doi:10.12068/j.issn.1005－3026.2025.20230252

基于MASAC最大熵强化学习的跳波束卫星系统资源适配方案

王译萱，刘 军
（东北大学 计算机科学与工程学院，辽宁 沈阳 110169）

摘 要：针对跳波束卫星系统中通信终端多样化的业务需求导致星-地资源供需失配，以及上行传输中机器类终端能量资源受限的挑战，提出一种基于MASAC（multi-agent soft actor-critic）最大熵强化学习的资源适配方案．首先构建了两阶段传输系统模型，在星-地资源供需失配问题的基础上，研究跳波束与非正交多址接入（non-orthogonal multiple access，NOMA）的协同作用．同时，引入能量采集与收集机制，优化了终端设备能量采集与信号传输之间的关系．在此基础上，将上下行传输过程进行整合，建立跳波束图样选择，时隙分配以及速率与功率控制的多目标优化问题，并采用MASAC算法进行优化求解，得到最优联合控制方案．实验结果表明，所提方案能够有效进行资源分配以实现星-地资源供需匹配，并满足能量受限终端的信号传输需求．与基准算法相比，所提算法具有良好的性能．

关 键 词：跳波束卫星；非正交多址；能量收集；资源适配；深度强化学习
中图分类号：TP 915 **文献标志码**：A **文章编号**：1005-3026(2025)02-0009-09

Resource Adaptation Scheme for Beam-Hopping Satellite System Based on MASAC Maximum Entropy Reinforcement Learning

WANG Yi-xuan, *LIU Jun*
(School of Computer Science & Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: WANG Yi-xuan, E-mail: 18742066986@163.com)

Abstract: To address the mismatch between space-to-ground resources supply and demand caused by the diversified traffic requirements of communication terminals in the beam-hopping satellite system, as well as the challenge of limited energy resources of machine-type devices in upward transmission, a resource adaptation scheme is proposed based on a multi-agent soft actor-critic (MASAC) approach utilizing maximum entropy reinforcement learning. Firstly, a two-stage transmission system model is constructed to investigate the synergistic effect of beam-hopping and non-orthogonal multiple access (NOMA) on the basis of the space-to-ground resource mismatch problem. Additionally, an energy harvesting and collection mechanism is introduced to optimize the relationship between terminal device energy harvesting and signal transmission. On this basis, a multi-objective optimization problem is established for beam-hopping pattern selection, time slot allocation, and rate and power control by integrating the uplink and downlink transmission processes. MASAC maximum entropy reinforcement learning is employed for optimization, obtaining an optimal joint control strategy. Experimental results show that the proposed scheme can effectively allocate resources for space-to-ground resource matching and meet the signal transmission requirements of energy-constrained machine terminals. Compared with the benchmark algorithm, the proposed algorithm exhibits superior performance.

Key words: beam-hopping satellite; non-orthogonal multiple access (NOMA); energy harvesting;

收稿日期：2023-08-29
基金项目：国家自然科学基金资助项目(61671141).
作者简介：王译萱(1999—)，女，河南郑州人，东北大学硕士研究生；刘 军(1969—)，男，辽宁沈阳人，东北大学教授．

resource allocation; deep reinforcement learning

随着全球物联网产业进入爆发式的发展时期,第三代合作伙伴计划(3rd generation partnership project,3GPP)已正式开始研究卫星通信与5G新无线电技术之间的集成,包括窄带物联网技术和面向机器类型通信的长期演进(long term evolution,LTE)技术[1].基于卫星的机器对机器(machine to machine,M2M)通信引起了越来越多研究者和研究机构的关注[2].

然而,基于卫星的M2M通信存在以下2个问题.首先,机器类通信终端具有多样化的业务类型及不同的服务质量需求,导致其所在的卫星波束小区间数据流量请求差异性大,在时间和空间上分布不均衡[3],致使卫星无法将所提供的星载容量与波束小区之间的异构请求流量分布相匹配,从而造成星-地资源供需失配[4].其次,机器类通信终端主要依赖于嵌入内部的微型电池供能,但在某些实际应用场景中,如智慧城市、环境监测、智能家居等以传感和数据采集为目标的应用场景,电池的更换成本较高或不能更换[5].这些能量受限机器类终端的接入,对其信号传输过程中的能量资源供给提出了新要求[6].因此,针对上述问题,本文提出一种基于跳波束(beam-hopping,BH)卫星系统的资源适配方案,以实现星-地资源供需匹配,并且满足能量受限的机器类设备的信号传输[7]需求.

1 相关工作

如何灵活高效地进行卫星资源分配已成为研究热点.传统方法采用固定资源分配方式,难以适应通信需求量动态变化的特性,极易造成资源的浪费.为了克服固定分配的缺点,各种动态资源分配算法应运而生.文献[8]提出了一种两阶段遗传算法和模拟退火算法来分配波束功率.文献[9]针对分布式卫星中资源有限以及能源效率低的问题,建立了功率频谱联合分配模型,提出基于凸优化理论的能效资源分配算法.随着相控阵天线技术的发展,跳波束技术已经被应用到多波束卫星系统中,其具有优越的灵活性、高效的资源利用率以及适应地面业务动态变化的能力[10].文献[11]提出了一种启发式算法来提高BH系统的容量;文献[12]根据整体流量需求设计卫星跳波束模式,进而实现BH的联合功率和带宽分配.

为了实现更高的频谱效率和边缘吞吐量,非正交多址(NOMA)接入技术支持在同一频谱/时间资源上多个设备的接入,有望在无线网络中提供高频谱效率和支持大规模连接的海量通信需求[13].现有研究将跳波束技术与NOMA相结合,使系统可在功率域进行不同用户的信号复用.文献[14]首次研究了多波束卫星系统NOMA和BH的潜在协同作用.为了缓解卫星提供容量和波束请求流量之间的失配问题,采用贪心算法解决联合BH调度和基于NOMA的功率分配问题.

以上研究主要基于传统算法和智能优化算法对于卫星资源进行分配.然而,随着卫星可用波束增加,存在计算时间长、算法复杂度高的缺点,在需求不断变化的情况下难以实现资源的快速动态分配[14-15].

随着人工智能技术的快速发展,深度强化学习(deep reinforcement learning,DRL)在信息领域得到了广泛的应用,为卫星的资源配置提供了一种新的方法.文献[16]提出了一种基于DRL的近端策略优化方法,通过动态分配卫星波束的功率,满足用户请求容量和功率有效利用率的需求.文献[15]研究了卫星系统中的联合跳波束选择和带宽分配问题,利用多智能体DRL解决启发式算法在卫星资源分配方面收敛速度慢和模型泛化能力弱的问题.基于策略梯度的强化学习方法目前成为研究的热点方向,深度确定性策略梯度(deep deterministic policy gradient,DDPG)已被广泛应用于解决资源分配问题[17-18].DDPG改进了启发式算法导致的计算复杂度相对较高的情况,提高了模型的泛化能力,但其存在探索能力弱、容易陷入局部最优的缺点,当智能体超过一定数量时,DDPG就不易收敛[19].

以上研究方案并不适用于未来大规模、多形态、深覆盖的机器类物联网通信场景.由于机器类通信场景下拥有能量受限类设备的接入,对传输过程中的能量供给提出了新要求,但是现有方案大多只考虑了跳波束模式对带宽、流量等单一因素的影响,并未充分挖掘系统中的能量供给关系,对资源的利用有待进一步地提高.

针对上述问题,本文提出了基于MASAC最大熵强化学习的跳波束卫星系统资源适配方案,整体框架如图1所示.其主要思想是将资源分配

问题进行拆解,通过下行跳波束图样选择,速率与功率分配达到星-地供需平衡.在选择好的跳波束下进行终端上行链路能量资源优化.通过分时参数对跳波束时隙进行分割,在分割时隙下进行能量采集与数据传输,满足能量受限终端的通信需求[20].最后,本文对优化问题进行整合,采用多智能体最大熵强化学习进行优化求解,得到最优联合控制.

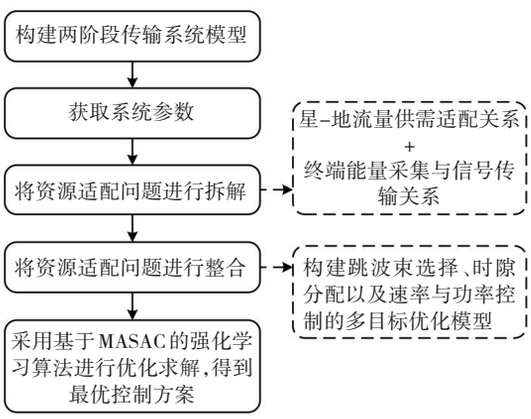

图1　跳波束卫星系统资源适配方案总体框架

Fig. 1　Overall framework of resource adaptation scheme for beam-hopping satellite system

2　系统模型及问题表述

2.1　系统模型

本文考虑跳波束卫星系统下行链路.卫星总功率为P,覆盖区域包括$N=\{n|1,2,\ldots,N\}$个小区,跳波束卫星可产生$K=\{k|1,2,\ldots,K\}(K<N)$个波束周期性地对地面进行覆盖.1个跳波束周期包含$T=\{t|1,2,\ldots,T\}$个时隙以及$H=\{H_1,H_2,\ldots,H_j,\ldots,H_J\}$个跳波束图样.在每个时隙选择$H$的1个子集作为1个跳波束图样$H_j$,所有波束使用相同的频段$B$.

为提高频谱利用率,卫星系统终端采用NOMA接入.由于机器类终端具有低功耗、低延迟等特点,因此,将其作为次用户(secondary user,SU)接入网络.同时,采用能量收集技术,收集环境中主用户(primary user,PU)产生的射频能量以供给用户自身通信需求[21].假设1个波束下用户数量为$L=\{M\cup N\}$,包括$M=\{m|1,2,\ldots,M\}$个PU和$N=\{n|1,2,\ldots,N\}$个SU.能量存储设备作为能量存储及释放单元被设置在每个波束小区内.由于终端地理分布的不均衡以及时变特性,每个小区的流量需求各不相同,假设跳波束图样H_j下用户的流量需求为$D_j=\{D_1,D_2,\ldots,D_b,\ldots,D_K\}$,所构建的系统模型如图2所示.

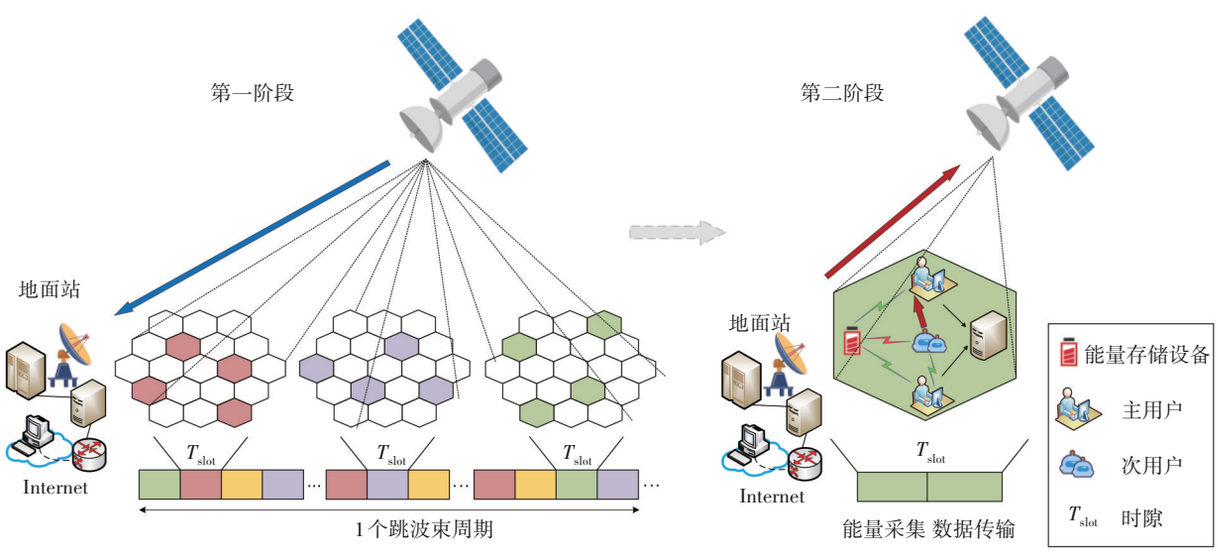

图2　两阶段传输系统模型

Fig. 2　Two-stage transmission system model

在跳波束时隙T、波束图样H_j被选定时,其下的波束b中终端用户i的信噪比可表示为

$$\gamma_{ibtj}=\frac{\left|h_{bi}\right|^2\mu_{ibtj}P}{I_{ibtj}^{\text{intra}}+I_{ibtj}^{\text{inter}}+\sigma^2}. \tag{1}$$

其中:h_{bi}为服务用户i的波束b与用户i之间的信道增益;μ_{ibtj}为功率分配系数,$0<\mu_{ibtj}<1$;P为卫星发射功率;I_{ibtj}^{intra}为用户i与同一波束内其他用户之间的波束内干扰;I_{ibtj}^{inter}为用户i与其他被照亮波

束之间的波束间干扰;σ^2为高斯白噪声.

波束内干扰I_{ibtj}^{intra}和波束间干扰I_{ibtj}^{inter}可分别表示如下:

$$I_{ibtj}^{\text{intra}}=\sum_{i'=1}^{L-1}\left|h_{bi}\right|^2\mu_{i'btj}P\,, \tag{2}$$

$$I_{ibtj}^{\text{inter}}=\sum_{a\in[1,2,\cdots,K]}\sum_{i'=1}^{L}\varepsilon_t^a\left|h_{bi}\right|^2\mu_{i'atj}P\,. \tag{3}$$

其中:$\varepsilon_t^a=\{0,1\}$表示在时隙T、跳波束图样H_j下与波束b相邻的波束a是否被照亮.

2.2 问题表述

2.2.1 星-地供需流量失配问题表述

在时隙T跳波束图样H_j中,波束b中终端i可获得的流量(R_{ibtj})为

$$R_{ibtj}=\varepsilon_t^b B\ln\left(1+\gamma_{ibtj}\right). \tag{4}$$

其中,B为卫星带宽.

因此,1个跳波束周期内,卫星提供给波束b中终端i的流量(R_{ib})及卫星提供给波束b的总流量(R_b)可分别计算如下:

$$R_{ib}=\sum_{t\in T}\sum_{j\in H}R_{ibtj}\,, \tag{5}$$

$$R_b=\sum_{i=1}^{L}R_{ib}\,. \tag{6}$$

为了使卫星提供容量与波束请求流量相匹配,消除小区之间不同需求量级的影响,将供需匹配关系(supply and demand matching relationship,SDMR)转化为未匹配的系统容量比与溢出系统容量比的加权值[16](R_{SDMR}),其表述如下:

$$R_{\text{SDMR}}=\min\sum_{b=1}^{K}\frac{\max[R_b-D_b,0]}{R_b+D_b}+\xi\frac{\max[D_b-R_b,0]}{R_b+D_b}. \tag{7}$$

其中,ξ为调和参数.

2.2.2 终端的能量采集与信号传输关系表述

在跳波束时隙T中,在$(1-\rho(t))T$时间内进行SU的数据传输;在$\rho(t)T$时间内,通过环境从PU处收集射频能量并存储于能量存储设备中.其中$\rho(t)$为分时参数,$0<\rho(t)<1$.能量收集与信号传输过程如图3所示,n_1,n_2为噪声.

假设从PU处收集射频能量时,PU额外的能量消耗忽略不计.在跳波束时隙T、波束b下,SU_n收集的射频能量可计算为

$$E_n(t)=\eta\rho(t)T\sum_{m=1}^{M}\left|h_{mn}\right|^2P_{\text{PU}}^m. \tag{8}$$

其中:η为能量收集效率;h_{mn}为SU_n与PU_m之间的信道增益;P_{PU}^m为PU_m的辐射功率.

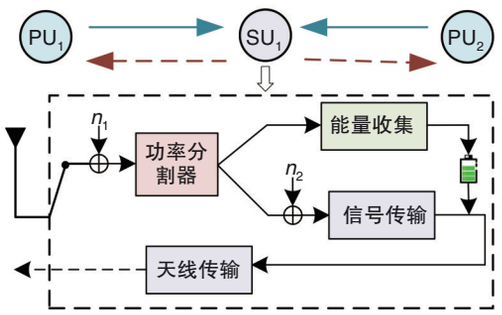

图3 能量收集与信号传输过程

Fig. 3 Energy harvesting and signal transmission process

SU_n收集完射频能量后,进行信号传输.其信号传输所需能量可表述如下:

$$E_n'(t)=(1-\rho(t))T\alpha_{\text{SU}}^n(t)P_{\text{SU}}^n. \tag{9}$$

其中:$\alpha_{\text{SU}}^n=\{0,1\}$为$\text{SU}_n$的当前状态,$\alpha_{\text{SU}}^n=1$表示$\text{SU}_n$当前为活跃状态,否则$\alpha_{\text{SU}}^n=0$;$P_{\text{SU}}^n$为$\text{SU}_n$进行数据传输所需发射功率.

SU_n在跳波束时隙T中经射频能量采集以及数据传输后,能量存储设备的剩余能量为

$$E_n^{\Delta}(t)=\min\{E_n(t)+E_n^{\Delta}(t-\Delta)-E_n'(t)\,,E_{\max}\}. \tag{10}$$

其中:$E_n^{\Delta}(t-\Delta)$为前1个时隙能量存储设备中的剩余能量;Δ为时间间隔;$E_{\max}$为能量存储上限.在下1个时隙中,SU利用射频能量采集获得的能量以及时隙T中能量存储设备剩余的能量进行数据传输,以此类推.因此,SU_n进行数据传输所需能量满足$0<E_n'(t)<E_n(t)+E_n^{\Delta}(t-\Delta)$.

将时隙T收集的射频能量以及前1个时隙$T-1$能量存储设备的剩余能量转化为SU自身通信所需的功率.在时隙T跳波束图样H_j下,波束小区b中能量存储设备提供的用于SU数据传输总功率为

$$P_{bjT}'=\sum_{n=1}^{N}\frac{E_n(t)+E_n^{\Delta}(t)}{(1-\rho(t))T}. \tag{11}$$

为保障SU的通信需求,能量存储设备可提供的功率应大于SU进行信号传输所需的发射功率:

$$P_{bjT}'\geqslant\sum_{n=1}^{N}P_{\text{SU}}^n\geqslant 0\,. \tag{12}$$

2.2.3 联合优化问题

将上述星-地供需流量失配问题以及终端能量采集与信号传输关系转化为跳波束图样选择,时隙分配以及速率与功率控制多目标优化数学模型为

$$\left.\begin{aligned}
&P\text{:Max}\sum_{b\in K}1-R_{\text{SDMR}},\\
&\text{s.t. } C_1\text{:}\sum_{b=1}^{K}\varepsilon_t^b=K,\varepsilon_t^b=\{0,1\},\\
&C_2\text{:}0<\mu_{ibtj}<1,\\
&C_3\text{:}\sum_{b=1}^{K}P_b<P,\\
&C_4\text{:}0<\rho\ (t)<1,\\
&C_5\text{:}0\leqslant\sum_{n=1}^{N}P_{\text{SU}}^n\leqslant P'_{bjT}.
\end{aligned}\right\}\tag{13}$$

在优化问题(13)中:C_1表示跳波束图样选择参数限制因素,每个跳波束时隙下只能有K个波束被照亮;C_2表示卫星与波束之间的功率分配参数取值范围应在(0,1)之间;C_3为波束功率限制条件,每个跳波束时隙下被照亮波束所分配的总功率不应大于卫星提供的总功率;C_4为分时参数范围;C_5保证了SU进行信号传输的通信需求.

3　多智能体最大熵强化学习算法

3.1　基于MASAC的深度强化学习框架

本文将最大熵学习结合到Actor-Critic框架当中,最大化资源分配满意度与流量缺口加权的累积奖励,同时最大化策略的熵.由于MASAC对超参数很敏感,引入熵正则化因子来提高训练过程的稳定性[22].将每个波束视为1个智能体,在每个智能体上部署1个参数为Φ的策略网络π_Φ,具有参数θ_1和θ_2的2个当前Q网络Q_{θ_1}和Q_{θ_2}以及具有参数θ'_1和θ'_2的2个目标Q网络$Q_{\theta'_1}$和$Q_{\theta'_2}$.所采用的算法架构如图4所示.系统的状态、动作和奖励设定如下.

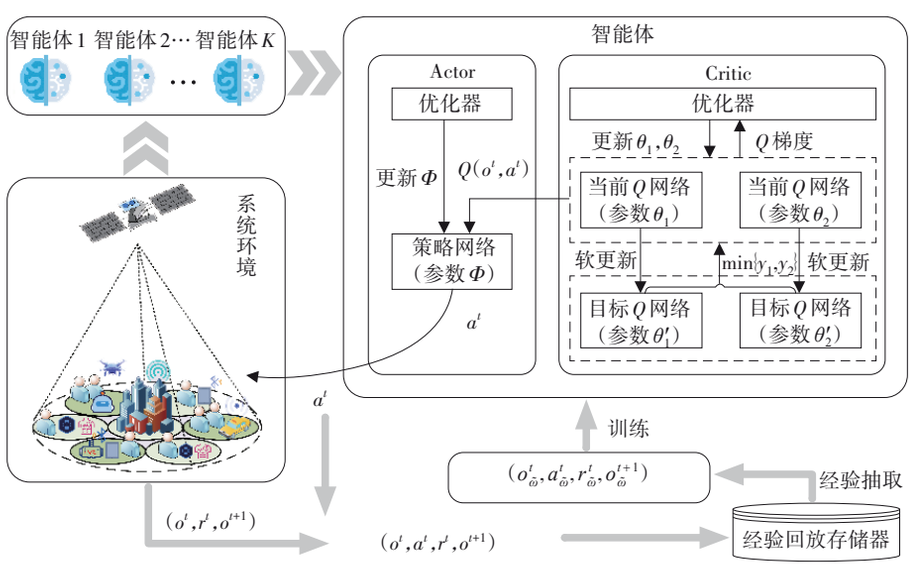

图4　基于MASAC的深度强化学习框架

Fig. 4　Deep reinforcement learning structure based on MASAC

3.1.1　状态

将观测状态$O_i^t\in\boldsymbol{O}$定义为智能体i在时隙T的局部观测信息,主要由时隙T波束i下用户的流量请求D_i^t、时隙T波束i下用户的信道信息$H_i^t=\{h_{i,1}^t,h_{i,2}^t,...,h_{i,m}^t\}$以及波束小区内能量存储设备剩余能量$E_i^t$组成:

$$O_i^t=\{D_i^t,H_i^t,E_i^t\}.\tag{14}$$

所有K个智能体的局部观测值的组合即为时隙T整个系统的状态空间O^t.

3.1.2　动作

智能体在观察环境后,通过相应状态O_i^t确定在该状态下的行为.将$a_i^t\in\boldsymbol{A}$定义为智能体i在时隙T中要执行的动作,包括跳波束图样选择参数$a_{i,\varepsilon}^t$,决定当前时隙波束是否被照亮;分时参数$a_{i,\rho}^t$,用于对跳波束时隙T进行划分;功率分配系数$a_{i,\mu}^t$,决定每个波束的功率分配情况:

$$a_i^t=\{a_{i,\varepsilon}^t,a_{i,\rho}^t,a_{i,\mu}^t\}.\tag{15}$$

所有K个智能体的动作值的组合即为时隙T整个系统的联合动作$\boldsymbol{A}^t$.

3.1.3　奖励

智能体执行动作后获得即时反馈.将智能体在时隙T中完成上述动作后的奖励函数$r^t\in\mathbf{R}$设计为与满意度、流量缺口有关的函数:

$$r^t=\frac{1}{K}\sum_{b=1}^{K}S_b-\frac{|\Delta_b|}{\omega}.\tag{16}$$

其中：$S_b=\frac{R_b}{D_b}$为波束的资源分配满意度；Δ_b为波束的流量缺口，$\Delta_b=D_b-R_b$；ω为常数，用于标准化Δ_b.

3.2 MASAC算法实现

3.2.1 初始化阶段

随机初始化网络参数Φ，θ_1和θ_2，并使用θ_1和θ_2对目标Critic网络参数θ_1'和θ_2'进行赋值；清空经验回放存储器.

3.2.2 智能体训练阶段

每个智能体$i(i=1,...,K)$单独观测其局部环境状态o_i^t，并依据当前局部环境状态，随机选择1个动作集$a_i^t\sim\pi_\Phi(\cdot|O_i^t)$作为输出，并执行联合动作$\boldsymbol{A}^t=(a_1^t,a_2^t,...,a_K^t)$. 智能体执行联合动作$\boldsymbol{A}^t$后，得到单步奖励$r^t$与策略的熵.在获得单步奖励后，将全局状态更新为$\boldsymbol{O}^{t+1}$.将环境全局状态、输出动作及获得单步奖励作为经验$\{\boldsymbol{O}^t,\boldsymbol{A}^t,r^t,\boldsymbol{O}^{t+1}\}$存储于经验回放存储器$W$中，并提取1个$\widetilde{\omega}\in W$的小批经验用于训练神经网络.在后续训练中智能体寻求最大化长期累积折扣奖励，同时最大化策略熵：

$$R(\boldsymbol{O}^{t+1},\boldsymbol{A}^t)=\max E[\sum_{t=1}^{\infty}\gamma^{t-1}[r^t(\boldsymbol{O}^t,\boldsymbol{A}^t)+\alpha H(\pi(\cdot|\boldsymbol{O}^t))]]. \tag{17}$$

其中：γ为衰变系数；α为熵正则化因子；$H(\pi(\cdot|\boldsymbol{O}^t))=-\log\pi_\Phi(\boldsymbol{A}^t|\boldsymbol{O}^t)$为计算$\pi(\cdot|\boldsymbol{O}^t)$的熵；$\pi(\cdot|\boldsymbol{O}^t)$为从状态到动作的映射概率分布.

智能体依据长期累积折扣奖励和最大化策略熵输出一组策略向量：

$$\boldsymbol{\pi}^*=\underset{\pi}{\operatorname{argmax}}\,R(\pi). \tag{18}$$

通过引入Q网络进行迭代，以改进输出策略，并利用当前Critic网络计算对动作的评估值：

$$Q(\boldsymbol{O}^t,\boldsymbol{A}^t)=\underset{\{\boldsymbol{O},\boldsymbol{A}\}\sim W}{E}[R(\boldsymbol{O}^t,\boldsymbol{A}^t)+\gamma^t(Q(\boldsymbol{O}^{t+1},\boldsymbol{A}^{t+1}))+\alpha H(\pi(\cdot|\boldsymbol{O}^t))]. \tag{19}$$

训练柔性Q网络以最小化贝尔曼残差：

$$J_Q(\theta)=\underset{\{\boldsymbol{O},\boldsymbol{A},\boldsymbol{O}'\}\sim W}{E}[(Q_\theta(\boldsymbol{O}^t,\boldsymbol{A}^t)-y(r^t,\boldsymbol{O}^{t+1}))^2]. \tag{20}$$

其中，

$$y(r^t,\boldsymbol{O}^{t+1})=r^t+\gamma\pi_\theta(\boldsymbol{A}^{t+1}|\boldsymbol{O}^{t+1})^{\mathrm{T}}\times(\min_{j=1,2}Q_{\theta_j'}(\boldsymbol{O}^{t+1},\boldsymbol{A}^{t+1})-\alpha\log\pi_\theta(\boldsymbol{A}^{t+1}|\boldsymbol{O}^{t+1})). \tag{21}$$

3.2.3 参数更新阶段

求解$J_Q(\theta)$关于θ的梯度$\nabla_\theta J_Q(\theta)$，进而对当前Critic网络中的参数$\theta(\theta_i,i=1,2)$进行更新：

$$\nabla_\theta J_Q(\theta)=\nabla_\theta\frac{1}{|\widetilde{\omega}|}\sum_{\{\boldsymbol{O},\boldsymbol{A},\boldsymbol{O}',\boldsymbol{A}'\in\widetilde{\omega}\}}(Q(\boldsymbol{O}^t,\boldsymbol{A}^t)-y(r^t,\boldsymbol{O}^{t+1}))^2. \tag{22}$$

更新当前Actor网络参数：

$$\nabla_\Phi J_\pi(\Phi)=\nabla_\Phi\frac{1}{|\widetilde{\omega}|}\sum_{\{\boldsymbol{O}\in\widetilde{\omega}\}}(\min_{i=1,2}Q(\boldsymbol{O}^t,\boldsymbol{A}^{t+1})-\alpha\log\pi_\theta(\boldsymbol{A}^{t+1}|\boldsymbol{O})). \tag{23}$$

对熵正则化因子进行更新：

$$\nabla_\alpha J(\alpha)=\pi_\theta(\boldsymbol{A}^t|\boldsymbol{O}^t)^{\mathrm{T}}[-\nabla_\alpha\alpha\log\pi_\theta(\boldsymbol{A}^t|\boldsymbol{O})+\overline{\boldsymbol{H}}]. \tag{24}$$

其中$\overline{\boldsymbol{H}}$表示目标熵的恒定向量.

采用滑动平均方式对目标Critic网络参数θ_i'进行更新：

$$\theta_i'\leftarrow\tau\theta_i+(1-\tau)\theta_i',\ i=1,2. \tag{25}$$

其中τ为学习率，$\tau\in(0,1)$.

经上述训练后，智能体获得最优联合跳波束图样选择、速率与功率控制策略$\boldsymbol{A}^*=[A_\varepsilon^*,A_\rho^*,A_\mu^*]$. 本文所采用的MASAC具体实现步骤如算法1所示.

算法1 基于MASAC的跳波束卫星系统资源适配算法

输入：初始化Actor网络参数Φ，Critic网络参数$\theta_1,\theta_2,\theta_1',\theta_2'$，熵正则化因子$\alpha$，经验回放存储器$W$

```
设置目标Critic网络参数：θ1', θ2' ← θ1, θ2
for each episode do
    重置初始环境；
    for t←1, ···, T do
        for i←1, ···, K do
            观测环境 O_i^t=[D_i^t, h_i^t]并根据策略网络选择策略 a_i^t ~ π_Φ(·|O_i^t)；
        end for
        执行联合动作 A^t=(a_1^t, ..., a_K^t)；
        获得奖励 r^t 以及下一时刻的环境状态 O^{t+1}
        将经验元组{O^t, A^t, r^t, O^{t+1}}存储到经验回放存储器 W；
        if O^{t+1} 到达最终状态 then
            重新初始化环境
        end if
```

if 网络更新时间步到达 **then**
　for $i \leftarrow 1,\cdots,K$ **do**
　从 W 中随机抽取一批数据 $\widetilde{\omega}$;
　根据式(22)更新Critic网络参数;
　根据式(23)更新Actor网络参数;
　根据式(25)更新目标网络参数;
　根据式(24)更新熵正则化因子;
　end for
end if
end for
end for
输出:最优联合控制策略 $\pi^*(\boldsymbol{A}^*)$

4　仿真验证

4.1　仿真参数

为评估模型和算法的有效性,本文选用Python3.6和TensorFlow1.0对所提方案进行了实验仿真.其场景设计如下:待服务区域被划分成30个规模相等的小区,每个小区内包含1个能量收集单元,用于存储收集的射频能量.假设每个小区的请求流量服从泊松分布,跳波束卫星系统包括5个波束,主要仿真参数设置如表1所示.

表1　主要仿真参数

Table 1　Main simulation parameters

仿真参数	取值
卫星轨道高度/km	1 000
下行链路工作频率/ GHz	20
系统带宽/MHz	500
卫星波束个数	5
服务小区总数/个	30
卫星最大天线增益/dBi	52
用户接收天线增益/dBi	21
卫星星载总功率/W	2 000
主用户发射功率/dBm	30
噪声功率密度/(dBm·Hz^{-1})	−174
时隙长度/ ms	2
能量存储单元最大容量/J	0. 6
能量收集效率系数	0. 7

本文采用MASAC算法解决跳波束卫星系统中的波束调度、功率与速率分配问题.因此需要对神经网络的参数进行训练,MASAC算法参数设置如表2所示.

4.2　仿真结果

图5对比了前30个被选择的跳波束中,用户

表2　MASAC算法参数设定

Table 2　MASAC algorithm parameters settings

MASCA算法参数	取值
训练轮次	400
训练步数	100
经验池容量	1 000
折扣因子γ	0. 9
学习率	0. 001
批量训练数目	32
优化器算法	Adam

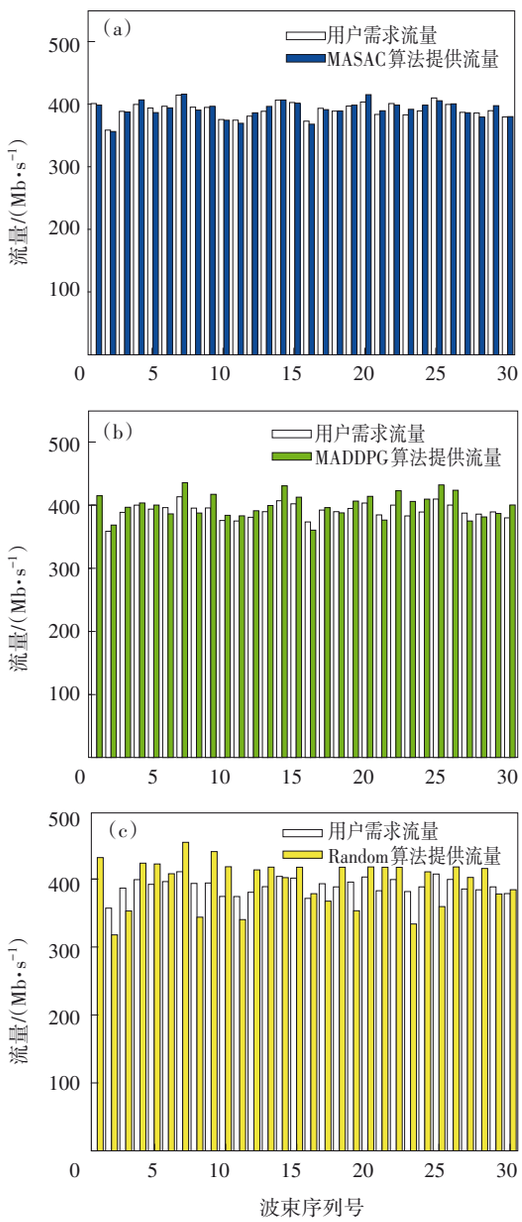

图5　采用不同算法时星-地供需流量关系的比较

Fig. 5　Comparison of the supply-demand flow relationship between satellite and ground with different algorithms

(a)—MASAC算法下的供需流量;(b)—MADDPG算法下的供需流量;(c)—Random算法下的供需流量.

需求流量与不同算法的容量供给分布情况.本文将所提算法与MADDPG(multi-agent deep deterministic policy gradient)算法和随机选择策略进行了比较.仿真结果显示,Random算法在某些情况下不能满足用户需求流量或者出现提供容量大于需求流量的情况,其算法的供需匹配误差较大,平均误差约为30.36 Mb/s.相比之下,MASAC算法与MADDPG算法可以较好地满足供需流量匹配.其中,MASAC算法供需流量平均误差约为4.04 Mb/s,远小于MADDPG算法供需流量平均误差11.45 Mb/s.由此可见,本文所提的MASAC算法在供需流量匹配方面更具有优越性,资源利用率较高.

图6验证了MASAC算法与MADDPG算法下SU平均吞吐量与PU辐射功率之间的关系.实验结果表明,随着PU辐射功率的增长,SU能收集更多射频能量,从而增强通信能力,导致吞吐量增加.但当PU辐射功率超过一定值后,由于SU能量存储容量的限制,使得SU吞吐量增长趋于饱和.

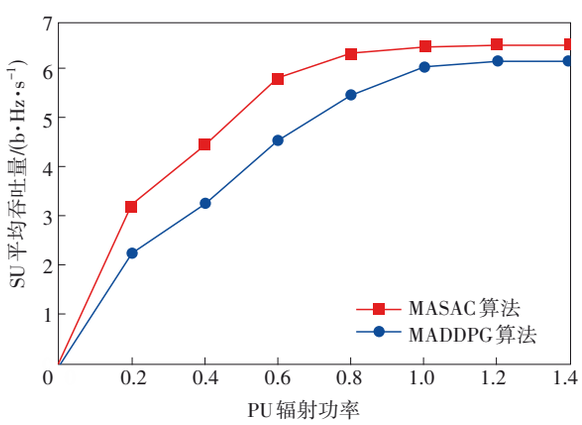

图6 SU的平均吞吐量与PU辐射功率的关系

Fig. 6 The relationship between the average throughput of SU and the radiated power of PU

为了证明所提方案的性能,图7比较了MASAC,MADDPG以及随机选择策略3种算法的收敛性能与稳定性.实验结果表明,MASAC算法在训练50轮左右趋于收敛,MADDPG算法在训练150轮左右趋于收敛.在训练速度方面,MASAC算法优于MADDPG算法,MASAC算法在训练中也更加稳定.

5 结 语

本文提出了基于MASAC最大熵强化学习的跳波束卫星系统资源适配方案.针对星-地资源供需失配和终端的能量受限问题,本文构建了两阶段传输系统模型,并探讨了资源分配策略.本文建立了跳波束图样选择、时隙分配以及速率与功率控制的多目标优化问题,并将SAC方法拓展到多智能体强化学习领域,采用MASAC框架进行优化问题的求解.实验结果表明,与两种基准方案相比本文所提方案具有良好的收敛性和稳定性.

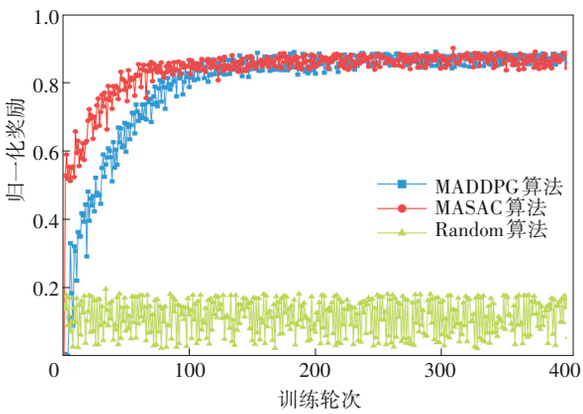

图7 3种算法的收敛性能

Fig. 7 Convergence performance of three algorithms

参考文献:

[1] Euler S, Fu X T, Hellsten S, et al. Using 3GPP technology for satellite communication [J]. *Ericsson Technology Review*, 2023, 2023(6): 2-12.

[2] 何炬良.卫星通信中基于载波协同的随机多址接入技术研究[D].北京:北京邮电大学,2018.
(He Ju-liang. Random multiple access based on carrier cooperation for satellite communication systerm [D]. Beijing: Beijing University of Posts and Telecommunications, 2018.)

[3] Hu X, Zhang Y C, Liao X L, et al. Dynamic beam hopping method based on multi-objective deep reinforcement learning for next generation satellite broadband systems[J]. *IEEE Transactions on Broadcasting*, 2020, 66(3): 630-646.

[4] Wang A Y, Lei L, Lagunas E, et al. Joint optimization of beam-hopping design and NOMA-assisted transmission for flexible satellite systems[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(10): 8846-8858.

[5] Kamalinejad P, Mahapatra C, Sheng Z G, et al. Wireless energy harvesting for the Internet of things [J]. *IEEE Communications Magazine*, 2015, 53(6): 102-108.

[6] 彭醇陵.基于射频能量收集的双向中继网络传输优化研究[D].重庆:重庆邮电大学,2019.
(Peng Chun-ling. Research on transmission optimization strategy in two-way relay networks with RF energy harvesting [D]. Chongqing: Chongqing University of Posts and Telecommunications, 2019.)

[7] OPPO研究院.零功耗通信白皮书[R/OL].(2022-01-19)[2023-04-18]. https://www.oppo.com/content/dam/oppo/en/mkt/newsroom/press/oppo-releases-zero-power-communication-whitepaper/white-paper_cn.pdf.
(OPPO Research Institute. Zero power communications

white paper[R/OL].(2022-01-19)[2023-04-18]. https://www. oppo. com/content/dam/oppo/en/mkt/newsroom/press/oppo-releases-zero-power-communication-whitepaper/white-paper_cn.pdf.)

[8] Aravanis A I, Bhavani S M R, Arapoglou P D, et al. Power allocation in multibeam satellite systems: a two-stage multi-objective optimization[J]. *IEEE Transactions on Wireless Communications*, 2015, 14(6): 3171-3182.

[9] Wang W L, Wei J, Zhao S H, et al. Energy efficiency resource allocation based on spectrum-power tradeoff in distributed satellite cluster network[J]. *Wireless Networks*, 2020, 26(6): 4389-4402.

[10] Zhang M Y, Yang X M, Bu Z Y. Resource allocation with interference avoidance in beam-hopping based LEO satellite systems [C]//The 4th Information Communication Technologies Conference (ICTC). Nanjing, 2023: 83-88.

[11] Zhang T, Zhang L X, Shi D Y. Resource allocation in beam hopping communication system [C]// IEEE/AIAA 37th Digital Avionics Systems Conference (DASC). London, 2018: 1-5.

[12] Shi S C, Li G X, Li Z Q, et al. Joint power and bandwidth allocation for beam-hopping user downlinks in smart gateway multibeam satellite systems [J]. *International Journal of Distributed Sensor Networks*, 2017, 13 (5): 155014771770946.

[13] Wu S W, Zhang S, Li Q, et al. Study of non-orthogonal multiple access technology for satellite communications [C]// IEEE 8th International Conference on Computer and Communications (ICCC). Chengdu, 2022: 771-775.

[14] Wang A Y, Lei L, Lagunas E, et al. Joint beam-hopping scheduling and power allocation in NOMA-assisted satellite systems [C]// IEEE Wireless Communications and Networking Conference (WCNC). Nanjing, 2021: 1-6.

[15] Lin Z Y, Ni Z Y, Kuang L L, et al. Dynamic beam pattern and bandwidth allocation based on multi-agent deep reinforcement learning for beam hopping satellite systems [J]. *IEEE Transactions on Vehicular Technology*, 2022, 71 (4): 3917-3930.

[16] 徐素洁,胡欣,王银,等. 基于深度强化学习的卫星动态功率控制技术[J]. 陆军工程大学学报,2022,1(2): 13-20.
(Xu Su-jie, Hu Xin, Wang Yin, et al. Dynamic power allocation technology for satellites based on deep reinforcement learning [J]. *Journal of Army Engineering University of PLA*, 2022, 1(2): 13-20.)

[17] Wang X M, Zhang Y H, Shen R J, et al. DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems[J]. *IEEE Internet of Things Journal*, 2020, 7(8): 7279-7294.

[18] Zhang H Y, Liu R K, Kaushik A, et al. Satellite edge computing with collaborative computation offloading: an intelligent deep deterministic policy gradient approach[J]. *IEEE Internet of Things Journal*, 2023, 10(10): 9092-9107.

[19] 张严心,孔涵,殷辰堃,等. 一类基于概率优先经验回放机制的分布式多智能体软行动-评论者算法[J]. 北京工业大学学报,2023,49(4):459-466.
(Zhang Yan-xin, Kong Han, Yin Chen-kun, et al. Distributed multi-agent soft actor-critic algorithm with probabilistic prioritized experience replay [J]. *Journal of Beijing University of Technology*, 2023, 49(4): 459-466.)

[20] Ghosh D, Hanawal M K, Zlatanov N. Learning to optimize energy efficiency in energy harvesting wireless sensor networks[J]. *IEEE Wireless Communications Letters*, 2021, 10(6): 1153-1157.

[21] Ding Z G, Schober R, Poor H V. No-pain No-gain: DRL assisted optimization in energy-constrained CR-NOMA networks[J]. *IEEE Transactions on Communications*, 2021, 69(9): 5917-5932.

[22] Wu D P, Liu T, Li Z D, et al. Delay-aware edge-terminal collaboration in green Internet of vehicles: a multiagent soft actor-critic approach [J]. *IEEE Transactions on Green Communications and Networking*, 2023, 7(2): 1090-1102.

(上接第8页)

[12] Li H Q, Lu J X , Wang J F, et al. SDM4IIoT: an SDN-based multicast algorithm for Industrial Internet of Things [J]. *IEICE Transactions on Communications*, 2022, 105 (5): 545-556.

[13] Liao L X, Leung V C M. LLDP based link latency monitoring in software defined networks [C]//The 12th International Conference on Network and Service Management(CNSM). Montreal, 2016: 330-335.