智能电网中基于多智能体强化学习的频谱分配算法

燕锋[1]，林晓薇[2]，李正浩[3]，徐霞[4]，夏玮玮[1]，沈连丰[1]

（1. 东南大学移动通信全国重点实验室，江苏 南京 210096；
2. 东南大学软件学院，江苏 南京 211100；
3. 国网山东省电力公司信息通信公司，山东 济南 250001；
4. 国网山东省电力公司济南供电公司，山东 济南 250012）

摘 要：针对智能电网中利用 5G 网络承载多样化电力终端的业务需求，提出了一种基于多智能体强化学习的频谱分配算法。首先，基于智能电网中部署的集成接入回程系统，考虑智能电网中轻量化和非轻量化终端业务的不同通信需求，将频谱分配问题建模为最大化系统总能效的非凸混合整数规划。其次，将前述问题构建为一个部分可观测的马尔可夫决策过程并转换为完全协作的多智能体问题，进而提出了一种集中训练分布执行框架下基于多智能体近端策略优化的频谱分配算法。最后，通过仿真验证了所提算法的性能。仿真结果表明，所提算法具有更快的收敛速度，通过有效减少层内与层间干扰、平衡接入与回程链路速率，可以将系统总速率提高 25.2%。

关键词：智能电网；集成接入回程；频谱分配；多智能体强化学习

中图分类号：TN92

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2023179

Spectrum allocation algorithm based on multi-agent reinforcement learning in smart grid

YAN Feng[1], LIN Xiaowei[2], LI Zhenghao[3], XU Xia[4], XIA Weiwei[1], SHEN Lianfeng[1]

1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
2. School of Software, Southeast University, Nanjing 211100, China
3. State Grid Shandong Information and Telecommunication Company, Jinan 250001, China
4. State Grid Jinan Power Supply Company, Jinan 250012, China

Abstract: In view of the fact that 5G networks are used to meet the service requirements of various power terminals in smart grid, a spectrum allocation algorithm based on multi-agent reinforcement learning was proposed. Firstly, for the integrated access backhaul system deployed in smart grid, considering the different communication requirements of services in lightweight and non-lightweight terminal, the spectrum allocation problem was formulated as a non-convex mixed-integer programming aiming to maximize the overall energy efficiency. Secondly, the above problem was modeled as a partially observable Markov decision process and transformed into a fully cooperative multi-agent problem, then a spectrum allocation algorithm was proposed which was based on multi-agent proximal policy optimization under the framework of centralized training and distributed execution. Finally, the performance of the proposed algorithm was verified by simulation. The results show that the proposed algorithm has a faster convergence speed and can increase the overall transmission rate by 25.2% through effectively reducing intra-layer and inter-layer interference and balancing the access and backhaul link rates.

Keywords: smart grid, integrated access and backhaul, spectrum allocation, multi-agent reinforcement learning

---

收稿日期：2023−05−25；**修回日期：**2023−09−04

基金项目：国家电网有限公司科技基金资助项目（No.520601220022）

Foundation Item: The Science and Technology Project of State Grid Corporation of China (No.520601220022)

0　引言

智能电网利用先进的信息通信技术、新能源技术和电力传输技术实现了电力系统的智能化、高效化、可靠化和安全化[1]。通信技术是智能电网实现实时监测和控制、精细化管理以及多点接入互联互通的关键，其可靠性和效率直接取决于通信基础设施的性能，但随着电力设备接入的密集度越来越高，传统的信息交互系统无法保证电力业务终端的服务需求[2]，满足智能电网中多样化服务需求是一项较大的挑战。随着移动通信技术的发展，5G 技术可提供超高带宽、超可靠低时延以及超大规模连接的用户体验[3]，可以很好地适应未来电力多场景、差异化业务灵活承载的需求，但通信业务规模的不断扩大，频谱资源变得越来越紧缺，使高效资源管理成为 5G 承载电力业务亟待解决的关键技术之一。

对于面向智能电网的 5G 资源管理，已有不少研究取得了相关进展。文献[4]为解决智能电网中配电网对可再生能源的管理，利用终端直通 D2D（device-to-device）传输解决实时定价问题。文献[5]在传输过程中部署智能反射面并利用非正交多址（NOMA, non-orthogonal multiple access）技术以满足智能电网中时延敏感型业务对通信资源的迫切需求。文献[6-7]分别将云计算和边缘计算技术引入智能电网中支持需求侧资源管理，通过优化带宽资源和计算能力进一步提高了通信质量。文献[8]利用波束成形技术优化智能电网中的家庭局域网，将其建模为异构多用户网络并对下行链路容量进行优化。文献[9-10]考虑智能电网中应用超密集异构网络能源消耗过大的问题分别提出资源优化方案。这些文献大多只关注智能电网中的一种业务，但智能电网中不同业务的差异化需求较明显，例如，广域态势感知系统、配网保护、机器人巡检等控制类和巡检类业务需要高数据速率来实现可靠性、准确性和实时性，而高级量测、智能电表等采集类业务并不需要同等条件的数据速率[11]，若采用同样的资源分配方案会使资源利用效率不高，因此对智能电网的多样化业务提供一个自适应的高效资源分配方案是智能电网发展的需要。

另一方面，5G 技术的一个方案是通过网络密集化和频率复用达到增加网络容量的目的，而部署大量有线光纤回程的运营成本高昂，并不适于智能电网中需要覆盖大范围多连接的场景。针对大量部署有线光纤成本高昂问题，3GPP 在 R16 中提出了集成接入回程（IAB, integrated access and backhaul）系统[12]，并在 R17 中进一步规范。IAB 系统中无线回程链路取代传统的有线光纤回程链路，有利于在智能电网复杂的地理环境中进行灵活部署和扩展，优化智能电网的运营和维护，减少电网的运营成本；接入链路和回程链路共享相同的无线资源，在实现资源高效利用的同时可以优化电网的负载平衡，因此 IAB 系统可以更好地满足电力业务实际场景的通信需求。然而使用 IAB 系统带来的频谱资源高度复用在网络中会引起严重干扰，回程链路容量有限也限制了整个网络的性能。

针对 IAB 系统，不少学者对其资源分配提出了不同的优化方法。文献[13]针对使用 IAB 系统的智能电网提出基于通信能力的流量管理框架。文献[14]考虑全双工模式的 IAB 系统带宽划分问题，在避免 IAB 节点自干扰的情况下尽可能提高频谱效率。文献[15]采用一种基于序列的凸规划方法解决 IAB 系统的频谱和功率分配问题。文献[16]提出一种基于最大权重匹配的半集中式方法以解决 IAB 系统中接入与回程链路的资源分配问题。

尽管关于 IAB 系统的资源分配研究已经取得不错的进展，但是以上方法大多通过制定并求解优化问题获得，高度依赖模型的准确性，需要瞬时全局信道状态信息进行集中处理，计算复杂度较大；另一方面，网络环境是动态变化的，需要一个更智能、更灵活的资源分配算法。近年来，随着人工智能技术的发展，强化学习在解决复杂问题时展现出优越的性能，将其应用到资源分配问题中将是一个不错的解决方案。文献[17]研究上行链路多用户 NOMA 系统中的联合子信道分配和功率分配问题，分别采用了深度 Q 网络（DQN, deep Q-network）、深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法以及联合两者的强化学习方法以最大化系统能效。文献[18]在移动边缘网络中利用双层深度 Q 网络（DDQN, double deep Q-network）解决多维资源分配以最大化系统能效。文献[19-20]结合近端策略优化（PPO, proximal policy optimization）算法分别在大规模机器通信设备和增强型移动宽带类型场景下实现资源的动态调度。这些工作将复杂的全局优化问题建模为单智能体强化学习问题，对整个系统做出全局决策，对于 IAB 系统而言，每个节点都需要根据局部观察单独做出决策，

更适合使用多智能体系统。文献[21]首次将深度强化学习（DRL, deep reinforcement learning）方法应用到IAB系统的频谱分配方案中，使用完全集中式的多智能体框架分别利用DDQN、演员−评论家（AC, actor-critic）算法进行求解，但是每个用户都需要向中央控制器报告本地信息，使系统产生了大量的信令开销，并随着网络规模的扩大而急剧增长。文献[22]在此基础上使用完全分布式框架进一步求解了IAB系统的频谱和功率分配方案，虽然极大减少了通信开销，但完全分布式框架不利于智能体之间的协作。结合集中式和分布式框架的优点，文献[23]使用了集中训练分布执行框架下多智能体深度确定性策略梯度（MADDPG, multi-agent deep deterministic policy gradient）解决超密集网络中多维资源分配问题。文献[24]为最大化多用户无线蜂窝网络中的和速率，同样利用集中训练分布执行框架，提出DQN、DDPG方法来解决动态下行链路功率控制问题。上述多智能体框架均基于DQN或DDPG算法拓展而来，DQN算法适于离散动作空间问题，IAB系统中频谱分配的解空间维数随IAB节点数与信道数的增加呈指数级增长，处理这样的高维状态空间与动作空间时容易面临维度灾难。DDPG算法根据值函数进行更新，IAB系统中信道状态信息动态变化，DDPG收敛性能容易受到采样数据的相关性以及环境的不稳定性影响。PPO在策略更新过程中引入了比例裁剪和投影梯度等技术用于限制策略更新的幅度，保证训练过程的稳定性实现收敛；此外，PPO充分利用采样数据进行策略更新，加速了训练过程。为了最大限度提高有限频谱资源下的系统性能，需要在保证一定的通信开销下考虑各个智能体之间的合作与交互，本文提出使用集中训练分布执行的多智能体近端策略优化（MAPPO, multi-agent proximal policy optimization）解决IAB系统中的频谱分配问题。

综上所述，本文从电力终端的多样化需求出发，提出一种智能电网中基于多智能体强化学习的频谱分配算法，建立了基于IAB系统的双层通信网络频谱分配模型，利用集中训练分布执行的MAPPO解决该优化问题，在满足不同电力终端业务需求的条件下进一步提高了网络性能。

本文的主要贡献总结如下。

1) 首先，针对智能电网中的IAB系统，面向不同电力终端的通信需求，将频谱分配问题建模为非凸混合整数规划，其目标是在满足多约束条件下最大化系统总能效。

2) 其次，将此问题构建为一个部分可观测的马尔可夫决策过程并转换为完全协作的多智能体问题，提出了一种集中训练分布执行框架下基于MAPPO的频谱分配算法。所提算法避免了IAB系统频谱分配产生的维度灾难并加快了收敛速度；集中训练利用全局信息加强基站间的协作，避免独立执行带来的潜在不稳定和恶性竞争，同时分布执行降低了通信开销和计算复杂度。

3) 最后，通过仿真比较了所提算法与MADDPG、粒子群优化（PSO, particle swarm optimization）等算法的性能，结果表明所提算法能够有效减少层内与层间干扰，平衡接入与回程链路速率；与MADDPG相比收敛速度更快，系统总速率可以提高21%；在不同用户数和不同信道数下的系统性能均显著优于PSO算法，系统总速率最高可以提高25.2%。

1 系统模型

智能电网中不同典型电力业务对应不同的通信需求，相关通信需求包括速率、时延、带宽等，本文为实现智能电网中资源的高效利用，定义两类电力终端：轻量化终端 $\mathcal{U}'=\{u_0',\cdots,u_L'\}$ 和非轻量化终端 $\mathcal{U}=\{u_0,\cdots,u_H\}$。在变电设备状态感知、配网保护等控制类业务，以及输电线路在线监测、机器人及无人机智能巡检等巡检类业务中，需要保证业务的实时性、安全性和可靠性，传输速率要求至少达到10 Mbit/s，资源需要尽可能对这两类业务倾斜，因此将其定义为非轻量化终端；智能电网中还有一类采集业务，如智能电表、高级量测、电动汽车充电等，其传输速率、时延要求较低，资源分配方案可进行轻量化处理，因此将其定义为轻量化终端。

在智能电网中，考虑如图1所示的IAB系统上行链路。其中，IAB供体 b_0 通常为宏基站，位于系统中心；IAB节点集合 $\mathcal{B}'=\{b_n \mid n=1,2,\cdots,N\}$ 通常为小基站，均匀分布在IAB供体覆盖范围内；IAB供体和IAB节点一起构成基站集合 $\mathcal{B}=b_0\cup\mathcal{B}'$。该IAB系统的第一层为电力终端到IAB节点的接入链路，接收端为 $\mathcal{L}_1=\mathcal{B}'$；第二层为电力终端到IAB供体的接入链路和IAB节点到IAB供体的回程链路，接收端为 $\mathcal{L}_2=b_0$。

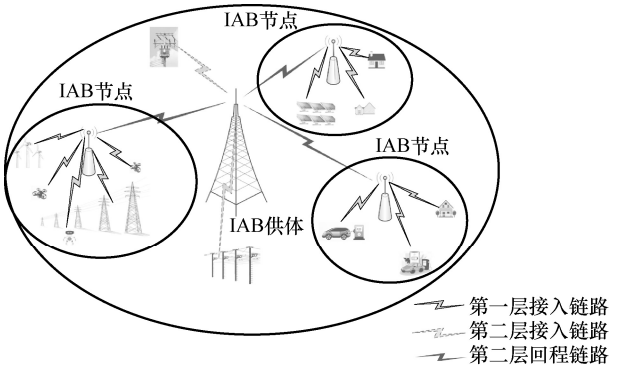

图 1　智能电网中的 IAB 系统上行链路

为了提高频谱资源利用率，IAB 节点以全双工正交频分复用方式工作，IAB 供体和每个 IAB 节点使用同样的频谱资源，可用带宽被划分为 M 个正交子信道，子信道集合 $\mathcal{M}=\{1,\cdots,M\}$，接入和回程链路通过 M 个正交子信道共享相同资源池。$c_u^m\in\{0,1\}$ 和 $z_u^m\in\{0,1\}$ 分别表示第一层和第二层的发射端 u 是否占用子信道 m。

在不失一般性的情况下，将几个连续的子载波形成频谱子信道，信道衰落在一个子信道内是相同的且不同子信道之间相互独立。第 m 个子信道处从发射端 i 到接收端 j 的上行链路信道增益可以表示为

$$g_{i,j}^m=\alpha_{i,j}h_{i,j}^m \tag{1}$$

其中，$h_{i,j}^m$ 是经历瑞利衰落的与频率相关的小尺度衰落，$\alpha_{i,j}$ 是与距离相关的大尺度衰落，包括路径损耗和阴影衰落（$i\in\mathcal{U}\cup\mathcal{U}'\cup\mathcal{B}'$, $j\in\mathcal{L}_1\cup\mathcal{L}_2$）。

对于第一层接入链路，电力终端向 IAB 节点上传数据时，IAB 节点会受到多层干扰。非轻量化终端 u_h 到 IAB 节点 b_n 在子信道 m 上的信干噪比可以表示为

$$\mathrm{SINR}_{u_h,b_n}^m=\frac{P_{u_h}g_{u_h,b_n}^m x_{u_h}^{b_n}c_{u_h}^m}{I_u+I_u'+I_b'+I_{\mathrm{self}}+I_{\mathrm{other}}+\delta^2} \tag{2}$$

其中，P_{u_h} 是非轻量化终端 u_h 的发射功率，δ^2 是噪声功率，$x_{u_h}^{b_n}\in\{0,1\}$ 是非轻量化终端 u_h 与 IAB 节点 b_n 的连接状态。I_u 是第一层接入链路中与其他 IAB 节点相连的电力终端的同层干扰，其计算式为

$$I_u=\sum_{\substack{b_j\notin\mathcal{B}',\\ j\neq n}}\sum_{\substack{u_i\notin\mathcal{U},\\ i\neq h}}P_{u_i}g_{u_i,b_n}^m x_{u_i}^{b_j}c_{u_i}^m+\sum_{\substack{b_j\in\mathcal{B}',\\ j\neq n}}\sum_{u_i\notin\mathcal{U}'}P_{u_i'}g_{u_i',b_n}^m y_{u_i'}^{b_j}c_{u_i'}^m \tag{2a}$$

其中，$y_{u_i'}^{b_j}\in\{0,1\}$ 是轻量化终端 u_l' 与 IAB 节点 b_j 的连接状态。I_u' 是第二层接入链路中与 IAB 供体相连的电力终端的跨层干扰，其计算式为

$$I_u'=\sum_{u_i\in\mathcal{U}}P_{u_i}g_{u_i,b_n}^m x_{u_i}^{b_0}z_{u_i}^m+\sum_{u_i'\in\mathcal{U}'}P_{u_i'}g_{u_i',b_n}^m y_{u_i'}^{b_0}z_{u_i'}^m \tag{2b}$$

其中，$x_{u_i}^{b_0}$ 和 $y_{u_i'}^{b_0}$ 分别是非轻量化终端 u_h 和轻量化终端 u_l' 与 IAB 供体 b_0 的连接状态。I_b' 是第二层回程链路中与 IAB 供体相连的其他 IAB 节点的跨层干扰，其计算式为

$$I_b'=\sum_{b_j\in\mathcal{B}',\ j\neq n}P_{b_j}g_{b_j,b_n}^m z_{b_j}^m \tag{2c}$$

因为 IAB 节点以全双工方式工作，所以会受到来自自身的干扰，可表示为

$$I_{\mathrm{self}}=P_{b_n}\varepsilon_{b_n}c_{b_n}^m \tag{2d}$$

其中，ε_{b_n} 是 IAB 节点 b_n 的自干扰消除系数。此外，其他小区的电力终端和 IAB 节点同样会对本小区的 IAB 节点产生干扰 I_{other}，为了简化模型，本文将小区间干扰 I_{other} 设置为常数。

同样地，可得到轻量化终端 u_l' 到 IAB 节点 b_n 在第 m 个子信道上的信干噪比 $\mathrm{SINR}_{u_l',b_n}^m$。

对于第二层接入链路，非轻量化终端 u_h 到 IAB 供体 b_0 在子信道 m 上的信干噪比可以表示为

$$\mathrm{SINR}_{u_h,b_0}^m=\frac{P_{u_h}g_{u_h,b_0}^m x_{u_h}^{b_0}z_{u_h}^m}{I_u+I_{\mathrm{other}}+\delta^2} \tag{3}$$

其中，I_u 是第一层接入链路中与 IAB 节点相连的电力终端的干扰，其计算式为

$$I_u=\sum_{b_j\in\mathcal{B}'}\sum_{u_h\in\mathcal{U}}P_{u_h}g_{u_h,b_0}^m x_{u_h}^{b_j}c_{u_h}^m+\sum_{b_j\in\mathcal{B}'}\sum_{u_l'\in\mathcal{U}'}P_{u_l'}g_{u_l',b_0}^m y_{u_l'}^{b_j}c_{u_l'}^m \tag{3a}$$

同理可得轻量化终端 u_l' 到 IAB 供体 b_0 在子信道 m 上的信干噪比 $\mathrm{SINR}_{u_l',b_0}^m$。

类似地，第二层回程链路 IAB 节点 b_n 到 IAB 供体 b_0 在第 m 个子信道上的信干噪比为

$$\mathrm{SINR}_{b_n,b_0}^m=\frac{P_{b_n}g_{b_n,b_0}^m z_{b_n}^m}{I_u+I_{\mathrm{other}}+\delta^2} \tag{4}$$

其中，I_u 与式(3a)保持一致。

根据香农公式，第二层接入链路中与 IAB 供体 b_0 直连的非轻量化终端 u_h 的传输速率可以表示为

$$\mu_{u_h}^{b_0}=\sum_{m=1}^{M}L_m\mathrm{lb}\left(1+\mathrm{SINR}_{u_h,b_0}^m\right) \tag{5}$$

其中，L_m 为第 m 个子信道的带宽。

同理可得第二层接入链路中轻量化终端的传输速率 $\mu_{u_l'}^{b_0}$ 以及 IAB 节点的传输速率 $\mu_{b_n}^{b_0}$。

第一层接入链路中与 IAB 节点 b_n 相连的非轻量化终端 u_h 的数据传输受到第二层回程链路中传输速率的限制，其传输速率可以表示为

$$\mu_{u_h}^{b_n}=\min\left(\mu_{u_h}^{b_n},\mu_{b_n}^{b_0}\right)=\min\left(\sum_{m=1}^{M}L_m\mathrm{lb}\left(1+\mathrm{SINR}_{u_h,b_n}^{m}\right),\sum_{m=1}^{M}L_m\mathrm{lb}\left(1+\mathrm{SINR}_{b_n,b_0}^{m}\right)\right) \tag{6}$$

同理可得第一层接入链路中轻量化终端的传输速率为 $\mu_{u_l'}^{b_n}$。

本文定义系统能量效率 η 为系统发射功率的产出投入比[25]，表示通信系统中单位能量所能传输的比特数，其计算式为

$$\eta=\frac{\sum_{b_n\in\mathcal{B}}\sum_{u_h\in\mathcal{U}}\mu_{u_h}^{b_n}+\sum_{b_n\in\mathcal{B}}\sum_{u_l'\in\mathcal{U}'}\mu_{u_l'}^{b_n}}{\sum_{u_h\in\mathcal{U}}P_{u_h}+\sum_{u_l'\in\mathcal{U}'}P_{u_l'}} \tag{7}$$

考虑到本文研究智能电网场景下 IAB 系统的上行链路频谱分配问题，智能电网中的轻量化和非轻量化终端通常采用恒定功率发射，不考虑功率的优化分配；非轻量化终端与轻量化终端相比需要更多的频谱资源以满足时延、速率等要求，将该能效问题转化为最大化系统上行链路非轻量化终端和速率问题，可描述为

$$\begin{aligned}
&\max\left(\sum_{u_h\in\mathcal{U},b_n\in\mathcal{B}}\mu_{u_h}^{b_n}\right)\\
&\text{s.t.}\quad \mathrm{C}_1: x_{u_h}^{b_n}\in\{0,1\},y_{u_l'}^{b_n}\in\{0,1\}\\
&\qquad \mathrm{C}_2: c_u^m\in\{0,1\},z_u^m\in\{0,1\}\\
&\qquad \mathrm{C}_3: \forall b_n\in\mathcal{B}',\sum_{u_h\in\mathcal{U}}x_{u_h}^{b_n}c_{u_h}^m+\sum_{u_l'\in\mathcal{U}'}y_{u_l'}^{b_n}c_{u_l'}^m\leqslant 1\\
&\qquad \mathrm{C}_4: \sum_{u_h\in\mathcal{U}}x_{u_h}^{b_0}z_{u_h}^m+\sum_{u_l'\in\mathcal{U}'}y_{u_l'}^{b_0}z_{u_l'}^m+\sum_{b_n\in\mathcal{B}'}z_{b_n}^m\leqslant 1\\
&\qquad \mathrm{C}_5: \mu_{u_h}^{b_n}\geqslant Q_h,\mu_{u_l'}^{b_n}\geqslant Q_l\\
&\qquad \mathrm{C}_6: \sum_{m=0}^{M}L_m\leqslant L
\end{aligned} \tag{8}$$

其中，C_1 约束了每个电力终端只能与一个基站相连接；C_2、C_3、C_4 约束了每个基站下的一个子信道只能被一个电力终端所占用；C_5 约束了非轻量化终端和轻量化终端的服务质量（QoS, quality of service）；C_6 约束了所有子信道的带宽之和不能超过总带宽。

2 算法设计

2.1 完全协作的多智能体问题

与 IAB 相关联的终端传输速率取决于回程链路和接入链路的传输速率，系统性能对频谱分配策略十分敏感，当部署更多的 IAB 节点时频谱分配的解空间维数呈指数级增加。此外，传统的基于模型的方法大多需要系统的完整信息，这在实践中往往不可行。结合深度神经网络（DNN, deep neural network）的 DRL 在解决 IAB 系统中的频谱分配问题上有着明显优势。一方面，DRL 可以通过与环境交互来学习优化策略，处理 IAB 系统中复杂的无线环境；另一方面，DRL 在处理非线性以及高维的优化问题上表现出优越的性能，根据不同的期望目标灵活设计奖励函数可以使系统性能朝有效方向发展。

将上述优化问题重新表述为部分可观测的马尔可夫过程。考虑到每个基站下都有同样的频谱资源可以分配，将每个基站 $b_n\in\mathcal{B}$ 都视为一个智能体。智能体只能观测到本地信息而不知道其他智能体信息，共同目标是寻找一个最优的频谱分配策略，这样一个完全协作的多智能体问题可以用一个部分可观测的马尔可夫决策过程 $\Gamma=<\mathcal{S},\mathcal{O},\mathcal{A},P,r>$ 表示。在时隙 t，$o_t^n\in\mathcal{O}$ 表示智能体 $n=0,\cdots,N$ 观察到的局部状态；$s_t\in\mathcal{S}$ 表示全局状态。智能体 n 根据策略函数 $\pi^n\left(a_t^n\mid o_t^n\right)$ 选择动作 $a_t^n\in\mathcal{A}$，所有智能体的联合动作为 $\boldsymbol{a}_t$；智能体根据状态转移函数 P 转移到下一个状态，即 $P(s_{t+1}\mid s_t,\boldsymbol{a}_t)$；所有智能体共享同一个奖励函数 $r(s_t,\boldsymbol{a}_t)$。

根据以上内容，将 DRL 应用到频谱分配问题中，上述马尔可夫决策过程的基本元素的详细设置如下。

1) 动作

智能体 n 的动作是将其下的 M 个子信道分配给与其连接的轻量化终端和非轻量化终端或 IAB 节点，即 $a^n=\{\boldsymbol{u}_0,\cdots,\boldsymbol{u}_M\}$。当智能体为 IAB 节点时，$\boldsymbol{u}_m=\left[c_{u_0}^m,\cdots,c_{u_H}^m,c_{u_0'}^m,\cdots,c_{u_L'}^m\right]$；否则，$\boldsymbol{u}_m=\left[z_{u_0}^m,\cdots,z_{u_H}^m,z_{u_0'}^m,\cdots,z_{u_L'}^m,z_{b_1}^m,\cdots,z_{b_N}^m\right]$，$\boldsymbol{u}_m$ 表示子信道 m 被占用情况。

2) 状态

每个智能体的状态由与系统传输速率相关的三部分组成，即 $o^n=\{\mathbf{QoS}_n,\mathrm{BH}_n,\mathbf{Rate}_n\}$。

$\mathbf{QoS}_n$ 记录与智能体 n 相连接的电力终端满足 QoS 需求的情况，定义为 $\mathbf{QoS}_n=\left[x_{u_0}^{b_n}q_0,\cdots,x_{u_H}^{b_n}q_H,y_{u_0}^{b_n}q_0',\cdots,y_{u_l'}^{b_n}q_L'\right]$，当 $q_h=1$ 时，非轻量化终端满足 QoS 需求；当 $q_l'=1$ 时，轻量化终端满足 QoS 需求。

BH_n 记录智能体 n 的回程链路信息，其计算式为

$$\mathrm{BH}_n=\begin{cases}0, & \mu^n_{\text{backhaul}}<\mu^n_{\text{access}}\\ \sum\limits_{u_h\in\mathcal{U}} x^{b_n}_{u_h}q_h+\sum\limits_{u'_l\in\mathcal{U}'} y^{b_n}_{u'_l}q'_l, & \mu^n_{\text{backhaul}}\geqslant\mu^n_{\text{access}}\end{cases} \tag{9}$$

其中，μ^n_{backhaul} 和 μ^n_{access} 分别是智能体 n 的回程链路和接入链路的传输速率，计算式分别为

$$\mu^n_{\text{backhaul}}=\mu^{b_0}_{b_n} \tag{9a}$$

$$\mu^n_{\text{access}}=\sum_{u_h\in\mathcal{U}}\mu^{b_n}_{u_h}+\sum_{u'_l\in\mathcal{U}'}\mu^{b_n}_{u'_l} \tag{9b}$$

当 IAB 节点为智能体时，BH_n 记录智能体 n 的回程链路是否为瓶颈链路，若回程链路传输速率小于接入链路传输速率，此时回程链路成为瓶颈链路，该项置 0；否则，记录该 IAB 节点下满足 QoS 需求的终端数量。当 IAB 供体为智能体时，BH_0 记录所有 IAB 节点的回程链路信息，其计算式为

$$\mathrm{BH}_0=\sum_{n=1}^{N}\mathrm{BH}_n \tag{10}$$

$\mathbf{Rate}_n$ 记录智能体 n 下每个子信道的传输速率，$\mathbf{Rate}_n=\left[v^1_n,\cdots,v^M_n\right]$，$v^m_n$ 是第 n 个智能体下子信道 m 的传输速率，计算式为

$$v^m_n=L_m\mathrm{lb}\left(1+\mathrm{SINR}^m_{u,b_n}\right) \tag{11}$$

干扰信息作为本文系统中的重要环境信息并未直接体现在状态设置中，但与状态设置的三项元素形成紧密关联。$\mathbf{QoS}_n$ 和 BH_n 由终端传输速率和 IAB 节点的回程链路传输速率直接决定，受干扰的间接影响；$\mathbf{Rate}_n$ 记录每个子信道的传输速率，受干扰的直接影响。各智能体依据各自的动作网络同时做出频谱分配决策，环境利用系统模型和频谱分配决策计算干扰信息进行更新并进入下一状态。

3) 奖励

频谱分配问题被建立为一个完全协作的多智能体问题，因此所有智能体共享同一个奖励。考虑到该优化问题需要在满足多个约束条件下最大化非轻量化电力终端的和速率，从 QoS 惩罚项和 QoS 奖励项 2 个方面设置具体的奖励函数。

① QoS 惩罚项

将当前环境中距离所有终端满足 QoS 需求的差距作为 QoS 惩罚项，其计算式为

$$r_p=\sum_{n=1}^{N}r_{p_n}+r_{p_0} \tag{12}$$

其中，r_{p_n} 是当 IAB 节点作为智能体时，将回程链路的影响考虑在内的惩罚因子。若回程链路成为瓶颈链路，将未满足 QoS 需求的终端数量放大至两倍作为惩罚，相对于回程链路已满足需求而部分终端未满足 QoS 需求的情况，惩罚较小。r_{p_n} 计算式为

$$r_{p_n}=\left(\sum_{u_h\in\mathcal{U}}x^{b_n}_{u_h}q_h+\sum_{u'_l\in\mathcal{U}'}y^{b_n}_{u'_l}q'_l\right)+\mathrm{BH}_n-2\left(\sum_{u_h\in\mathcal{U}}x^{b_n}_{u_h}+\sum_{u'_l\in\mathcal{U}'}y^{b_n}_{u'_l}\right) \tag{12a}$$

其中，r_{p_0} 表示与 IAB 供体直连的电力终端中未满足 QoS 需求的数量，其计算式为

$$r_{p_0}=\left(\sum_{u_h\in\mathcal{U}}x^{b_0}_{u_h}q_h+\sum_{u'_l\in\mathcal{U}'}y^{b_0}_{u'_l}q'_l\right)-\left(\sum_{u_h\in\mathcal{U}}x^{b_0}_{u_h}+\sum_{u'_l\in\mathcal{U}'}y^{b_0}_{u'_l}\right) \tag{12b}$$

② QoS 奖励项

针对 QoS 约束，除了上述设置的 QoS 惩罚项，同样为满足 QoS 需求的终端设置了奖励项。将整个系统中满足 QoS 需求的终端数量作为 QoS 奖励项，其计算式为

$$r_{e_1}=\left(\sum_{u_h\in\mathcal{U}}x^{b_0}_{u_h}q_h+\sum_{u'_l\in\mathcal{U}'}y^{b_0}_{u'_l}q'_l\right)+\mathrm{BH}_0 \tag{13}$$

这里假设只有当回程链路能够将接入链路数据完全回传时，IAB 节点下的终端才能满足 QoS 需求。

为使非轻量化终端获得更多频谱资源，将非轻量化终端的传输速率作为速率奖励项 r_{e_2}，计算式为

$$r_{e_2}=\sum_{b_n\in\mathcal{B}}\sum_{u_h\in\mathcal{U}}\mu^{b_n}_{u_h} \tag{14}$$

结合上述奖励项和惩罚项，获得最终的奖励函数，其计算式为

$$r=\begin{cases}\alpha_1 r_p+\alpha_2 r_{e_1}, & r_p\neq 0\\ \alpha_1 r_p+\alpha_2 r_{e_1}+\alpha_3 r_{e_2}, & r_p=0\end{cases} \tag{15}$$

其中，α_1、α_2、α_3 是归一化因子，可使 r_p、r_{e_1}、r_{e_2} 处于同一量纲上。若 QoS 惩罚项 $r_p\neq 0$，表示还有终端未满足 QoS 需求，分配频谱资源时优先考虑满足终端的 QoS 需求；相反，当所有终端都满足要求时，频谱分配策略将速率奖励项 r_{e_2} 考虑在内对速率进行优化。

2.2 近端策略优化

为获得可靠的学习性能和数据速率，本文使用 PPO 算法进行求解。PPO 的基本思想是采用一种重要性采样的方法，通过比较当前策略和旧策略之间的差异，在一个可接受的最大步长范围内对策略进行更新，有效避免策略更新过大或过小的情况，从而保证算法的稳定性和收敛性[26]。

在基于策略梯度方法的基础上，对目标函数进行变换，PPO 的目标函数可以表示为

$$\hat{J}(\theta)=\hat{\mathbb{E}}_t\left[\frac{\pi_\theta(\boldsymbol{a}_t\mid o_t)}{\pi_{\theta_{\text{old}}}(\boldsymbol{a}_t\mid o_t)}\hat{\boldsymbol{A}}(o_t,\boldsymbol{a}_t)\right]=\hat{\mathbb{E}}_t\left[\psi(\theta)\hat{\boldsymbol{A}}(o_t,\boldsymbol{a}_t)\right] \tag{16}$$

其中，$\psi(\theta)$ 是当前策略 $\pi_\theta(\boldsymbol{a}_t\mid o_t)$ 与旧策略 $\pi_{\theta_{\text{old}}}(\boldsymbol{a}_t\mid o_t)$ 的比值，$\hat{\boldsymbol{A}}(o_t,\boldsymbol{a}_t)$ 是利用广义优势估计（GAE, general advantage estimation）[27]的联合优势函数。

通过对当前策略和旧策略之间的差异施加约束，即一个裁剪后的损失函数 $\mathcal{L}(\theta)$ 来更新网络参数 θ，其计算式为

$$\mathcal{L}(\theta)=\hat{\mathbb{E}}_t\left[\min(\psi(\theta)\hat{\boldsymbol{A}}_t,\text{clip}(\psi(\theta),1-\varepsilon,1+\varepsilon)\hat{\boldsymbol{A}}_t\right] \tag{17}$$

其中，clip 函数将 $\psi(\theta)$ 限制在 $[1-\varepsilon,1+\varepsilon]$ 内以避免策略更新过大，其计算式为

$$\text{clip}(\psi(\theta),1-\varepsilon,1+\varepsilon)=\begin{cases}1+\varepsilon\ ,\ \psi(\theta)\geqslant 1+\varepsilon\\ 1-\varepsilon\ \ ,\psi(\theta)<1-\varepsilon\end{cases} \tag{17a}$$

2.3 基于 MAPPO 的 IAB 系统频谱分配

将 PPO 算法拓展到多智能体问题中可得到 MAPPO 算法[28]，基于此，本文提出一种集中训练分布执行的 MAPPO 算法以解决 IAB 系统中的频谱分配问题。如图 2 所示，MAPPO 算法中包括 N 个智能体，每个基站都作为一个智能体拥有一个动作网络 π 和一个评价网络 V，动作网络作为策略网络指导智能体做决策，评价网络作为值网络评价动作网络质量；θ^n 和 θ^n_{old} 分别是当前策略网络和旧策略网络的网络参数；φ^n 和 φ^n_{old} 分别是当前值网络和旧值网络的网络参数。MAPPO 算法包含集中训练阶段和分布执行阶段。在集中训练阶段，智能体 n 从环境观测到本地状态信息 o^n 作为动作网络的输入，动作网络根据 o^n 做出动作 a^n 后得到环境的反馈即奖励 r^n。所有智能体做完动作后，分别将状态信息、动作信息以及奖励上传至各智能体的评价网络，各个评价网络依据全局信息计算联合优势函数 $\hat{\boldsymbol{A}}$ 指导各自的动作网络更新。在分布执行阶段，图中虚线部分可去除，仅剩下灰色区域中动作网络与环境的交互过程，即每个智能体观察到本地信息后可根据训练完成的动作网络直接做出决策。

在集中训练阶段，动作网络根据智能体观察到的局部状态信息做出决策，为尽量做出最佳决策，不断根据评价网络的全局信息对网络进行更新。利用裁剪后的损失函数 $\text{loss}(\theta^n)$ 更新策略网络参数 θ^n，其计算式为

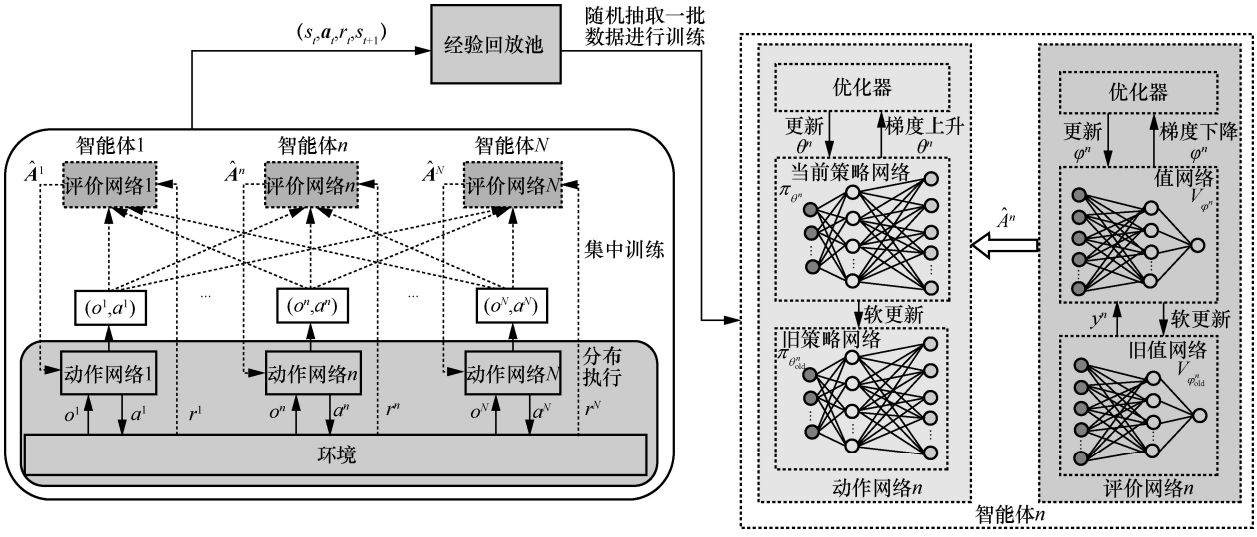

图 2 集中训练分布执行的 MAPPO 算法

$$\text{loss}(\theta^n)=\hat{\mathbb{E}}_t\left[\min(\psi(\theta^n)\hat{\boldsymbol{A}}_t^n,\text{clip}(\psi(\theta^n),1-\varepsilon,1+\varepsilon)\hat{\boldsymbol{A}}_t^n)\right] \tag{18}$$

其中，$\hat{\boldsymbol{A}}_t^n$ 是评价网络利用全局信息采用 GAE 的联合优势函数，其计算式为

$$\hat{\boldsymbol{A}}_t^n=\boldsymbol{\delta}_t^n+(\gamma\lambda)\boldsymbol{\delta}_{t+1}^n+\cdots+(\gamma\lambda)^{T-t+1}\boldsymbol{\delta}_{T-1}^n \tag{18a}$$

$$\boldsymbol{\delta}_t^n=r_t+\gamma V_{\varphi_{\text{old}}^n}(s_{t+1},\boldsymbol{a}_{t+1})-V_{\varphi_{\text{old}}^n}(s_t,\boldsymbol{a}_t) \tag{18b}$$

其中，γ 为折扣率，λ 为平滑系数。

每个智能体的评价网络根据全局信息估计联合优势函数，从全局角度评估动作网络，指导动作网络更新。评价网络通过最小化损失函数 $\text{loss}(\varphi^n)$ 更新网络参数 φ^n，其计算式为

$$\text{loss}(\varphi^n)=\hat{\mathbb{E}}_t\left[(\boldsymbol{y}_t^n-V_{\varphi^n}(s_t,\boldsymbol{a}_t))^2\right] \tag{19}$$

其中，$\boldsymbol{y}_t^n=\hat{\boldsymbol{A}}_t^n+V_{\varphi_{\text{old}}^n}(s_t)$ 是折扣回报。

算法 1 具体描述了利用 MAPPO 进行频谱分配的过程。步骤 1)~步骤 2)设置训练回合 episode，并初始化状态信息；步骤 3)~步骤 9)为数据收集阶段，每个基站利用旧策略网络 $\pi_{\theta_{\text{old}}}$ 与环境交互，智能体 n 在一个 episode 中收集 T 步的轨迹 τ^n，根据式(18a)估计联合优势函数，并将这些轨迹存入经验回放池。步骤 10)~步骤 18)为训练阶段，使用上述收集到的数据在 K 个 epoch 里迭代地更新参数。步骤 19)在每训练完一个 episode 后更新旧策略网络及旧值网络参数。

算法 1　基于 MAPPO 的频谱分配算法

初始化　评价网络 V_{φ_n}、动作网络 π_{θ_n}、旧策略网络参数 $\theta_{\text{old}}^n\leftarrow\theta^n$、旧值网络参数 $\varphi_{\text{old}}^n\leftarrow\varphi^n$、经验回放池 $\mathcal{D}$

1) for episode = 1, 2, ⋯, L
2)　$s_1=$ 初始状态
3)　for $t=1,2,\cdots,T$
4)　　智能体 n 根据 $\pi_{\theta_{\text{old}}^n}\left(a_t^n\mid o_t^n\right)$ 分配子信道
5)　　获得奖励 r_t^n、下一时刻的本地状态 o_{t+1}^n 以及全局状态 s_{t+1}
6)　end for
7)　获得在 T 个时隙内每个基站与环境交互的一条轨迹 $\tau^n=\left\{o_t^n,a_t^n,r_t^n,o_{t+1}^n\right\}_{t=1}^T$
8)　根据式(18a)计算优势函数 $\left\{\hat{\boldsymbol{A}}_t^n(s_t,\boldsymbol{a}_t)\right\}_{t=1}^T$
9)　将 $\left\{\left[o_t^n,a_t^n,r_t^n,o_{t+1}^n,\hat{\boldsymbol{A}}_t^n(s_t,\boldsymbol{a}_t)\right]_{n=1}^N\right\}_{t=1}^T$ 存入经验回放池 $\mathcal{D}$ 中
10)　for $k=1,2,\cdots,K$
11)　　for $j=0,1,\cdots,\dfrac{T}{B-1}$
12)　　$D_j=\left\{\left[o_t^n,a_t^n,r_t^n,o_{t+1}^n,\hat{\boldsymbol{A}}_t^n(s_t,\boldsymbol{a}_t)\right]_{n=1}^N\right\}_{t=1+Bj}^{B(j+1)}$
13)　　for n =1, 2, ⋯, N
14)　　根据式(18)计算 $\text{loss}(\theta^n)$，利用优化器和反向传播对 θ^n 进行梯度上升更新
15)　　根据式(19)计算 $\text{loss}(\varphi_n)$，利用优化器和反向传播对 φ_n 进行梯度下降更新
16)　　end for
17)　end for
18)　end for
19)　智能体 n 更新旧策略网络参数 $\theta_{\text{old}}^n\leftarrow\theta^n$，旧值网络参数 $\varphi_{\text{old}}^n\leftarrow\varphi^n$
20) end for

在分布执行阶段，不再需要评价网络，每个基站都拥有一个经过训练的动作网络，根据观察到的本地状态信息即可做出决策，在整个执行过程中仅进行了一个正向传播过程，与训练阶段相比，极大地减少了时间消耗和计算资源消耗。

2.4　对比算法介绍

为验证本文提出的基于 MAPPO 的频谱分配算法性能，将平均分配、PSO 算法、MADDPG 算法作为对比算法，以下是这几种算法的简要介绍。

1) 平均分配。随机将每个基站下的信道平均分配给其下连接的各个用户，直到找到一个可行解，即满足约束条件。

2) PSO。该算法是一种启发式优化算法，模拟鸟群或鱼群中个体之间的协作与信息共享，通过迭代优化来搜索问题的最优解。粒子可以看作问题的一个可能的解决方案，每个粒子都具有位置、速度和适应度函数这 3 个特征。其中，位置代表问题的一个解向量，每次迭代粒子通过局部最优值与全局最优值更新速度；速度用来调整位置移动方向以优化适应度值；适应度函数将问题的解映射到一个适应度值，用于评估解的质量。在每轮的迭代中，第 i 个粒子的位置为

$\boldsymbol{d}_i=[\boldsymbol{d}_i^0,\boldsymbol{d}_i^1,\cdots,\boldsymbol{d}_i^N]$，速度表示为$\boldsymbol{v}_i=[\boldsymbol{v}_i^0,\boldsymbol{v}_i^1,\cdots,\boldsymbol{v}_i^N]$，第 i 个粒子的局部最优位置为$\boldsymbol{\kappa}_i=[\boldsymbol{\kappa}_i^1,\boldsymbol{\kappa}_i^2,\cdots,\boldsymbol{\kappa}_i^N]$，全局最优位置为$\boldsymbol{\kappa}_g=[\boldsymbol{\kappa}_g^1,\boldsymbol{\kappa}_g^2,\cdots,\boldsymbol{\kappa}_g^N]$。其中，粒子速度的更新式为

$$\boldsymbol{v}_i^n=\omega\boldsymbol{v}_i^n+e_1f_1\left(\boldsymbol{\kappa}_i^n-\boldsymbol{v}_i^n\right)+e_2f_2\left(\boldsymbol{\kappa}_g^n-\boldsymbol{v}_i^n\right) \tag{20}$$

其中，ω 为惯性因子，e_1 和 e_2 分别为个体加速度和群体加速度，f_1 和 f_2 分别为 2 个[0,1]内的随机数。适应度函数与式(15)定义的奖励一致。

3) MADDPG。该算法同样基于集中训练分布执行框架，将基于策略的 DDPG 拓展至多智能体问题中得到 MADDPG 算法[29]。在 MADDPG 中，每个智能体同样各自训练一个接收局部信息的动作网络和一个接收全局信息的评价网络，其损失函数的计算式为

$$\text{loss}(\theta^n)=\hat{\mathbb{E}}_t\left(\nabla_{\theta^n}\pi_{\theta^n}(o_t^n)\nabla_{\boldsymbol{a}^n}V_{\varphi^n}(s_t,\boldsymbol{a}_t)\right)\Big|_{\boldsymbol{a}_t^n=\pi_{\theta^n}(o_t^n)} \tag{21}$$

2.5 时间复杂度分析

1) 平均分配。设找到可行解的迭代步数为 T_{AA}，则平均分配的时间复杂度为 $O(N(L+H)T_{\text{AA}})$，其中 T_{AA} 随机性较大。

2) PSO。设总迭代步数为 T_{PSO}，则 PSO 算法的时间复杂度为 $O(XNMT_{\text{PSO}})$，其中 X 为粒子个数。

3) MADDPG、MAPPO。训练一次神经网络的时间复杂度为[30]

$$2\sum_{j=0}^{J}n_{\text{actor},j}n_{\text{actor},j+1}+2\sum_{k=0}^{K}n_{\text{critic},k}n_{\text{critic},k+1}=O\left(\sum_{j=0}^{J}n_{\text{actor},j}n_{\text{actor},j+1}+\sum_{k=0}^{K}n_{\text{critic},k}n_{\text{critic},k+1}\right) \tag{22}$$

其中，$n_{\text{actor},j}$ 为动作网络第 j 层的神经元个数，$n_{\text{critic},l}$ 为评价网络第 k 层的神经元个数。设每次训练的迭代步数为 T_{upd}，训练回合数为 T_{episode}，每个回合的步长为 T_{step}，则 MADDPG 与 MAPPO 在训练阶段的时间复杂度为

$$O\left(T_{\text{episode}}T_{\text{step}}T_{\text{upd}}N\left(\sum_{j=0}^{J}n_{\text{actor},j}n_{\text{actor},j+1}+\sum_{k=0}^{K}n_{\text{critic},k}n_{\text{critic},k+1}\right)\right) \tag{23}$$

分布执行阶段 MADDPG 与 MAPPO 仅需使用训练好的模型，时间复杂度为 $O(T_{\text{step}}N)$，满足实时网络条件下在线决策时间的要求。

综上所述，各算法的时间复杂度如表 1 所示。

表 1 各算法的时间复杂度

算法	时间复杂度
平均分配	$O(N(L+H)T_{\text{AA}})$
PSO	$O(XNMT_{\text{PSO}})$
MAPPO、MADDPG	$O(T_{\text{step}}N)$

3 仿真分析

3.1 系统参数设置

本文在 Python3.7 和 PyTorch 环境下对 IAB 系统中基于 MAPPO 的频谱分配算法进行了仿真实验。考虑由一个 IAB 供体和 2 个 IAB 节点组成的 IAB 系统，IAB 供体的覆盖半径为 400 m，IAB 节点和电力终端均匀分布在 IAB 供体的覆盖范围内，IAB 节点的覆盖半径为 150 m，电力终端根据与 IAB 供体与 IAB 节点的距离以及 IAB 节点的覆盖范围就近选择基站进行连接。路径损耗模型使用 3GPP 标准中的 5GCM Uma，具体参数设置如表 2 所示。

表 2 仿真场景参数

参数	设置值
每个基站信道总带宽/MHz	50
IAB 节点路径损耗/dB	$28+22\lg(d)+20\lg(f_{\text{c}})$
电力终端路径损耗/dB	$32.4+21\lg(d)+20\lg(f_{\text{c}})$
载波频率 f_{c} /GHz	2.4
IAB 节点发射功率/dBm	33
电力终端发射功率/dBm	20
噪声功率谱密度/ (dBm·Hz^{-1})	−174
自干扰/dB	−70
子信道衰落模型	瑞利衰落

为方便对比 MAPPO 和 MADDPG 性能，设定两者动作网络与评价网络参数一致。动作网络和评价网络均为 3 层全连接层，动作网络神经元个数为 128、128、64，评价网络神经元个数为 256、128、128。折扣率 γ 设置为 0.9，使用 Adam 作为优化器，每次训练 1 000 个回合，每个回合的步长为 $T=400$。

3.2 仿真结果分析

考虑 12 个电力终端，其中，轻量化终端和非轻量化终端的数量比例为 1:1，可用子信道数为 25。MAPPO 与 MADDPG 在训练过程中所获奖励如图 3 所示。从图 3 可以看到，所提 MAPPO 算法显著优于 MADDPG

算法，MAPPO 更早达到收敛状态且在 MADDPG 基础上将奖励值提高了 15.3%。尽管 MADDPG 也采用集中训练分布执行的框架，但确定性策略的探索能力较差，容易陷入局部最优解，此外，MADDPG 容易高估值函数，从而导致训练曲线的波动。因此在本文考虑的 IAB 系统中，MADDPG 的性能相对低于 MAPPO。MAPPO 采用具有全局信息的评价网络和本地信息的动作网络来实现各个基站之间的合作，并增加了行动熵奖励，以鼓励基站对分配子信道的探索，与 MADDPG 相比，MAPPO 获得的高性能充分证明了本文所提频谱分配算法的有效性。

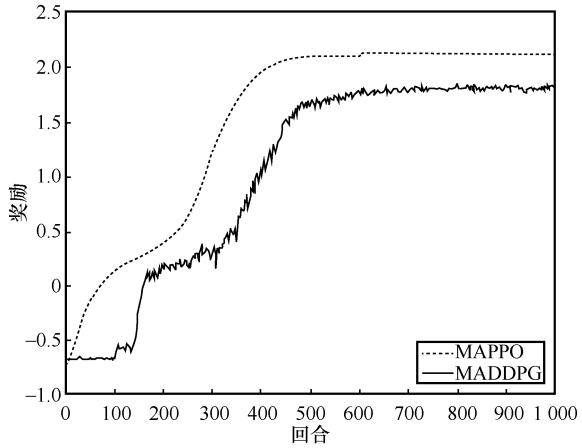

图 3　MAPPO 与 MADDPG 在训练过程中所获奖励

最大化系统能效的优化目标在问题建模时被转换为在满足轻量化终端的 QoS 需求前提下，最大化非轻量化终端的传输速率，下面将从终端速率、系统总速率以及非轻量化终端的速率占比几个方面分析算法的性能。

6 个轻量化终端和 6 个非轻量化终端随着训练过程的速率变化如图 4 所示。图 4 仅结合优化目标着重分析轻量化终端与非轻量化终端在整个训练过程中的整体变化趋势，不对各终端速率进行区别。从图 4 可以看出，在前 300 回合中，轻量化终端和非轻量化终端的速率相当，初期训练先满足各电力终端的 QoS 需求，终端速率上下波动，随着 QoS 需求得到满足，训练目标转为优化非轻量化终端速率，在这个过程中频谱资源在满足约束条件下逐渐向非轻量化终端倾斜，非轻量化终端速率逐渐上升，轻量化终端速率逐渐下降，在 600 回合左右，各终端速率收敛到一个较平稳的状态。

IAB 系统中 2 个 IAB 节点的接入链路和回程链路的速率变化如图 5 所示。初始时刻，2 个节点的回程链路速率均小于接入链路速率，回程链路成为瓶颈链路限制 IAB 系统中两跳链路的传输。经过几十回合的训练后，智能体将更多的子信道分配给 IAB 节点，IAB 节点为保证回程链路速率的平衡选择丢弃部分产生过大干扰的子信道。当回程链路不再成为瓶颈链路后，各智能体逐步选择更合适的信道优化非轻量化终端的速率，节点 2 的接入链路速率随着回程链路速率逐步上升，并逐渐减少两者之间的差距，在 500 回合左右达到一个稳定的状态。节点 1 的回程链路速率首先经历了一个快速的上升过程，但是接入链路速率并未跟上，导致节点 1 回程链路占用的信道资源大部分将会浪费，因此在优化非轻量化终端速率过程中，回程链路速率逐步下降而接入链路速率逐渐上升并在 500 回合左右稳定。

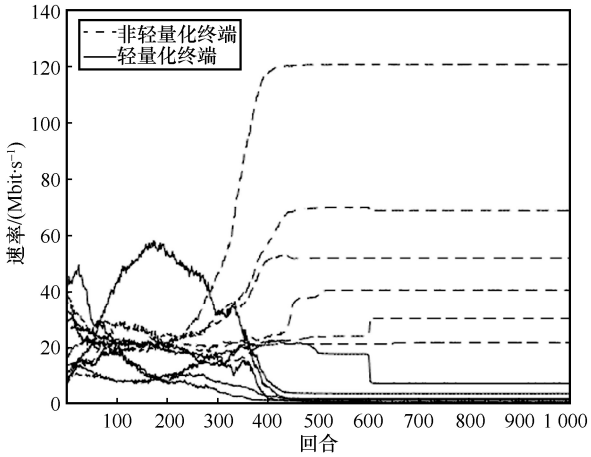

图 4　各电力终端随着训练过程的速率变化

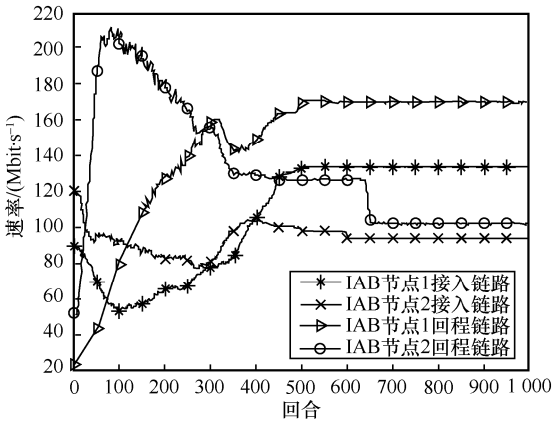

图 5　IAB 节点的接入链路和回程链路的速率变化

考虑总带宽不变，改变子信道数是否可以获得更高的速率。不同算法下改变子信道数的系统总速率如图 6 所示。从图 6 可以看出，与随机性较大的平均分配相比，使用 MAPPO、MADDPG 以及 PSO 获得的

系统总速率得到显著提升。MAPPO 与 MADDPG 相比，系统总速率最高可以提高 13.2%，与 PSO 相比系统总速率最高可以提高 24.3%。进行平均分配时子信道全利用，每个基站将子信道平均分配给其下所连接的电力终端或 IAB 节点，不考虑终端与信道质量的适配性，接入链路和回程链路速率也未得到平衡，IAB 节点下的电力终端数据无法向上回传。使用 MAPPO 或 MADDPG 时，IAB 节点观察接入链路和回程链路状态，根据每条链路状态信息分配信道质量与之匹配的信道，选择丢弃部分对其他链路干扰较大的子信道，总速率得到极大提升。PSO 在逐次迭代的过程中利用适应度函数逐渐探索到干扰较小的分配方案，获得较高的总速率，当子信道数为 25 时，PSO 算法的性能优于 MADDPG。从图 6 中可以看到，使用 MAPPO 或 MADDPG 分配子信道时，随着子信道数的增加，系统总速率明显上升，这是因为总带宽不变，当子信道数变多时，带宽利用率进一步增加，但 PSO 受环境影响较大，在某一状态易陷入局部最优解，当信道数为 40 时速率反而下降。

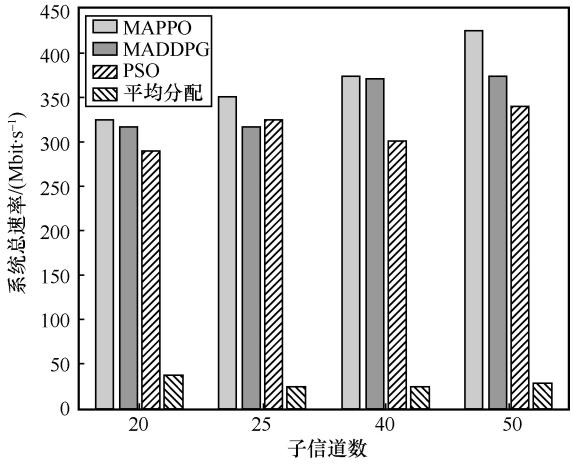

图 6 不同算法下改变子信道数的系统总速率

图 7 展示了不同算法下改变子信道数的非轻量化终端总速率占比。从图 7 可以看出，使用 MAPPO 算法时，资源尽可能向非轻量化终端倾斜，但 MADDPG 性能随子信道数的增加而下降。信道数增加导致动作空间变得更复杂。在处理复杂动作空间上的性能方面，MADDPG 不及 MAPPO 稳定。此外，从图 6 和图 7 可以明显看到，使用 PSO 时非轻量化终端速率占比显著高于 MADDPG，但系统总速率却不及 MADDPG，进一步体现了 PSO 易陷入局部最优的问题。

图 8 展示了不同用户数下不同算法获得的系统总速率。从图 8 可以看出，使用 MAPPO、MADDPG 以及 PSO 算法获得的系统总速率依旧远优于平均分配。MAPPO 与 MADDPG 相比，系统总速率最高可以提升 21%，与 PSO 相比系统总速率最高可以提升 25.2%。增加用户数的同时降低了 MAPPO、MADDPG 和 PSO 的性能，当用户数增加时，用户之间对资源的竞争更加强烈，每个用户能获得的子信道变少从而导致性能下降，但本文所提的 MAPPO 算法所获性能一直处于最优状态。

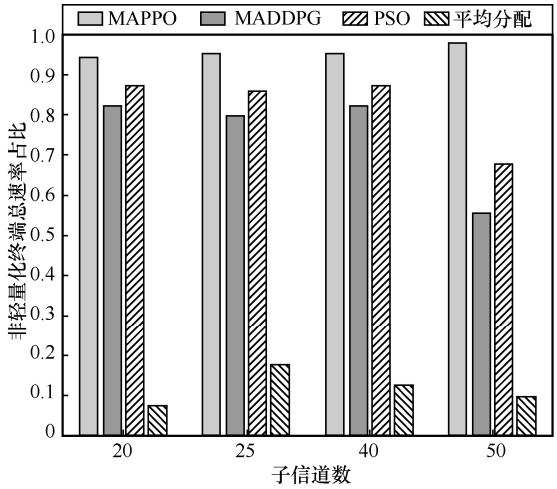

图 7 不同算法下改变子信道数的非轻量化终端总速率占比

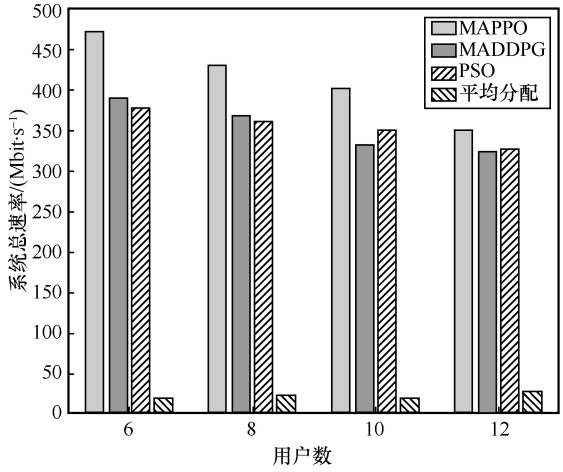

图 8 不同用户数下不同算法获得的系统总速率

图 9 给出了不同算法下改变用户数非轻量化终端总速率占比。从图 9 可以看出，使用 MAPPO 算法时，非轻量化终端速率占比始终领先于其他 3 种算法，不管是从获得最大系统总速率方面，还是优先给非轻量化终端分配资源方面，MAPPO 算法都表现出最优的性能。此外，当用户数变化时，PSO 算法在系统总速率上与 MADDPG 性能相当，但在非轻量化终端总速率占比中却优于 MADDPG，这

在一定程度上反映了 MADDPG 更易受到环境影响从而导致性能不稳定的问题。

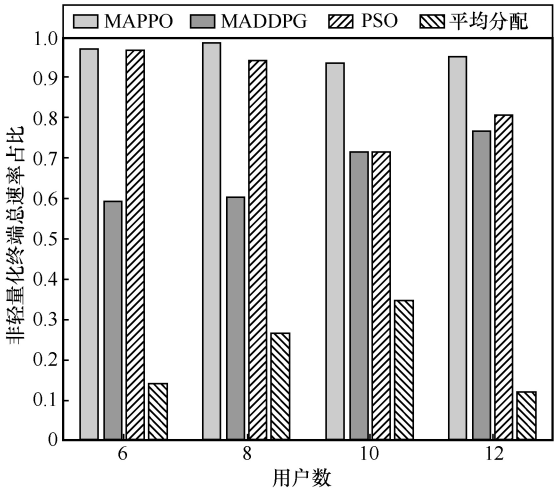

图 9　不同算法下改变用户数非轻量化终端总速率占比

4　结束语

本文针对智能电网中部署的 IAB 系统，考虑了轻量化和非轻量化终端业务的不同通信需求，将频谱分配问题建模为最大化系统总能效的非凸混合整数规划。为求解前述问题，将其构建为一个部分可观测的马尔可夫决策过程并转换为完全协作的多智能体问题，进而提出了一种集中训练分布执行框架下基于 MAPPO 的频谱分配算法。从理论上分析了算法的时间复杂度，并通过仿真验证了所提算法的性能。仿真结果表明，所提算法能够有效减少层内与层间干扰，平衡接入链路与回程链路速率；与 MADDPG 相比收敛速度更快，系统总速率可以提高 21%；在不同用户数和不同子信道数下的系统性能均显著优于 PSO，系统总速率最高可以提高 25.2%。

参考文献：

[1] DILEEP G. A survey on smart grid technologies and applications[J]. Renewable Energy, 2020, 146: 2589-2625.

[2] CHI Y Y, ZHANG Y, LIU Y, et al. Deep reinforcement learning based edge computing network aided resource allocation algorithm for smart grid[J]. IEEE Access, 2022, 11: 6541-6550.

[3] ANDREWS J G, BUZZI S, CHOI W, et al. What will 5G be?[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1065-1082.

[4] KONG P Y. Radio resource allocation scheme for reliable demand response management using D2D communications in smart grid[J]. IEEE Transactions on Smart Grid, 2020, 11(3): 2417-2426.

[5] YANG J J, LIU G, REN J, et al. Resource allocation for intelligent reflecting surface-assisted cooperative NOMA-URLLC networks in smart grid[C]//Proceedings of 2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids. Piscataway: IEEE Press, 2022: 83-89.

[6] CAO Z J, LIN J, WAN C, et al. Optimal cloud computing resource allocation for demand side management in smart grid[J]. IEEE Transactions on Smart Grid, 2017, 8(4): 1943-1955.

[7] SUN M Y, YUAN Y Z, MA K, et al. Spectrum allocation and computing resources optimization for demand-side cooperative communications in smart grid[J]. IEEE Transactions on Smart Grid, 2022, 13(3): 1967-1975.

[8] LI Z, LIANG Q L. Capacity optimization in heterogeneous home area networks with application to smart grid[J]. IEEE Transactions on Vehicular Technology, 2016, 65(2): 699-706.

[9] YIN F F, ZENG M Y, ZHANG Z L, et al. Coded caching for smart grid enabled HetNets with resource allocation and energy cooperation[J]. IEEE Transactions on Vehicular Technology, 2020, 69(10): 12058-12071.

[10] LIU L L, ZHANG Z Z, WANG N, et al. Online resource management of heterogeneous cellular networks powered by grid-connected smart micro grids[J]. IEEE Transactions on Wireless Communications, 2022, 21(10): 8416-8430.

[11] GUNGOR V C, SAHIN D, KOCAK T, et al. A survey on smart grid potential applications and communication requirements[J]. IEEE Transactions on Industrial Informatics, 2013, 9(1): 28-42.

[12] 3GPP. Study on integrated access and backhaul (release 16): TR 38.874[S]. 2018.

[13] BELAID M N, AUDEBERT V, DENEUVILLE B, et al. Smart grid critical traffic routing and link scheduling in 5G IAB networks[C]//Proceedings of 2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grid. Piscataway: IEEE Press, 2022: 76-82.

[14] YASHIMA T, NISHIYAMA H. Analysis of optimal bandwidth partitioning ratio in full-duplex integrated access and backhaul[C]//Proceedings of 2022 IEEE International Conference on Communications Workshops. Piscataway: IEEE Press, 2022: 1-6.

[15] ZHANG S M, XU X D, SUN M Y, et al. Joint spectrum and power allocation in 5G integrated access and backhaul networks at mmWave band[C]//Proceedings of 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications. Piscataway: IEEE Press, 2020: 1-7.

[16] PAGIN M, ZUGNO T, POLESE M, et al. Resource management for 5G NR integrated access and backhaul: a semi-centralized approach[J]. IEEE Transactions on Wireless Communications, 2022, 21(2): 753-767.

[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] 喻鹏, 张俊也, 李文璟, 等. 移动边缘网络中基于双深度 Q 学习的高能效资源分配方法[J]. 通信学报, 2020, 41(12): 148-161.
YU P, ZHANG J Y, LI W J, et al. Energy-efficient resource allocation

method in mobile edge network based on double deep Q-learning[J]. Journal on Communications, 2020, 41(12): 148-161.

[19] ZHU X Y, WANG J, LI J M, et al. A scheme for uplink NOMA communication with intelligent resource allocation for mMTC traffic over eMBB traffic[C]//Proceedings of 2022 IEEE 95th Vehicular Technology Conference. Piscataway: IEEE Press, 2022: 1-5.

[20] 徐思雅, 邢逸斐, 郭少勇, 等. 基于深度强化学习的能源互联网智能巡检任务分配机制[J]. 通信学报, 2021, 42(5): 191-204.
XU S Y, XING Y F, GUO S Y, et al. Deep reinforcement learning based task allocation mechanism for intelligent inspection in energy Internet[J]. Journal on Communications, 2021, 42(5): 191-204.

[21] LEI W L, YE Y, XIAO M. Deep reinforcement learning-based spectrum allocation in integrated access and backhaul networks[J]. IEEE Transactions on Cognitive Communications and Networking, 2020, 6(3): 970-979.

[22] CHENG Q Q, WEI Z Q, YUAN J H. Deep reinforcement learning-based spectrum allocation and power management for IAB networks[C]//Proceedings of 2021 IEEE International Conference on Communications Workshops (ICC Workshops). Piscataway: IEEE Press, 2021: 1-6.

[23] CHENG Z P, MIN M H, GAO Z B, et al. Joint task offloading and resource allocation for mobile edge computing in ultra-dense network[C]//Proceedings of 2021 IEEE Global Communications Conference. Piscataway: IEEE Press, 2021: 1-6.

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

[25] 周凡, 王鸿, 宋荣方. 密集异构蜂窝网络中基于深度强化学习的下行链路功率分配算法[J]. 南京邮电大学学报(自然科学版), 2021, 41(2): 12-19.
ZHOU F, WANG H, SONG R F. Deep reinforcement learning based downlink power allocation algorithm in dense heterogeneous cellular networks[J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science), 2021, 41(2): 12-19.

[26] KIM Y, LIM H. Multi-agent reinforcement learning-based resource management for end-to-end network slicing[J]. IEEE Access, 2021, 9: 56178-56190.

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

[28] YU C, VELU A, VINITSKY E, et al. The surprising effectiveness of ppo in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems, 2022, 35: 24611-24624.

[29] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[J]. arXiv Preprint, arXiv: 1706.02275, 2017.

[30] HOU W J, WEN H, SONG H H, et al. Multiagent deep reinforcement learning for task offloading and resource allocation in cybertwin-based networks[J]. IEEE Internet of Things Journal, 2021, 8(22): 16256-16268.

[作者简介]

燕锋（1983− ），男，湖北天门人，博士，东南大学副教授，主要研究方向为无人机自组网、卫星互联网、无线传感器网络等。

林晓薇（1999− ），女，广西桂林人，东南大学硕士生，主要研究方向为无线网络资源管理、强化学习应用等。

李正浩（1991− ），男，山东济南人，国网山东省电力公司信息通信公司工程师，主要研究方向为云计算、5G 通信、数字化等方面。

徐霞（1986− ），女，山东成武人，国网山东省电力公司济南供电公司高级工程师，主要研究方向为电力系统智慧物联网、网络资源优化等。

夏玮玮（1975− ），女，江苏句容人，博士，东南大学副研究员，主要研究方向为无线网络资源管理、边缘计算、泛在网络与短距离无线通信等。

沈连丰（1952− ），男，江苏邳州人，东南大学教授、博士生导师，主要研究方向为宽带移动通信、短距离无线通信和泛在网络等。