虑的问题。本文定义GU n的优化目标为

$$\left.\begin{aligned}
\max\quad & \sum_{t\in\mathrm{T}}\sum_{m\in\mathcal{M}}c_{m,n}(t)\\
\text{s. t.}\quad & \sigma_{m,n}(t)\in\{0,\ 1\},\ \forall t\in T,\ n\in\mathcal{N}\\
& \sum_{m\in\mathcal{M}}\sigma_{m,n}(t)=1,\ \forall t\in T,\ n\in\mathcal{N}\\
& \boldsymbol{u}^{\mathrm{GU}}_n(t+1)=y[\boldsymbol{u}^{\mathrm{GU}}_n(t)],\ \forall t\in T,\ n\in\mathcal{N}\\
& \boldsymbol{u}^{\mathrm{GU}}_n(t)\in\mathcal{D},\ \forall t\in T,\ n\in\mathcal{N}
\end{aligned}\right\} \tag{7}$$

其中，$y(\boldsymbol{u}^{\mathrm{GU}}_n)$定义为GU运动的状态转移函数，所有GU在区域$\mathcal{D}$内随机运动。与此同时，每个GU在同一时刻$t$最多选择一架UAV接入，即在[$t$, t+1]时间区间内只能选择一架UAV作为接入点。因此，每个GU在t时刻的决策动作为该时隙选择接入的UAV编号。

对于每架UAV来说，t时刻会被不同数量GU选择接入，而不同的资源分配策略会给服务的GU带来不同的服务体验。为使每个GU拥有尽量相同的通信速率，UAV需要根据当前拓扑及信道条件调节功率及带宽分配策略，以保证GU之间通信速率的公平性。为衡量GU之间通信速率的差异，本文引入Jain公平指数$f_m(t)$作为对UAV m的评价指标

$$f_m(t)=\frac{\left(\sum\limits_{n\in\mathcal{N}_m}c_{m,n}(t)\right)^2}{s_m(t)\sum\limits_{n\in\mathcal{N}_m}c^2_{m,n}(t)},\ \forall t\in T,\ m\in\mathcal{M} \tag{8}$$

其中，$s_m(t)=\sum_{n\in\mathcal{N}}\sigma_{m,n}(t)$是第$m$架UAV在$t$时隙服务的GU数量，即在$t$时刻有$s_m(t)$个GU选择UAV m作为自己的接入基站，$\mathcal{N}_m$为UAV m服务的GU集合，$\mathcal{N}_m\subseteq\mathcal{N}$。$f_m(t)$越大，则代表UAV m资源分配的公平性越高，GU之间通信速率的差异越小。

与此同时，为了最大化服务时段内通信系统的总吞吐量，它们需要分布式调整自己的飞行策略，改变当前空间拓扑结构以减少总体路径损耗，在保证GU公平通信的同时最大化系统吞吐量。所有UAV有共同目标并以分布式的方式进行合作，本文定义UAV的优化目标为

$$\left.\begin{aligned}
\max\quad & \sum_{t\in T}\sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}}c_{m,n}(t),\ \sum_{t\in T}\sum_{m\in\mathcal{M}}f_m(t)\\
\text{s. t.}\quad & \sigma_{m,n}(t)\in\{0,\ 1\},\ \forall t\in T\\
& \sum_{m\in\mathcal{M}}\sigma_{m,n}(t)=1,\ \forall t\in T,\ n\in\mathcal{N}\\
& \sum_{n\in\mathcal{N}}\sigma_{m,n}(t)=s_m(t),\ \forall t\in T,\ m\in\mathcal{M}\\
& \sum_{n\in\mathcal{N}}p_{m,n}(t)=P_{\mathrm{total}},\ \forall t\in T,\ m\in\mathcal{M}\\
& \sum_{n\in\mathcal{N}}\sum_{m\in\mathcal{M}}b_{m,n}(t)=B_{\mathrm{total}},\ \forall t\in T\\
& b_{m,n}(t)\geq b_{\min},\ \sigma_{m,n}(t)=1,\ \forall t\in T,\\
& m\in\mathcal{M},\ n\in\mathcal{N}\\
& \boldsymbol{u}^{\mathrm{UAV}}_m(t+1)=g\left(\boldsymbol{u}^{\mathrm{UAV}}_m(t),\ \vartheta^m_t(t)\right),\\
& \forall t\in T,\ m\in\mathcal{M}\\
& \boldsymbol{u}^{\mathrm{UAV}}_m(t)\in\mathcal{D},\\
& \forall t\in T,\ m\in\mathcal{M}
\end{aligned}\right\} \tag{9}$$

其中，$p_{m,n}(t)$和$b_{m,n}(t)$分别为UAV m 给GU n分配的发射功率和带宽资源，$b_{\min}$为最小可分带宽，$\vartheta^m_t(t)\in[-\pi,\ \pi]$定义为UAV m在[t, t+1]间隔内的飞行方位角，$g(\boldsymbol{u}^{\mathrm{UAV}}_m,\ \vartheta^m_t(t))$定义为UAV的坐标转移函数。每架UAV在$t$时刻的决策动作包含了给服务GU的功率、带宽分配方案与自己的飞行角度。

3 多无人机辅助通信系统设计

3.1 MARL原理

不同于单智能体强化学习(Single-Agent Reinforcement Learning, SARL)，在MARL中由于奖励受到智能体联合动作的影响，并且随着智能体数量的增加，训练难度及复杂度指数增长。本文用$\mathcal{G}=(\mathcal{I},\ \mathcal{S},\ \{\mathcal{A}_i\},\ \{\mathcal{R}_i\},\ \mathcal{P},\ \gamma)$来描述随机博弈[16]，其中$\mathcal{I}$为智能体数量，$\mathcal{S}$为状态空间。$\{\mathcal{A}_i\}$, $\{\mathcal{R}_i\}$, $i\in\mathcal{I}$分别表示所有智能体的动作空间集合与奖励函数集合，$\mathcal{A}=\mathcal{A}_1\times\mathcal{A}_2\times\cdots\times\mathcal{A}_i=\mathcal{A}_i\times\mathcal{A}_{-i}$为联合动作空间，$-i$表示为除第$i$个智能体以外的其他智能体。$\mathcal{P}$是基于概率的状态转移函数，$\mathcal{P}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,\ 1]$，$\gamma\in[0,\ 1)$为折扣因子代表未来奖励与当前奖励的权衡。每个智能体拥有自己的随机策略函数$\pi_i(a_i|o_i)$以最大化长期折扣回报

$$\begin{aligned}
J_i(\pi_i,\pi_{-i})=&\mathbb{E}_{s_{t+1}\sim\mathcal{P}(s_{t+1}|s_t,a^i_t,a^{-i}_t),a^i_t\sim\pi_i(a^i_t|o^i_t)}\\
&\cdot\left[\sum_{t=0}^{\infty}\gamma^t R_i(s_t,a^i_t,a^{-i}_t)\right],s_t\in\mathcal{S},\\
&a^i_t\in\mathcal{A}_i,a^{-i}_t\in\mathcal{A}_{-i}
\end{aligned} \tag{10}$$

由于部分可观测性，$o_t^i = O_i(s_t)$表示为智能体i在环境状态s_t时的观测状态，O_i为其观测函数。a_t^i是t时刻智能体i从策略$\pi_i(a_i|o_i)$的采样动作，而策略函数通常定义为参数化的高斯分布。分别定义其状态动作值函数与状态值函数为

$$Q^i_{\pi_i,\,\pi_{-i}}(s_t,\,a_t^i,\,a_t^{-i}) = \mathbb{E}_{s_{t+1},\,a^i_{t+1},\,a^{-i}_{t+1},\cdots} \cdot\left[\sum_{l=0}^{\infty}\gamma^l R_i(s_{t+l},\,a^i_{t+l},\,a^{-i}_{t+l})\right] \tag{11}$$

$$V^i_{\pi_i,\,\pi_{-i}}(s_t) = \mathbb{E}_{a_t^i,\,a_t^{-i},\,s_{t+1},\,\cdots} \cdot\left[\sum_{l=0}^{\infty}\gamma^l R_i(s_{t+l},\,a^i_{t+l},\,a^{-i}_{t+l})\right] \tag{12}$$

于是可以用优势函数$A^i_{\pi_i,\,\pi_{-i}}(s_t,\,a_t^i) = Q^i_{\pi_i,\,\pi_{-i}}(s_t,\,a_t^i,\,a_t^{-i}) - V^i_{\pi_i,\,\pi_{-i}}(s_t)$来衡量$t$时刻动作的优劣。

在马尔可夫博弈中，用$\pi_\theta = \{\pi_{\theta_i}(a_i|s)\},\ i\in\mathcal{I}$表示智能体的联合策略，$\theta_i$为智能体$i$的策略参数。当参与博弈个体策略$\pi_{\theta_i}$达到纳什均衡时所有智能体将达到一个联合的最优策略$\pi^*_\theta = \{\pi^*_{\theta_1}(a_i|s),\ \pi^*_{\theta_2}(a_i|s),\ \cdots,\pi^*_{\theta_i}(a_i|s)\}$，此时对于任何$s\in\mathcal{S}$有$V^i_{\pi^*_{\theta_i},\,\pi^*_{\theta_{-i}}}(s) \geq V^i_{\pi_{\theta_i},\,\pi^*_{\theta_{-i}}}(s)$成立。对于策略梯度类方法，智能体$i$的策略梯度可计算为

$$\nabla_{\theta_i} J_i(\theta) = \mathbb{E}_{s_{t+1}\sim\mathcal{P}(s_{t+1}|s_t,\,a_t^i,\,a_t^{-i}),\,a_t^i\sim\pi_{\theta_i}(a_t^i|o_t^i)} \cdot\left[\nabla_{\theta_i}\ln\pi_{\theta_i}(a_t^i|o_t^i)Q_{\pi_{\theta_i}}(s_t,a_t^i,a_t^{-i})\right],\ i\in\mathcal{I} \tag{13}$$

3.2 从PPO到MAPPO

PPO作为信任域策略优化(Trust Region Policy Optimization, TRPO)算法的改良版本，在其基础上将共轭梯度法的方式简化，计算复杂度下降的同时性能并未减少，其中Actor的目标函数定义为

$$J^{\mathrm{KL}}(\theta) = \mathbb{E}_{\pi_\theta} \cdot\left\{\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}\tilde{A}_t - \beta D_{\mathrm{KL}}\left[\pi_{\theta_{\mathrm{old}}}(a_t|s_t)\parallel\pi_\theta(a_t|s_t)\right]\right\} \tag{14}$$

其中，θ为Actor网络参数，$\tilde{A}_t$为Critic估计的优势函数，$\pi_{\theta_{\mathrm{old}}}$代表收集经验的原始策略，$\pi_\theta$为利用旧策略样本更新之后的策略，$D_{\mathrm{KL}}(\pi_{\theta_{\mathrm{old}}}\parallel\pi_\theta)$代表$\pi_{\theta_{\mathrm{old}}}$与$\pi_\theta$之间的KL散度控制每轮策略更新的差异，$\beta$根据预设的KL散度阈值自适应变化调节。在后续版本中，Clip函数将概率比限制在一个合理的范围，ε为一个超参数。同样以$\dfrac{\pi_\theta}{\pi_{\theta_{\mathrm{old}}}}\tilde{A}_t$作为优化目标，$\tilde{A}_t>0$时增加$\pi_\theta(a_t|s_t)$的概率，反之则减小$\pi_\theta(a_t|s_t)$的概率

$$J^{\mathrm{clip}}(\theta) = \mathbb{E}_{\pi_\theta}\left\{\min\left[\mathrm{clip}\left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)},1-\varepsilon,1+\varepsilon\right)\tilde{A}_t,\ \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}\tilde{A}_t\right] + \psi S(s_t)\right\} \tag{15}$$

其中，$S(s_t)$表示s_t状态的策略熵，在训练阶段增加探索性可以有助于避免收敛至局部最优，ψ是平衡探索与利用的超参数。对于$\tilde{A}_t$通常采用广义优势估计(Generalized Advantage Estimation, GAE)的方式利用状态值$V(s_t)$进行估算，通过引入参数λ以加权的方式平衡估计的偏差与方差

$$\tilde{A}_t^{\mathrm{GAE}}(s_t,a_t) = \sum_{l=0}^{\infty}(\gamma\lambda)^l\sigma^V_{t+l} \tag{16}$$

$$\sigma_t^V = r_t + \gamma V(s_{t+1}) - V(s_t) \tag{17}$$

更新Critic网络的参数ω则采用回归的方式减小$\hat{V}_\omega(s_t)$的估计误差

$$J^{\mathrm{target}}(\omega) = \left[\hat{V}_\omega(s_t) - V_{\mathrm{target}}(s_t)\right]^2 \tag{18}$$

其中，$V_{\mathrm{target}}(s_t)$为依据环境反馈奖励计算出来的目标状态值。

本文采用当前主流的集中训练分布式执行(Centralised Training with Decentralised Execution, CTDE)机制与Actor-Critic框架，训练具有全局信息的集中Critic网络来估计每个智能体状态值与优势函数$\tilde{A}_t^i$，将PPO算法扩展到MARL领域。与独立近端策略优化(Independent PPO, IPPO)算法处理方式不同，在文献[18]中首次提出利用全局状态信息的MAPPO来解决合作环境下的动态博弈问题，在多种环境中取得了当前最优(State Of The Art, SOTA)的性能。而在混合博弈模式下，由于GU与UAV各自拥有不同的优化目标，因此对应的值函数不尽相同。在集中训练阶段，本文采用两个全局Critic网络分别对GU和UAV的状态值进行评估并通过GAE方法计算$\tilde{A}_t^i$。在分散执行阶段，每个GU和UAV依靠自己局部观测状态o_t^i实现分布式决策交互环境。全局Critic在集中训练时可以了解全局信息，估计每个智能体的状态值函数。然后根据每个智能体动作轨迹所获奖励R_t或者R_t^i，用GAE的方式估计一个共同的$\tilde{A}_t$或者多个各自的$\tilde{A}_t^i$。于是对于智能体i的优化目标，可以重新定义为

$$J_i(\theta) = \mathbb{E}_{\pi_\theta}\left\{\min\left[\mathrm{clip}\left(\frac{\pi_\theta(a_t^i|o_t^i)}{\pi_{\theta_{\mathrm{old}}}(a_t^i|o_t^i)},1-\varepsilon,1+\varepsilon\right)\tilde{A}_t^i,\ \frac{\pi_\theta(a_t^i|o_t^i)}{\pi_{\theta_{\mathrm{old}}}(a_t^i|o_t^i)}\tilde{A}_t^i\right] + \psi S(o_t^i)\right\} \tag{19}$$

而全局Critic依旧按照式(18)所定义的损失函数方式，用特定于智能体的全局状态s_t^i更新。

3.3 多智能体混合博弈算法设计

(1) GU状态及动作设置。由于现实约束，GU通常无法直接获取其他GU的位置及状态信息，因此它们无法判断彼此的分布情况，但UAV基站的位置及关联负载情况对GU可以开放访问。当UAV部署高度h固定时$d_{m,n}$越小路径损耗越小，因此GU会倾向于选择距离近的UAV接入。但UAV的负载情况也同样制约着GU可以得到的通信资源，于是GU为获取更高的通信速率，应该综合考虑UAV的位置与负载。本文对t时刻GU n的观测状态定义为

$$\boldsymbol{o}_t^n = \left\{\boldsymbol{u}_n^{\mathrm{GU}}, \boldsymbol{u}_m^{\mathrm{UAV}}, \{s_m(t-j)\}\right\},\ m \in \mathcal{M} \tag{20}$$

其中，第1项为GU n自己的坐标，后两项分别为所有可接入UAV的坐标与近期每架UAV服务的用户数目。在集合$\{s_m(t-j)\}$中，$j=1,2,\cdots,w$代表GU会考虑前w个时刻UAV的负载以此预测近期基站过载的概率。

在每个时隙t的开始，每个GU需要根据观测状态选择一架UAV接入。GU n的动作空间等于无人机的集合$\mathcal{M}$且为离散动作，$\mathcal{A}_n=\mathcal{M},\ n\in\mathcal{N}$。本文用One-Hot向量来表示$t$时刻GU n的动作

$$a_t^n = \{\sigma_{m,n}\},\ m \in \mathcal{M} \tag{21}$$

对于策略梯度类算法，通常由一个参数化的连续概率密度分布表示动作被选中的概率。由于GU动作是离散的，本文用Softmax函数定义GU n的输出策略

$$\pi_\theta(a_t^n|\boldsymbol{o}_t^n) = \frac{\exp\left(\chi_m(\theta)\right)}{\sum\limits_{m\in\mathcal{M}}\exp\left(\chi_m(\theta)\right)},\ m\in\mathcal{M} \tag{22}$$

其中，$\chi_m(\theta)$是参数为θ的GU策略网络输出层预激活分量，输出维度为M，并且$a_t^n\sim\pi_\theta(a_t^n|\boldsymbol{o}_t^n)$。

(2) UAV状态及动作设置。UAV作为接入基站，它们之间可以在一条额外的信道上通信以交换一些必要的信息，而GU可通过GPS上传自身位置信息。如2.2所述，UAV共享总带宽资源B_{total}，因此UAV m在t时刻可支配的带宽资源为$b_m(t)=B_{\mathrm{total}}\dfrac{s_m(t)}{N}$。于是UAV m的观测状态定义为

$$\boldsymbol{o}_t^m = \left\{\boldsymbol{u}_m^{\mathrm{UAV}}, \boldsymbol{u}_{-m}^{\mathrm{UAV}}, \boldsymbol{u}_n^{\mathrm{GU}}, \{\sigma_{m,n}\}\right\},\ n\in\mathcal{N} \tag{23}$$

其中，$\boldsymbol{u}_m^{\mathrm{UAV}}$为UAV m的坐标，$\boldsymbol{u}_{-m}^{\mathrm{UAV}}$为其他UAV的坐标。$\{\sigma_{m,n}\},n\in\mathcal{N}$定义为选择接入UAV m的GU列表，同样以向量形式表示。

UAV m在连续的动作空间中既要为不同GU分配功率及带宽，也要输出自己的飞行方位角ϑ_t^m。用$\mathcal{A}_m$表示UAV m的动作空间，$\mathcal{A}_m=\{p_{m,n}\}\times\{b_{m,n}\}\times\{\vartheta_m\},\ n\in\mathcal{N}$，定义其$t$时刻的动作为

$$\boldsymbol{a}_t^m = \{p_{m,n}(t),\ b_{m,n}(t),\ \vartheta_t^m\},\ n\in\mathcal{N} \tag{24}$$

用$\bar{p}_{m,n}(\phi),\ \bar{b}_{m,n}(\phi)$分别表示参数为$\phi$的UAV策略网络关于功率及带宽的采样动作输出，其维度均为N。由于在不同时刻UAV服务用户数目不同，并且需要将全部资源分给服务用户并满足最小带宽条件。定义$\kappa_m^{\mathrm{mask}}=\{\sigma_{m,n}\},\ n\in\mathcal{N}$作为UAV m的动作掩码，用来屏蔽掉无关用户维度上的信息。令$\tilde{p}_{m,n}(\phi)=\bar{p}_{m,n}(\phi)\kappa_m^{\mathrm{mask}}$, $\tilde{b}_{m,n}(\phi)=\bar{b}_{m,n}(\phi)\kappa_m^{\mathrm{mask}}$，并用$\{x_i\}$与$\{x_j\}$表示$\tilde{p}_{m,n}(\phi)$和$\tilde{b}_{m,n}(\phi)$中非0元素的集合，其中$i,\ j\in[1,\ s_m]$，于是$\boldsymbol{a}_t^m$中$p_{m,n}(t)$与$b_{m,n}(t)$动作分量可以表述为

$$p_{m,n}(t) = \begin{cases} P_{\mathrm{total}}\dfrac{\exp(x_i)}{\sum\limits_{i=1}^{s_m}\exp(x_i)}, & \sigma_{m,n}(t)=1 \\ 0, \text{其它} \end{cases} \tag{25}$$

$$b_{m,n}(t) = \begin{cases} \left(b_m(t)-s_m(t)b_{\min}\right)\dfrac{\exp(x_j)}{\sum\limits_{j=1}^{s_m}\exp(x_j)}+b_{\min}, \\ \qquad \sigma_{m,n}(t)=1 \\ 0, \text{其它} \end{cases} \tag{26}$$

ϑ_t^m可以由归一化的采样输出$\bar{\vartheta}_t^m$放大得到，$\vartheta_t^m=\pi\bar{\vartheta}_t^m$，其中$\bar{\vartheta}_t^m$是由Tanh函数激活所得均值生成的高斯分布的采样值。值得注意的是，在算法中计算动作概率及策略熵时同样需要以κ_m^{mask}屏蔽掉无关维度上的信息。

(3)奖励函数设置。由式(7)中优化目标可知，每个GU都独立追求自身的长期累计奖励，于是可以直接定义GU n的奖励函数为

$$R_n(t) = r_c\sum_{m\in\mathcal{M}}c_{m,n}(t) \tag{27}$$

其中r_c是奖励系数。

所有UAV有共同目标，即最大化所有GU的总吞吐量。同时在每个时刻t，UAV需要合理分配功率及带宽以保证资源调度的公平性。本文用$f_m(t)$作为UAV m个体策略的评价指标，由于$f_m(t)$达到一定高度后提升困难，因此用其指数形式增加不同$f_m(t)$下的奖励差异。$c_{\mathrm{mean}}=\dfrac{1}{N}\sum_{m\in\mathcal{M}}\sum_{n\in\mathcal{N}}c_{m,n}$作为UAV联合动作的即时回报，以此项来促使UAV合作，于是UAV m的奖励函数定义为

$$R_m(t) = r_d f_m(t)^{\kappa_r}c_{\mathrm{mean}}(t) - R_m^b(t) \tag{28}$$

其中，r_d是奖励系数，κ_r为$f_m(t)$的指数参数。而$R_m^b(t)$作为UAV m的边界惩罚项，当UAV越过边

界约束时会受到相应惩罚。但是当违反约束时直接给定一个较大的负奖励会导致阶跃的奖励函数，继而带来爆炸的更新梯度。这很容易导致UAV畏惧惩罚而不敢探索，为使策略梯度更新更加平滑，定义$R_m^b(t)$为连续函数

$$R_m^b(t)=r_b\left(\frac{1}{1+\exp\left[-\kappa_b(u_m-u_{\max})\right]}+\frac{1}{1+\exp\left[\kappa_b(u_m-u_{\min})\right]}\right) \tag{29}$$

其中，r_b是惩罚项系数，κ_b是梯度因子用来决定边界函数的平缓程度。$u_m\in\boldsymbol{u}_m^{\mathrm{UAV}}$为UAV m的坐标，而$u_{\max}$和$u_{\min}$分别是边界的上界与下界坐标。

(4)通信系统全局状态设置。如3.2所述，全局Critic需要估计当前全局状态s_t的状态值$\hat{V}_\omega(s_t)$用于GAE估算$\tilde{A}_t$。值得注意的是，为了避免全局状态输入维度过高并保留所有有效信息，本文没有采取常用的直接拼接所有智能体观测的方式。而是针对当前通信系统的全局信息，设置属于每个智能体的全局状态$\boldsymbol{s}_t^i,\ i\in\{\mathcal{M}\cup\mathcal{N}\}$

$$\boldsymbol{s}_t^m=\left\{\boldsymbol{u}_m^{\mathrm{UAV}},\ \boldsymbol{u}_{-m}^{\mathrm{UAV}},\ \boldsymbol{u}_n^{\mathrm{GU}},\ c_n,\ \{p_{m,n},b_{m,n}\}_{m\in\mathcal{M}}\right\},\quad n\in\mathcal{N} \tag{30}$$

其中，$\boldsymbol{s}_t^m$表示为UAV m的全局状态，c_n为第n个GU的通信速率。同样的方式，可以定义GU n的全局状态为$\boldsymbol{s}_t^n$

$$\boldsymbol{s}_t^n=\left\{\boldsymbol{u}_n^{\mathrm{GU}},\ \boldsymbol{u}_{-n}^{\mathrm{GU}},\ \boldsymbol{u}_m^{\mathrm{UAV}},\ \{c_n\}_{n\in\mathcal{N}},\ \{p_{m,n},b_{m,n}\}_{n\in\mathcal{N}}\right\},\quad m\in\mathcal{M} \tag{31}$$

3.4 采用Beta分布的UAV策略网络

对于连续动作处理，通常是以参数化的高斯分布拟合策略函数并进行采样得到执行动作，$a_t\sim\mathbb{N}(\mu_\phi,\sigma_\phi^2)$，$\mu_\phi$与$\sigma_\phi^2$分别是均值和方差。但在大多现实问题中，动作具有边界约束。例如UAV输出的飞行方位角应该被限制在[−π, π]，而高斯分布是定义在[−∞, ∞]。在此类场景中高斯分布不可避免会由于动作截断的边界效应而引入概率估计的偏差[19]，高斯分布会赋予所有动作大于0的概率密度，而在现实中边界外的动作概率应该为0。

在高斯策略的一般实现中，当采样的动作发生在边界外时，它会被环境强行裁剪为区间内的边界值以满足约束条件。当用裁剪前的动作计算其对数似然概率时，策略梯度的估计量为$\mathbb{E}_{\pi_\phi}[\nabla_\phi\ln\pi_\phi(a|s)Q_{\pi_\phi}(s,a_{\mathrm{clip}})]$，其中$a_{\mathrm{clip}}$代表裁剪后的动作，$\phi$为UAV策略网络参数。此时策略梯度估计量的偏差$\Delta p$可表示为

$$\begin{aligned}\Delta p=&\nabla_\phi J'(\pi_\phi)-\nabla_\phi J(\pi_\phi)\\=&\mathbb{E}\left[\int_{-\infty}^{\infty}\pi_\phi(a|s)\nabla_\phi\ln\pi_\phi(a|s)Q_{\pi_\phi}(s,a_{\mathrm{clip}})\,\mathrm{d}a\right]\\&-\mathbb{E}\left[\int_{-\infty}^{\infty}\pi_\phi(a|s)\nabla_\phi\ln\pi_\phi(a|s)Q_{\pi_\phi}(s,a)\,\mathrm{d}a\right]\\=&\mathbb{E}\left[\int_{-\infty}^{-k}\pi_\phi(a|s)\nabla_\phi\ln\pi_\phi(a|s)[Q_{\pi_\phi}(s,-k)\right.\\&-Q_{\pi_\phi}(s,a)]\,\mathrm{d}a+\int_k^{\infty}\pi_\phi(a|s)\nabla_\phi\ln\pi_\phi(a|s)\\&\left.\cdot[Q_{\pi_\phi}(s,k)-Q_{\pi_\phi}(s,a)]\,\mathrm{d}a\right]\end{aligned} \tag{32}$$

假设动作的边界约束为[–k, k]，当a_{clip}被裁剪到–k或者k时，$\Delta p\neq 0$，即此时产生了策略梯度的偏差。如果用裁剪后的动作计算其对数概率，此时$\nabla_\phi\ln\pi_\phi(a|s)Q_{\pi_\phi}(s,a)$变成了$\nabla_\phi\ln\pi_\phi(a_{\mathrm{clip}}|s)Q_{\pi_\phi}(s,a_{\mathrm{clip}})$同样带有偏差。

使用高斯策略的另外一个弊端，则是在多峰奖励环境下，由于次优峰可能占据更大的概率密度，继而容易导致次优收敛。这与策略网络的参数初始化密不可分，但在训练前期通常不具备对于最优策略的先验信息，因此无法设置一个最佳的参数初始化。

Beta分布是定义在[0, 1]区间上的概率密度分布，由参数α与β共同决定其形状

$$f(x,\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{33}$$

其中，$\Gamma(\cdot)$为伽马函数。用参数化的α_ϕ和β_ϕ表示Beta策略，$\pi_\phi(a|s)=Be(\alpha_\phi,\beta_\phi)$，此时$a_t\sim\mathrm{Be}(\alpha_\phi,\beta_\phi)$。在Beta分布下，策略由一个有限区间的概率密度分布进行描述，避免了上述梯度估计偏差问题。在策略网络参数初始化时，通过让$\alpha_\phi\approx\beta_\phi\approx 1$近似为均匀分布，可以让智能体在初始阶段更加随机地探索环境，从而缓解高斯分布因不良初始化而造成的局部最优收敛。

4 算法训练流程及仿真实验设置

4.1 算法训练流程

由于所有GU具有同质性，并且在达到纳什均衡时所有GU应保持同样的接入策略，因此可以设置一个GU共享的策略网络可以加快收敛速度，其参数记为θ。同理所有UAV可以共享同一策略网络，其参数记为ϕ，GU与UAV的全局Critic网络参数分别记为ω_1与ω_2。

每个回合开始时，所有UAV从出发点起飞，而GU随机分布在区域$\mathcal{D}$并以随机的方向和速度运动。在每个t时刻，所有GU将自身观测$\boldsymbol{o}_t^n$输入参数为θ的GU策略网络，得到选项的离散概率分布并采样确定接入的UAV编号。当所有GU上传自身接入

请求后，UAV将自身观测$\boldsymbol{o}_t^m$输入到参数为ϕ的UAV策略网络，得到连续概率密度分布并采样动作。此时环境依据奖励函数反馈所有智能体的单步奖励$r_t^i,\ i\in\{\mathcal{M}\cup\mathcal{N}\}$，并且所有智能体的信息$(\boldsymbol{o}_t^i,\ \boldsymbol{a}_t^i,\ \boldsymbol{s}_t^i,\ r_t^i, \mathrm{pr}_t^i)$将被保存在一个经验缓存区$\mathcal{B}$中，其中$\mathrm{pr}_t^i$定义为旧策略下采样动作的概率。当满足设定的环境交互次数后，所有经验被一次性取出分批次更新各部分网络参数。在训练过程中由于使用重要性采样的技巧，可以将数据分成若干批次并重复使用多次，直到策略更新到信任区域边界。具体训练流程如表1所示。

4.2 实验环境及参数设置

在本文中，区域$\mathcal{D}$设置为2 ×2 km的正方形。并以此构建直角坐标系，因此区域边界坐标$u_{\max}=2000,\ u_{\min}=0$。所有UAV均部署在高度$h=500$ m的空中，每个回合开始所有UAV统一从坐标原点起飞，速度为15 m/s。每个GU以[0 m/s, 5 m/s]区间的速度，以随机的轨迹运动。所有实验中每架UAV额定发射功率$P_{\mathrm{total}}=10$ dBm，共享总带宽$B_{\mathrm{total}}=30$ MHz，通信载频$f_c=2$ GHz，噪声功率谱密度$n_0=-170$ dBm/Hz，最小可分带宽$b_{\min}=0.1$ MHz。

假定每回合的服务持续时间为$T_{\max}=1000$ s，每个决策时间间隔为1 s。根据研究表明，使用更多数据估计策略梯度可以更容易获得策略提升，并且使用较低的数据复用次数能够避免性能损失及过拟合[18,20,21]。为此，权衡收敛速度与算法性能，使用4096作为经验缓存区$\mathcal{B}$的大小，分4个Mini_batch复用5次更新网络参数。在实现Beta策略参数输出时采用Softplus函数激活后与1相加的方式，来近似初始的均匀分布策略以促进探索。此外根据文献[20]的部分建议，对于高斯策略MAPPO的实现采用了正交初始化作为层权值的初始方案。其余算法相关参数设置如表2所示。

4.3 实验结果及对比分析

为验证所提混合博弈模式下多无人机辅助地面通信的Beta-MAPPO算法性能及所提改进的有效性，本文将以下3种算法作为基线进行对照：

(1) Ga-MAPPO：采用高斯策略的MAPPO算法，但是遵循实现Beta-MAPPO时所采用的其他所有技巧。

(2) NM-MAPPO：同样采用Beta策略实现，但不使用文中所提动作掩码的方案，即在计算概率与策略熵时不屏蔽无关用户维度信息。

(3) IPPO：采用Beta策略的普通分布式PPO算法实现，每个智能体仅依靠本地观测得到其状态值，并且相互独立学习。

此外，在实现上述算法时均采用了相同的网络结构及超参数配置。

表 1 多无人机辅助通信的Beta-MAPPO算法

输入：初始化各类神经网络参数，经验缓存区$\mathcal{B}$，以及其余相关参数、超参数
输出：训练完善的GU, UAV策略网络

```
for episode do
    初始化经验缓存B以及所有GU与UAV位置
    for step t do
    所有GU根据当前状态o_t^n选择UAV接入
    所有UAV根据当前状态o_t^m选择飞行方向和资源分配方案
    各自得到奖励反馈r_t^i，储存经验元组(o_t^i, a_t^i, s_t^i, r_t^i, pr_t^i)到B，然后转移到下一个状态s_{t+1}
    if buffer B is full then
        for all agent i do
            用GAE方式计算优势函数Ã_t^i与目标状态值函数V_target(s_t^i)
            for update epoch n_r do
                从B中采样数据依据式(19)更新Actor网络参数θ或φ
                依据式(18)更新Critic网络参数ω1或ω2
            end for
        end for
        清空经验缓存区B
        end if
        end for
    end for
```

图2—图4绘制了以上所提算法训练过程中每回合累计奖励与平均公平指数变化情况，其中$M=3$, $N=20$，每个数据点为20个回合的最大值。根据结果显示，Beta-MAPPO表现出优于其他几种基线的性能，能够较为稳定收敛到较高得分。在训练初始阶段，Beta策略通过初始化能够近似为均匀分布，使得探索更加随机以获取更多样性的经验样本，从而避免次优收敛。而高斯分布虽然能以较大方差的形式近似初始化为随机策略，但是在动作有界条件下高斯分布的截断效应会更加严重，更多的概率密度分布将会处于动作边界之外，因此本文采用了正交初始化方案[20]。而在高斯策略下，由于动作边界效应的影响，截断高斯分布所产生的策略梯度误差导致算法性能损失从而效率降低。在同等训练回合下高斯策略的表现不如Beta策略。而NM-MAPPO算法由于不采用所提动作掩码方案，在计算动作概率及策略熵时包含了无关维度用户信息。

表 2　算法相关参数

参数名称	值
裁剪参数ε	0.2
熵奖励系数ψ	0.01
GAE参数λ	0.97
折扣因子γ	0.998
数据复用次数n_r	5
优化器	Adam
GU激活函数	Tanh, Softmax
UAV激活函数	Tanh, Softplus
Mini_batch数量	4
缓存区大小	4096
学习率	3e-4
奖励系数r_c	1
奖励系数r_d	2
考虑前w个时刻的UAV负载	3
指数参数κ_r	5
UAV边界惩罚函数参数r_b	20
UAV边界惩罚函数参数κ_b	8e-2

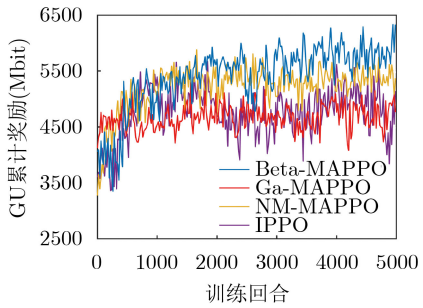


图 2 GU累计奖励变化对比

而这些无关信息与奖励的获取不直接相关甚至会干扰最优策略的学习，虽然在较大batch_size设置下能够减轻这一影响，但是依旧效率不如Beta-MAPPO。在IPPO算法的实现中，每个智能体仅依靠自身观测状态独立学习，是PPO算法到多智能体领域的简单扩展。虽然训练前期IPPO可以快速提升策略，但是独立学习带来的非平稳性直接导致了算法性能损失及学习曲线震荡。

从图2结果来看，Beta-MAPPO最终能够使多无人机辅助的通信系统在服务时段内每个GU平均吞吐量收敛在6100 Mbit左右，均高于其他基线算法。虽然GU的策略学习方式在上述算法中均未改变，但是在混合博弈模式下由于UAV与GU策略互相影响，UAV不同的飞行策略与资源分配方式直接影响了每个GU可获取的通信速率。如图3所示，每架UAV的累计奖励收敛在8300左右，同时平均每架UAV资源分配的公平指数收敛约为0.93。而Ga-MAPPO在训练后期各项得分呈现出了上升趋势，但是效率相对较低，为解决无先验知识情况下不良初始化与动作截断效应的影响，它需要更多的数据去训练策略。同时，公平指数由于越靠近1提升越困难，而且在通信双方均是动态移动的情况下，导致了图4中训练后期公平指数提升愈发缓慢的现象。

图5展示了其中3种随机测试场景下3架UAV在适应20个GU动态拓扑条件下的飞行策略，其中蓝色为GU运动轨迹，红色为UAV飞行轨迹，五角星

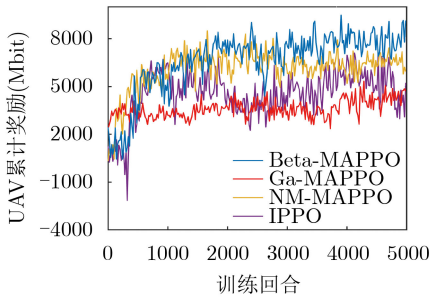


图 3 UAV累计奖励变化对比

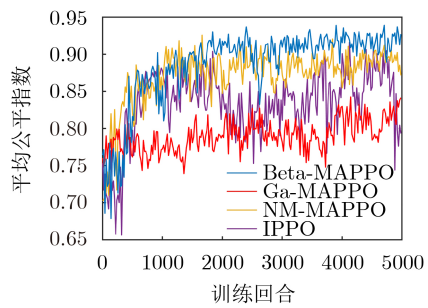
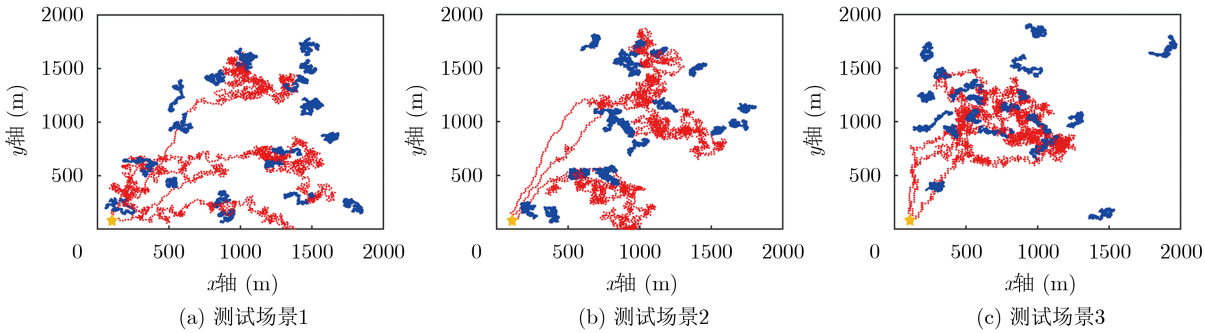


图 4 平均公平指数变化对比



(a) 测试场景1 (b) 测试场景2 (c) 测试场景3

图 5 UAV与GU运动轨迹

为出发点。可以发现在服务初始阶段，UAV可以分布式快速飞向GU区域。训练完善后的每架UAV不光会考虑选择接入自己的GU位置，同时还会考虑其它UAV的飞行策略合作完成部署，其飞行轨迹能够尽量覆盖大部分的GU，并且能够在GU集中的区域盘旋服务。除此之外，UAV的飞行策略还与GU的分布特性有关。在图5(c)所示情况下，某些极少数GU位置偏远，如果UAV飞向它们则会带来更多路径损耗，因此UAV会选择盘旋在绝大多数GU的区域，并通过资源调度来保证偏远地区GU的通信公平性。

为验证Beta-MAPPO在不同GU场景下的性能，图6绘制了在一个测试回合中，各时刻GU平均速率的变化情况。随着GU数量增加，在总资源一定条件下每个GU的平均通信速率势必会减少。在该服务周期内，一开始由于UAV从出发点起飞，而GU随机散布在区域$\mathcal{D}$，此时GU的平均通信速率较小。在前约100 s内UAV迅速部署到区域中，在适应GU选择的同时快速减小系统总体路径损耗，并通过合理调度资源使得GU平均速率快速稳定在一个较高水平。但由于该通信系统的动态性，GU的平均通信速率不可避免会产生波动，这是因为UAV需要实时调整自己的飞行策略以适应GU的选择和移动。此外，GU的平均速率与它们的实际分布情况有较大关系，当GU分布较为密集时，UAV传输信号的能量损失更少可以带来更高的通信速率；而当GU分布较为稀疏时，UAV无法照顾到每一个选择自己的GU，只能均匀减少到每个GU的距离。

图7将所提算法与基线算法所得策略在不同GU场景下进行了对照，所有实验均采用了3架UAV。如图所示，在不同GU场景下，Beta-MAPPO表现出优于其它基线算法的性能。随着GU数量增加总吞吐量由于总路径损耗的增加而略微减少，因为UAV无法同时靠近每一个选择自己的GU，但Beta-MAPPO的减少趋势相较其它方法更为平缓，

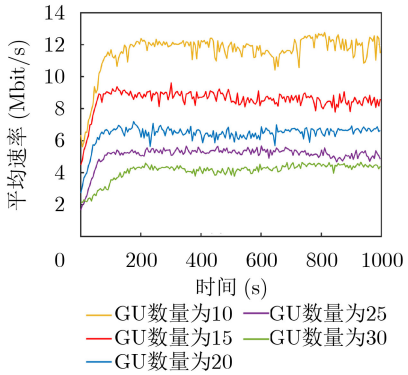
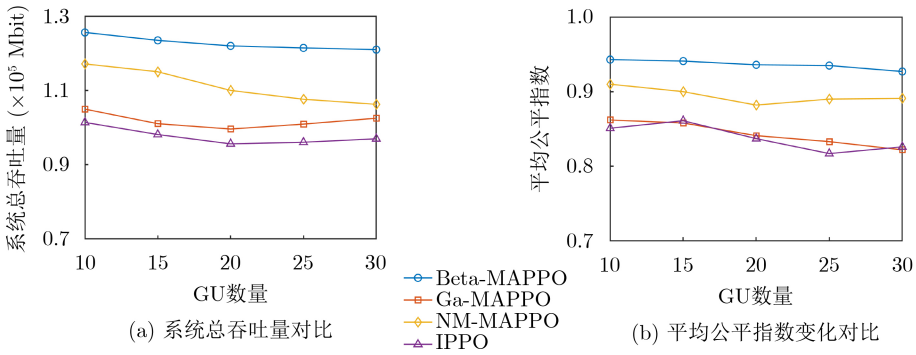


图 6 在不同GU场景下服务时段内GU平均速率变化情况

证明了其更加鲁棒。除此之外，Ga-MAPPO和IPPO方法得到的总吞吐量较低，因为它们在有限训练回合下只收敛到了次优飞行策略，无法很好应对GU的随机动态拓扑进行部署，因此总吞吐量方面会存在一定波动性和随机性。而在反映资源分配公平性的平均公平指数方面，Beta-MAPPO方法更接近于全局最优解。但随着GU数量增加，需要决策的动作维度进一步增长，资源分配的难度增加导致了平均公平指数略微下降，但依旧高于其它方法。而Ga-MAPPO和IPPO在资源分配方面表现出了相似的性能，一方面是因为它们都处于次优策略导致结果的随机性较大，另一方面是由于在UAV的资源分配策略中与其它UAV合作的成分较少，UAV只需关注选择接入自己的GU信息即可获得该部分奖励，CTDE方法对于该部分策略的学习优势不明显。而NM-MAPPO由于在计算动作概率和策略熵时包含了所有GU的信息，随着GU数量增加无用信息的干扰更加明显，因此在优化吞吐量方面下降较快。

5 结束语

针对多无人机辅助地面通信场景，本文从多智能体混合博弈层面提出了一种采用Beta策略的MAPPO算法解决GU与UAV动态博弈条件下优化通信系统吞吐量及公平通信问题。通过同时优化



(a) 系统总吞吐量对比　　(b) 平均公平指数变化对比

图 7 不同数量GU场景下系统总吞吐量及平均公平指数对比

GU接入策略、UAV多维资源分配以及飞行轨迹设计，将PPO算法在信任区域内稳定提升策略的优势扩展到混合博弈模式下的MARL领域，使得GU与UAV策略在互相适应调整的同时逼近纳什均衡，分布式决策的方式将高维的联合动作空间在不同智能体层面解耦，相比SARL集中决策解决多智能体问题的方式，大为减少了决策的动作维度。最后通过仿真实验验证了所提算法的有效性。未来的工作将研究在异构多智能体环境下，解决更复杂通信系统的联合博弈优化问题。

参 考 文 献

[1] YOU Xiaohu, WANG Chengxiang, HUANG Jie, *et al*. Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts[J]. *Science China Information Sciences*, 2021, 64(1): 110301. doi: 10.1007/s11432-020-2955-6.

[2] SEKANDER S, TABASSUM H, and HOSSAIN E. Multi-tier drone architecture for 5G/B5G cellular networks: Challenges, trends, and prospects[J]. *IEEE Communications Magazine*, 2018, 56(3): 96–103. doi: 10.1109/MCOM.2018.1700666.

[3] MISHRA D and NATALIZIO E. A survey on cellular-connected UAVs: Design challenges, enabling 5G/B5G innovations, and experimental advancements[J]. *Computer Networks*, 2020, 182: 107451. doi: 10.1016/j.comnet.2020.107451.

[4] ZHAO Jianwei, LIU Jun, JIANG Jing, *et al*. Efficient deployment with geometric analysis for mmWave UAV communications[J]. *IEEE Wireless Communications Letters*, 2020, 9(7): 1115–1119. doi: 10.1109/LWC.2020.2982637.

[5] LI Bin, FEI Zesong, and ZHANG Yan. UAV communications for 5G and beyond: Recent advances and future trends[J]. *IEEE Internet of Things Journal*, 2019, 6(2): 2241–2263. doi: 10.1109/JIOT.2018.2887086.

[6] 赵太飞, 林亚茹, 马倩文, 等. 无人机编队中无线紫外光隐秘通信的能耗均衡算法[J]. 电子与信息学报, 2020, 42(12): 2969–2975. doi: 10.11999/JEIT190965.

ZHAO Taifei, LIN Yaru, MA Qianwen, *et al*. Energy balance algorithm for wireless ultraviolet secret communication in UAV formation[J]. *Journal of Electronics & Information Technology*, 2020, 42(12): 2969–2975. doi: 10.11999/JEIT190965.

[7] WANG Yining, CHEN Mingzhe, YANG Zhaohui, *et al*. Deep learning for optimal deployment of UAVs with visible light communications[J]. *IEEE Transactions on Wireless Communications*, 2020, 19(11): 7049–7063. doi: 10.1109/TWC.2020.3007804.

[8] 徐勇军, 刘子腱, 李国权, 等. 基于NOMA的无线携能D2D通信鲁棒能效优化算法[J]. 电子与信息学报, 2021, 43(5): 1289–1297. doi: 10.11999/JEIT200175.

XU Yongjun, LIU Zijian, LI Guoquan, *et al*. Robust energy efficiency optimization algorithm for NOMA-based D2D communication with simultaneous wireless information and power transfer[J]. *Journal of Electronics & Information Technology*, 2021, 43(5): 1289–1297. doi: 10.11999/JEIT200175.

[9] ZHAN Cheng, ZENG Yong, and ZHANG Rui. Trajectory design for distributed estimation in UAV-enabled wireless sensor network[J]. *IEEE Transactions on Vehicular Technology*, 2018, 67(10): 10155–10159. doi: 10.1109/TVT.2018.2859450.

[10] SHEN Xinyue and GU Yuantao. Nonconvex sparse logistic regression with weakly convex regularization[J]. *IEEE Transactions on Signal Processing*, 2018, 66(12): 3199–3211. doi: 10.1109/TSP.2018.2824289.

[11] CUI Fangyu, CAI Yunlong, QIN Zhijin, *et al*. Multiple access for mobile-UAV enabled networks: Joint trajectory design and resource allocation[J]. *IEEE Transactions on Communications*, 2019, 67(7): 4980–4994. doi: 10.1109/TCOMM.2019.2910263.

[12] SAWALMEH A, OTHMAN N S, SHAKHATREH H, *et al*. Providing wireless coverage in massively crowded events using UAVs[C]. 2017 IEEE 13th Malaysia International Conference on Communications (MICC), Johor Bahru, Malaysia, 2017: 158–163. doi: 10.1109/MICC.2017.8311751.

[13] SHAKHATREH H, KHREISHAH A, ALSARHAN A, *et al*.

Efficient 3D placement of a UAV using particle swarm optimization[C]. 2017 8th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2017: 258–263. doi: 10.1109/IACS.2017.7921981.

[14] SAXENA V, JALDÉN J, and KLESSIG H. Optimal UAV base station trajectories using flow-level models for reinforcement learning[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2019, 5(4): 1101–1112. doi: 10.1109/TCCN.2019.2948324.

[15] LIU Xiao, LIU Yuanwei, and CHEN Yue. Reinforcement learning in multiple-UAV networks: Deployment and movement design[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(8): 8036–8049. doi: 10.1109/TVT.2019.2922849.

[16] WANG Qiang, ZHANG Wenqi, LIU Yuanwei, *et al*. Multi-UAV dynamic wireless networking with deep reinforcement learning[J]. *IEEE Communications Letters*, 2019, 23(12): 2243–2246. doi: 10.1109/LCOMM.2019.2940191.

[17] CAO Yang, ZHANG Lin, and LIANG Yingchang. Deep reinforcement learning for multi-user access control in UAV networks[C]. ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019: 1–6. doi: 10.1109/ICC.2019.8761794.

[18] YU Chao, VELU A, VINITSKY E, *et al*. The surprising effectiveness of PPO in cooperative, multi-agent games[J]. arXiv preprint arXiv: 2103.01955, 2021.

[19] CHOU P W, MATURANA D, and SCHERER S. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the Beta distribution[C]. Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 2017: 834–843.

[20] ENGSTROM L, ILYAS A, SANTURKAR S, *et al*. Implementation matters in deep policy gradients: A case study on PPO and TRPO[C]. International Conference on Learning Representations(ICLR), Addis Ababa, Ethiopia, 2020: 1–14.

[21] SMITH S L, KINDERMANS P J, YING C, *et al*. Don't decay the learning rate, increase the batch size[C]. International Conference on Learning Representations (ICLR), Vancouver, Canada, 2018: 1–11.

吴官翰：男，1993年生，助理工程师，研究方向为无人机通信组网和深度强化学习.

贾维敏：女，1971年生，教授，研究方向为遥测技术.

赵建伟：男，1989年生，讲师，研究方向为无人机蜂群通信.

高飞飞：男，1980年生，副教授，研究方向为通信原理和智能信号处理.

姚敏立：男，1966年生，教授，研究方向为卫星动中通技术.

责任编辑：余 蓉