异构云无线接入网下基于功率域NOMA的能效优化算法

唐 伦 李子煜* 管令进 陈前斌
(重庆邮电大学通信与信息工程学院 重庆 400065)
(重庆邮电大学移动通信技术重点实验室 重庆 400065)

摘 要：针对异构云无线接入网络的频谱效率和能效问题，该文提出一种基于功率域-非正交多址接入(PD-NOMA)的能效优化算法。首先，该算法以队列稳定和前传链路容量为约束，联合优化用户关联、功率分配和资源块分配，并建立网络能效和用户公平的联合优化模型；其次，由于系统的状态空间和动作空间都是高维且具有连续性，研究问题为连续域的NP-hard问题，进而引入置信域策略优化(TRPO)算法，高效地解决连续域问题；最后，针对TRPO算法的标准解法产生的计算量较为庞大，采用近端策略优化(PPO)算法进行优化求解，PPO算法既保证了TRPO算法的可靠性，又有效地降低TRPO的计算复杂度。仿真结果表明，该文所提算法在保证用户公平性约束下，进一步提高了网络能效性能。

关键词：异构云无线接入网络；资源分配；网络能效；深度强化学习
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2021)06-1706-09
DOI: 10.11999/JEIT200327

Energy Efficiency Optimization Algorithm Based On PD-NOMA Under Heterogeneous Cloud Radio Access Networks

TANG Lun LI Ziyu GUAN Lingjin CHEN Qianbin
(*School of Communication and Information Engineering*, *Chongqing University of Posts and Telecommunications*, *Chongqing* 400065, *China*)
(*Key Laboratory of Mobile Communication Technology*, *Chongqing University of Post and Telecommunications*, *Chongqing* 400065, *China*)

Abstract: In view of the spectrum efficiency and energy efficiency of Heterogeneous Cloud Radio Access Networks (H-CRAN), an energy efficiency optimization algorithm based on Power Domain Non-Orthogonal Multiple Access (PD-NOMA) is proposed. First, the algorithm takes queue stability and forward link capacity as constraints, jointly optimizes user association, power allocation and resource block allocation, and it establishes a joint optimization model of network energy efficiency and user fairness. Secondly, because the state space and action space of the system are both high-dimensional and continuity, the research problem is the NP-hard problem of the continuous domain, and then Trust Region Policy Optimization (TRPO) algorithm is introduced to solve efficiently the continuous domain issue. Finally, the amount of calculations generated by the standard solution for the TRPO algorithm is too large, and Proximal Policy Optimization (PPO) algorithm is used to optimize the solution. The PPO algorithm not only ensures the reliability of the TRPO algorithm, but also reduces effectively the TRPO calculation complexity. Simulation results show that the algorithm proposed in this paper improves further the energy efficiency performance of the network under the constraint of ensuring user fairness.

Key words: Heterogeneous Cloud Radio Access Networks(H-CRAN); Resource allocation; Network energy efficiency; Deep Reinforcement Learning(DRL)

收稿日期：2020-04-28；改回日期：2020-10-05；网络出版：2020-10-12
*通信作者： 李子煜 lzy395682410@qq.com
基金项目：国家自然科学基金(62071078)，重庆市教委科学技术研究项目(KJZD-M201800601)，重庆市重大主题专项项目(cstc2019jscx-zdztzxX0006)
Foundation Items: The National Natural Science Foundation of China (62071078), The Science and Technology Research Project of Chongqing Education Commission (KJZD-M201800601), The Major Theme Projects in Chongqing (cstc2019jscxzdztzxX0006)

1　引言

随着智能设备的爆炸性增长，诸如增强现实和虚拟现实等新兴高速率服务以及构建物联网(Internet of Things, IoT)的海量设备，使得设计高效的能效通信系统迫在眉睫，进而实现绿色经济和可持续发展的运营。与4G系统相比，5G系统需要达到1 ms的时延、10倍的频谱效率、100倍的能效以及1000倍的系统容量。作为有前景的新技术和网络体系结构，异构云无线接入网(Heterogeneous Cloud Radio Access Networks, H-CRAN)引起了业界和学术界的极大关注。可以预见，在H-CRAN中将采用各式的多址接入技术，以减轻小区间和小区内的干扰，并改善网络频谱效率和能效。作为一种新的多址方案，非正交多址接入(Non-Orthogonal Multiple Access, NOMA)被认为是有望显著地改善5G移动通信网络的频谱效率和能效的候选方案。文献[1]采用混合多址接入技术提高频谱效率，NOMA技术中的非正交性具有高频效、能效以及低传输时延的潜在优势。因此，本文在H-CRAN的下行传输场景下利用NOMA技术来最大化网络能效。

文献[2]在H-CRAN下行传输场景下研究网络能效性能，联合优化基站选择、子载波分配和功率分配，构建网络能效最大化的目标函数，利用连续凸近似理论进行求解，进而提高H-CRAN的能效性能。文献[3]在异构云无线接入网络的场景下提出一种能效优化算法，利用李雅普诺夫优化理论和拉格朗日对偶分解方法对优化问题进行求解。文献[4]在H-CRAN下行链路场景下，建立了网络总吞吐量最大化的随机优化模型，通过深度强化学习和迁移学习算法，智能化分配无线资源，提高网络的稳定性。

尽管上述的文献在无线资源分配上都取得了较好的研究成果，但仍然需要进一步的改进，主要存在3方面的问题：(1)多数工作忽略了NOMA技术带来的频谱效率和能效优势，同时没有考虑前传容量受限给接入网带来的吞吐量瓶颈，进而与实际的网络场景相脱离，无法取得合适的资源分配方案；(2)大多数研究仍采用传统非线性优化算法，当优化问题出现高维状态空间或动作空间时，可能会导致维度灾问题，使得优化算法陷入局部最优解；(3)尽管深度Q学习对无线资源的自优化具有一定的帮助，但其需要对动作空间进行离散化处理，导致求解的资源分配策略非常不稳定。此外，基于连续域的置信域策略优化(Trust Region Policy Optimization, TRPO)算法产生的计算量较为庞大，导致算法性能得不到有效的提升。

针对上述提出的问题，本文在H-CRAN下提出一种基于功率域-非正交多址接入(Power Domain Non-Orthogonal Multiple Access, PD-NOMA)的能效优化算法。所提算法的主要创新点如下：(1)为提高网络的频谱效率和能效，联合优化用户关联、功率分配和资源块(Resource Block, RB)分配，构建用户公平性和网络能效的优化模型；(2)针对无线网络资源分配的复杂性和动态性难题，引入基于自学习的置信域策略优化算法，大大降低了动作空间的维度，进而避免维度灾问题；(3)针对TRPO算法的标准解法产生的计算量较为庞大，采用近端策略优化(Proximal Policy Optimization, PPO)算法进行优化求解，进一步提高算法效率。

2　问题描述与系统模型

2.1　基于PD-NOMA的异构云无线接入网架构

考虑H-CRAN下行传输场景，如图1所示，建立了一个基于NOMA的H-CRAN架构，远端无线射频单元(Remote Radio Head, RRH)具有天线模块，只需执行射频处理以及简单的基带处理，主要的基带信号处理以及上层协议功能均在集中式基带单元(Base Band Unite, BBU)池中执行，RRH通常部署在热点区域负责海量数据业务的高速传输[5]。高功率节点(High Power Node, HPN)用于全网的控制信息分发，突发业务以及即时信息等低速率数据信息也由HPN承载，确保业务的无缝覆盖[6]。以此同时，采用基于PD-NOMA来提升频谱效率和网络能效，PD-NOMA允许不同用户占用相同的频谱、时间和空间等资源，通过主动引入干扰进一步地提升单用户速率和系统的和速率，尤其是保障了小区边缘用户速率。

2.2　无线通信模型

为了提高网络的能效性能，研究由1个HPN和M个RRHs组成的基于PD-NOMA的H-CRAN系统。用$m=0,1,\cdots,M$来表示基站的集合，其中0表示HPN，$\{1,2,\cdots,M\}$表示RRHs的集合，$l=1,2,\cdots,L$来表示用户的集合，系统中有k个RB来保障用户的通信。用$p_{m,k,l}(t)$表示第t时隙RRH m在资源块k上分配给用户l的功率；$h_{m,k,l}(t)$表示第t时隙RRH m在资源块k到用户l的信道系数，满足$\mathrm{E}[|h_{m,k,l}(t)|^2]=1/d_{m,k,l}^{\psi_{m,k,l}}$，其中$d_{m,k,l}$为用户$l$到RRH m的距离、$\psi_{m,k,l}$表示传输链路的路径损失指数；$n_{m,k,l}\sim\mathrm{CN}(0,\sigma_{m,k,l}^2)$为接收机在用户$i$处的均值为0、方差为$\sigma_{m,k,l}^2$的加性高斯白噪声；基站$m$在资源块$k$上的传输信号$x$为

$$x_{m,k}(t)=\sum_{l\in L}A_{m,k,l}(t)\phi_{m,k,l}(t)\sqrt{p_{m,k,l}(t)}s_{m,k,l}(t) \tag{1}$$

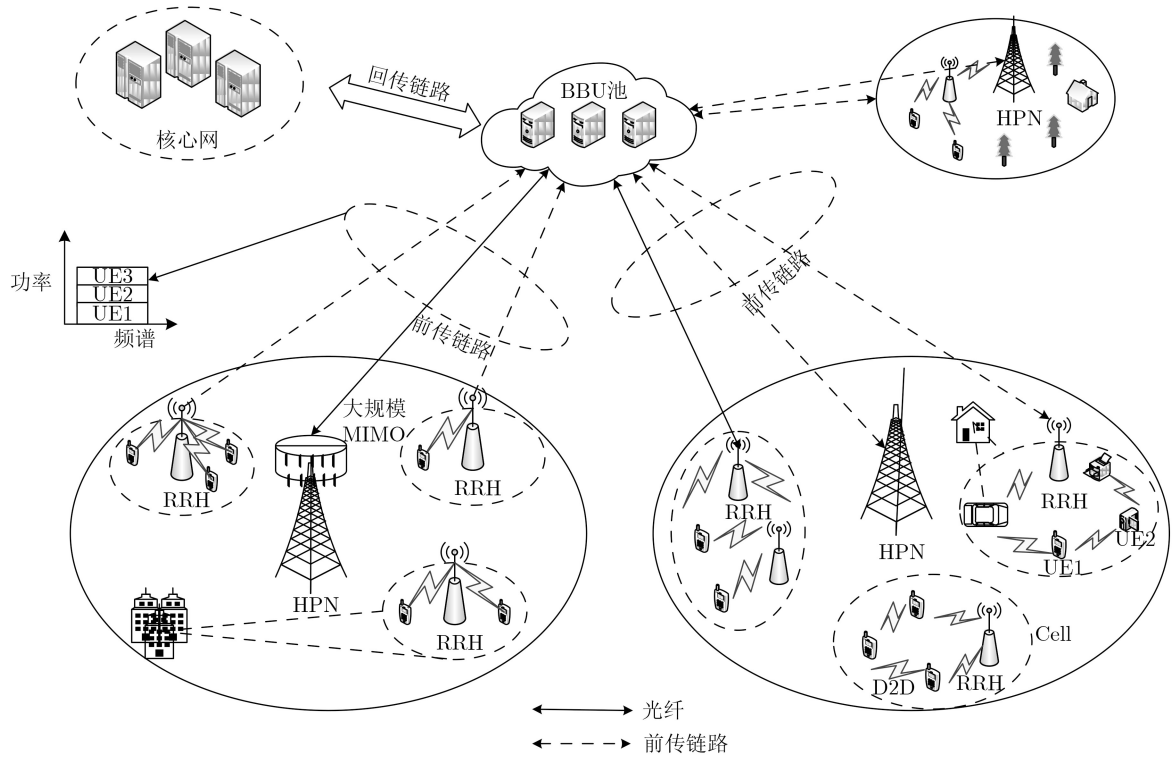

图 1 基于PD-NOMA的异构云无线接入网架构

其中，$A_{m,k,l}(t)$是二进制变量，若使用RB k的用户l关联到RRH m时，$A_{m,k,l}(t)=1$；反之$A_{m,k,l}(t)=0$。$\phi_{m,k,l}(t)$也是二进制变量，当RRH m分配第k个RB给用户l时，$\phi_{m,k,l}(t)=1$，反之$\phi_{m,k,l}(t)=0$。因此，用户l在第t时隙接收到的信号为

$$y_{m,k,l}(t)=h_{m,k,l}(t)\cdot x_{m,k}(t)+n_{m,k,l} \tag{2}$$

为了方便研究，假设用户的噪声是一致的，即$\sigma^2_{m,k,l}=\sigma^2_{m,k},\forall m=1,2,\cdots,M,\ k=1,2,\cdots,K,\ l=1,2,\cdots,L$.并且假设在RB k上信道增益的大小为$h_{m,k,1}>h_{m,k,2}>\cdots>h_{m,k,L}$。NOMA通过结合串行干扰消除(Successive Interference Cancellation, SIC)技术才能取得容量极限，获得更高的频效和能效。本文考虑在不完美的SIC场景下进行资源调度。因此，用户l在第t时隙的信噪比为

$$\gamma_{m,k,l}(t)=\frac{p_{m,k,l}(t)|h_{m,k,l}(t)|^2}{I_{m,k,l}(t)+I_{\text{unsic}}(t)+\sigma^2_{m,k}(t)} \tag{3}$$

其中，$I_{\text{unsic}}(t)=\vartheta\sum_{j=l+1}^{L}p_{m,k,j}(t)|h_{m,k,j}(t)|^2$，$\vartheta$为不完美SIC的影响因子；$\sigma^2_{m,k}(t)$为噪声功率；$I_{m,k,l}(t)=\sum_{i\in L,h_{m,k,l(t)}<h_{m,k,i}(t),i\neq l}C_{m,k,i}(t)\phi_{m,k,i}(t)p_{m,k,i}(t)|h_{m,k,l}(t)|^2+\sum_{v\in M,j\neq m}C_{v,k,i}(t)\phi_{v,k,i}(t)p_{v,k,i}(t)|h_{v,k,l}(t)|^2$，$I_{m,k,l}(t)$中的首项表示由于NOMA技术产生的干扰，后一项表示其他基站产生的干扰，整个网络的总速率为

$$R_{\text{tot}}(t)=\sum_{m\in M}\sum_{k\in K}\sum_{l\in L}\omega_{m,k,l}(t-1)\frac{B}{K}C_{m,k,l}(t)\phi_{m,k,l}(t)\cdot\log_2\left(1+\frac{p_{m,k,l}(t)|h_{m,k,l}(t)|^2}{I_{m,k,l}(t)+I_{\text{unsic}}(t)+\sigma^2_{m,k}(t)}\right) \tag{4}$$

其中，$\omega_{m,k,l}(t-1)$为用户的公平性因子，等价于

$$\omega_{m,k,1}(t-1)=\frac{(|R_{1,\text{QoS}}(t-1)-R_{1,\text{Obt}}(t-1)|^{2\tau})}{\sum_{l\in L,l=1,2\cdots L}(|R_{l,\text{QoS}}(t-1)-R_{l,\text{Obt}}(t-1)|^{2\tau})} \tag{5}$$

2.3 前传链路模型

随着移动设备的大量普及，移动流量也急剧增加，需要一种大容量、高可靠和低时延的传输网络作为前传网络，以此来满足移动用户越来越多的业务需求。在目前的前传网络选择中，无源光网络(Passive Optical Network, PON)具备低成本、大容量的特性，是一种高效可行的前传网络解决方案[7]。PON作为云无线接入网络(Cloud-Radio Access Network, C-RAN)的前传网络，不仅能够满足C-RAN架构对前传链路的传输要求，同时还能应对5G网络带来的高可靠、低时延和低损耗的无线网络需求。

如图2所示，PON是典型的一对多传输网络，其固有无源特性能够为前传链路提供极大的带宽容量和较长距离覆盖等优势，PON称为H-CRAN中

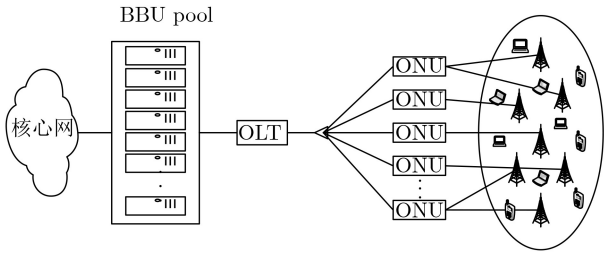

图 2 前传链路框图

光前传网络的最佳选择。因此，前传容量限制的模型为

$$\sum\sum R_{m,k,l}(t) \le \vartheta_m \tag{6}$$

其中，ϑ_m表示第m个RRH的有效前传容量。

2.4 网络能耗模型

由于H-CRAN和传统移动网络的架构不一样，传统网络的能耗模型不一定适用于H-CRAN。因此，本文在H-CRAN中建立了完善的网络能耗模型来描述RRHs, HPN, BBU池和前传链路的能耗

$$P_{\text{tot}}(t) = \sum_{m\in M}(P_m^{\text{RRH}}(t) + P_m^{\text{FH}}(t)) + P^{\text{HPN}}(t) + P^{\text{BBU}}(t) \tag{7}$$

其中，$P_m^{\text{RRH}}(t)$, $P_m^{\text{FH}}(t)$, $P^{\text{HPN}}(t)$和$P^{\text{BBU}}(t)$分别表示第t时隙RRH m的能耗、RRH m的前传能耗，HPN的能耗以及BBU池的能耗。基站(Base Station, BS)的能量消耗为

$$P_{\text{BS}}(t) = \sum_m [P_m^{\text{RRH},S}(t) + P^{\text{HPN},S}(t) + \psi_m \sum_{k\in K}\sum_{l\in L} C_{m,k,l}(t)\phi_{m,k,l}(t)p_{m,k,l}(t)] \tag{8}$$

其中，$P_m^{\text{RRH},S}(t)$和$P^{\text{HPN},S}(t)$分别表示RRH和HPN的静态能耗，第3项表示基站的动态能耗；ψ_m的值表示基站类型，由于基带信号处理功能，HPN比RRH消耗更多的动态能耗，因此HPN具有更大的能耗因子。

在建模前传链路的能耗时，本文考虑的是基于时分复用的无源光传输网络，PON包括一个光线路终端(Optical Line Terminal, OLT)，该终端通过单个光纤连接一组相关光网络单元(Optical Network Unit, ONU)。根据文献[8]的分析，前传链路的总功耗为

$$P_{\text{fh}}(t) = P_{\text{olt}}(t) + \sum_{m=1}^{M} P_m^{\text{tl}}(t) \tag{9}$$

其中，$P_{\text{olt}}(t)$表示第t时隙OLT的功耗，$P_m^{\text{tl}}(t) = P_{a,l}^{\text{tl}}(t)$和$P_m^{\text{tl}}(t) = P_{s,m}^{\text{tl}}(t)$分别表示第$t$时隙ONU在运行模式和休眠模式下的功耗。根据文献[8]可知，它们的一般值分别为：$P_{\text{olt}}(t) = 20$ W, $P_{a,m}^{\text{tl}}(t) = 3.85$ W, $P_{s,m}^{\text{tl}}(t) = 0.75$ W，因此，将一些传输链路置于休眠模式是降低H-CRAN功耗的一种有效方法。

BBU池的功耗与基带信号处理的计算工作量密切相关[9]，本文假设对于H-CRAN中的每一个无线接入节点都存在一个相应的虚拟机(Virtual Machine, VM)。因此，BBU池的能耗为$P^{\text{BBU}}(t) = \sum_{m\in M} P_m^B(t)$，其中$P_m^{\text{BBU}}(t)$表示在第$t$时隙虚拟机VM用来处理第$m$个RRH的基带信号产生的功耗，RAP m提供服务的虚拟机的功耗表示为

$$P_m^{\text{BBU}}(t) = P_m^{\text{BBU},S}(t) + \phi_m \sum_{k\in K}\sum_{l\in L} R_{m,k,l}(t) \tag{10}$$

其中，$P_m^{\text{BBU},S}(t)$表示第t时隙下第m个RRH对应的虚拟机的静态功率，ϕ_m为能耗因子，它用来描述VM的功耗与无线资源效用之间的关系[10]，$\sum_{k\in K}\sum_{l\in L} R_{m,k,l}(t)$表示第 t 时隙所占用的无线资源。

2.5 两级队列模型

根据文献[9]的分析，本文使用两级队列模型来描述从核心网传输数据给用户。如图3所示，核心网传输给用户的业务数据首先进入基带资源池，首先分配给BBUs内的每个虚拟机。在VM的队列长度中处理后，数据将被传输到服务于用户的RRHs，再通过无线通道传输到用户。

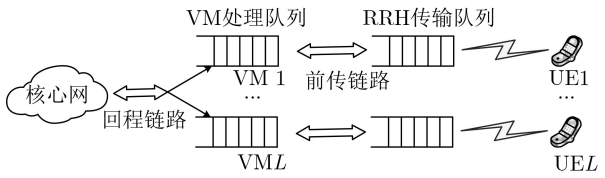

图 3 两级队列架构

用$G_l(t)$表示在BBU池中虚拟机l在第t时隙的队列长度，使用$A_l(t)$表示在第t时隙从核心网到用户l的数据到达率，其假设服从独立同分布并且满足$\lambda_l = \text{E}\{A_l(t)\}$。在处理速率为$\mu_l(t)$条件下，BBU中VM可以传输数据给服务于用户$l$的RRH，虚拟机$l$的队列模型表示为

$$G_l(t+1) = \max[G_l(t) - \mu_l(t)\tau, 0] + A_l(t) \tag{11}$$

此外，使用$Q_l(t)$表示用户l的在第t时隙的队列长度，用户l的业务队列动态更新过程为

$$Q_l(t+1) = \max[Q_l(t) - R_l(t)\tau, 0] + \mu_l(t) \tag{12}$$

根据网络稳定性[11]的定义：当离散时间队列过程$Q(t)$满足式(13)，则它是强稳定的。

$$Q(t)=\lim_{t\to\infty}\sup\frac{1}{t}\sum_{t=0}^{t-1}\mathrm{E}\{|Q(t)|\}<\infty \tag{13}$$

本文将能效η_{EE}定义为整个网络长期时间下的和速率与长期的能量消耗的比值。在业务队列稳定的前提下，基于PD-NOMA技术的H-CRAN中能效问题被建模为如下随机优化问题

$$\left.\begin{aligned}\max_{\{C,\phi,p,\omega\}}\ &\eta_{\mathrm{EE}}=\frac{\lim\limits_{t\to\infty}\frac{1}{t}\sum\limits_{t=0}^{t-1}\mathrm{E}\{R_{\mathrm{tot}}(t)\}}{\lim\limits_{t\to\infty}\frac{1}{t}\sum\limits_{t=0}^{t-1}\mathrm{E}\{P_{\mathrm{tot}}(t)\}}\\ \text{s.t.}\quad &\mathrm{C1}:\ G_l(t)<\infty,Q_l(t)<\infty\\ &\mathrm{C2}:\ p_{m,k,l}(t)<p_m^{\max},k\in K,l\in L\\ &\mathrm{C3}:\ C_{m,k,l}(t)\in\{0,1\},\phi_{m,k,l}(t)\in\{0,1\}\\ &\mathrm{C4}:\ \sum\sum R_{m,k,l}(t)\le\vartheta_m,m\in M\\ &\mathrm{C5}:\ P_m^{\mathrm{RRH}}(t)\ge0,P^{\mathrm{HPN}}(t)\ge0\end{aligned}\right\} \tag{14}$$

3 问题转化与算法描述

3.1 基于TRPO的能效优化算法

本文除了考虑约束条件外，还综合考虑网络功耗，于是资源分配问题变成了NP-hard问题，难以求出最优解。根据文献[12]的分析，深度强化学习(Deep Reinforcement Learning, DRL)可以通过与动态环境进行交互获取最优解，从而提升系统的总能效，但它只能处理低维和离散的动作空间，不能直接应用于连续域。因此，本节将引入基于连续性DRL的能效优化算法，利用RL与无线网络进行交互，并通过DL的非线性函数近似特征，使得基站做出满足优化目标的最佳决策。

为了在队列稳定，功率和前传容量受限的约束下最大化网络的能效，本文将基站关联、RB分配及功率分配描述为置信域策略优化问题，即通过引入KL散度定义的信赖域约束，通过选取合适的步长，保证策略的优化总是朝着不变坏的方向进行。首先，需定义为一个6元组$(S,A,P_{s_t,a}^{s_{t+1}},\rho_0,r,\gamma)$，其中$S$和$A$分别是状态和动作的连续且有界空间；$P_{s_t,a}^{s_{t+1}}$描述在执行动作$a\in A$时，从状态$s_t\in S$到$s_{t+1}\in S$的概率密度函数；$\rho_0$包含了可能的初始状态$s_0$的独立分布；$r:S\times S\times A\to\Re$表示奖励函数；$\gamma$为折扣因子。

通过上述的定义，本文将$\mathrm{CSI}_1(t),\mathrm{CSI}_2(t),\cdots,\mathrm{CSI}_l(t)$，$G_1(t),G_2(t),\cdots,G_l(t)$，$Q_1(t),Q_2(t),\cdots,Q_l(t)$以及$p_1(t),p_2(t),\cdots,p_M(t),p_{HPN}(t)$组合成向量表达式，作为系统的连续性状态空间$\boldsymbol{s}(t)$。此外，将$C_{m,k,1}(t),C_{m,k,2}(t),\cdots,C_{m,k,L}(t)$，$\phi_{m,1,l}(t),\phi_{m,2,l}(t),\cdots,\phi_{m,K,l}(t)$和$p_{m,k,1}(t),p_{m,k,2}(t),\cdots,p_{m,k,L}(t)$组合成向量，作为系统的连续性动作空间$\boldsymbol{a}(t)$。为了更好地解决连续性的控制难题，所提的TRPO算法兼具了值函数学习和策略学习的优势，能够有效克服离散策略空间带来的性能损失，获得更好的系统性能。

基站与无线网络进行交互时，在用户与基站进行关联的条件下，在第t时隙通过观察系统的状态空间s_t，基站采取相应的动作a_t后将得到即时奖励r_t。为了最大化H-CRAN的能效，在第t时隙的奖励函数定义为

$$r_t=\eta(t)=\frac{R_{\mathrm{tot}}(t)}{P_{\mathrm{tot}}(t)} \tag{15}$$

因此，TRPO算法的目标就是在基站与无线网络进行交互的条件下，找到一个最佳的策略$\pi:S\times A\to[0,1]$来表示基站分配无线资源行为的概率密度。最优策略π与累积折扣回报奖励具有相关性

$$\begin{aligned}\eta(\pi):=&E_{s_0,a_0,\cdots}\left\{\sum_{t=0}^{\infty}\gamma^t r_{t+1}\right\},s_0\sim P_0(s_0),\\ &a_t\sim\pi(a_t|s_t),s_{t+1}\sim P_{s_t,a_t}^{s_{t+1}}\end{aligned} \tag{16}$$

其中，$r_{t+1}:=r(s_{t+1},s_t,a_t)$，在Actor-Critic中策略$\pi$是在Actor中决定的，优势函数定义为

$$A^{\pi}(s_t,a_t):=Q^{\pi}(s_t,a_t)-V^{\pi}(s_t) \tag{17}$$

策略梯度算法的缺陷在于更新步长难以确定，当步长不合适时，更新的参数所对应的资源分配策略是一个更不好的策略。因此，合适的步长对于整个H-CRAN系统是非常关键。本文的TRPO算法通过寻找使得回报奖励函数单调递增的步长，进而逐步完善网络的资源分配策略，将新策略所对应的回报函数分解成旧的策略所对应的回报函数加上优势函数项，如式(18)所示

$$\eta(\pi^*)=\eta(\pi)+E_{s_0,a_0,\cdots,\pi^*}\left[\sum_{t=0}^{\infty}\gamma^t A_{\pi}(s_t,a_t)\right] \tag{18}$$

其中，π表示旧策略，π^*表示新策略，式(18)右侧的第2项表示新旧资源优化策略的回报函数差值，利用优势函数以及状态-动作值函数和值函数的定义，式(18)可以转化为基于无线网络状态s分布的函数

$$\begin{aligned}\eta(\pi^*)=&\eta(\pi)+\sum_{t=0}^{\infty}\sum_{s}P(s_t=s|\pi^*)\\ &\cdot\sum_{s}\pi^*(a|s)\gamma^t A_{\pi}(s,a)\end{aligned} \tag{19}$$

其中，$P(s_t=s|\pi^*)\pi^*(a|s)$为(s,a)的联合概率；$\pi^*(a|s)\gamma^t A_{\pi}(s,a)$表示对资源分配行为的边际分布，

即在状态s对整个动作空间求和。状态s的分布是由$L_{\pi_{\theta_{\text{old}}}}(\pi_{\theta_{\text{old}}})=\eta(\pi_{\theta_{\text{old}}})$且$\nabla_\theta L_{\pi_{\theta_{\text{old}}}}(\pi_\theta)|_{\theta=\theta_{\text{old}}}=\nabla_\theta\eta(\pi_{\theta_{\text{old}}})|_{\theta=\theta_{\text{old}}}$新策略$\pi^*$产生的，但是只知道旧策略$\pi$的参数，可以通过用旧的状态分布来代替新的状态分布

$$L(\pi^*)=\eta(\pi)+\sum_s\rho_\pi(s)\sum_s\pi^*(a|s)A_\pi(s,a) \tag{20}$$

式(20)中资源分配动作a是由新策略π^*产生的，利用重要性采样对资源分配动作空间的分布进行处理。为了证明策略的单调性，引入平均KL散度：$\overline{D}_{\text{KL}}^{\rho_{\theta_{\text{old}}}}(\theta_{\text{old}},\theta)\le\phi$，可以将问题转化为

$$\left.\begin{aligned}&\max_\theta E_{s\sim\pi_{\theta_{\text{old}}},a\sim\pi_{\theta_{\text{old}}}}\left[\frac{\pi_\theta(a|s_t)}{\pi_{\theta_{\text{old}}}(a|s_t)}A_{\theta_{\text{old}}}(s_n,a)\right]\\&\text{C1}: E_{s\sim\pi_{\theta_{\text{old}}}}[D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s)||\pi_\theta(\cdot|s))]\le\phi\end{aligned}\right\} \tag{21}$$

3.2 近端策略优化算法

对于上述的目标函数，为了求解出最佳的资源分配策略。首先，需要将目标函数进行1阶近似；其次，利用泰勒级数对约束条件进行2次展开；最后利用共轭梯度的方法求解更新的参数。当选用深度神经网络表示策略参数时，TRPO的标准解法产生的计算量较为庞大[13]，较难应用到H-CRAN网络的策略更新中。根据2017年OpenAI提出的PPO算法，其策略参数通过梯度估算进行迭代优化。PPO不需要估算状态转移函数，可以应用于大规模的连续域控制问题，因此，本文将使用PPO算法对式(21)的目标函数进行优化，获得最优的基站关联策略、RB分配策略以及用户功率分配策略。PPO是基于参考策略$\pi_{\theta_{0\text{ld}}}$的$\eta$1阶近似值来对$\pi_\theta$的参数进行局部优化，两者间的概率比表示为

$$\xi_t(\theta)=\frac{\pi_\theta(a|s_t)}{\pi_{\theta_{\text{old}}}(a|s_t)} \tag{22}$$

若$\xi_t(\theta)>1$，则新策略的行为比旧策略的更有可能发生。因此，TRPO的目标函数将转化为

$$L_{\text{CPI}}(\theta)=\widehat{E_t}[\xi_t(\theta)\widehat{A_t}] \tag{23}$$

其中，$\widehat{A}=\text{E}[A_{\theta_{\text{old}}}(s_t,a)]$和$\widehat{E}$分别表示优势函数和期望的估计值。在没有约束的条件下，上述问题中策略参数将会产生过大的更新，PPO算法通过修改式(21)的优化目标来惩罚策略过度更新的问题

$$\begin{aligned}L_{\text{CLIP}}(\theta)=&\widehat{E_t}[\min(\xi_t(\theta)\widehat{A_t},\text{clip}(\xi_t(\theta),\\&1-\sigma,1+\sigma)\widehat{A_t})],\quad\sigma=0.2\end{aligned} \tag{24}$$

其中，$\text{clip}(\xi_t(\theta),1-\sigma,1+\sigma)$防止策略陷入局部最优值。

为了提高网络的能效性能，Actor和Critic中神经网络的权重值都需要通过反复的学习来拟合复杂的环境特征，具体的训练模型如图4所示。该过程通过最小化Critic神经网络的损失函数来训练神经网络参数：

$$L_{\text{closs}}(\kappa_v)=r_t+V^\pi(s_{t+1};\kappa_v)-V^\pi(s_t;\kappa_v) \tag{25}$$

为了使Actor神经网络获得最佳的策略，通过最大化Actor的神经网络损失函数来训练其权重参数

$$\begin{aligned}L_{a\text{loss}}(\theta)=L_{\text{CLIP}}(\theta)=&\widehat{E_t}[\min(\xi_t(\theta)\widehat{A_t},\\&\text{clip}(\xi_t(\theta),1-\sigma,1+\sigma)\widehat{A_t})]\end{aligned} \tag{26}$$

其中，σ为超参数，取值为0.2，即：缩减该变化率在0.8～1.2之间，以保证策略更新不会过大。在

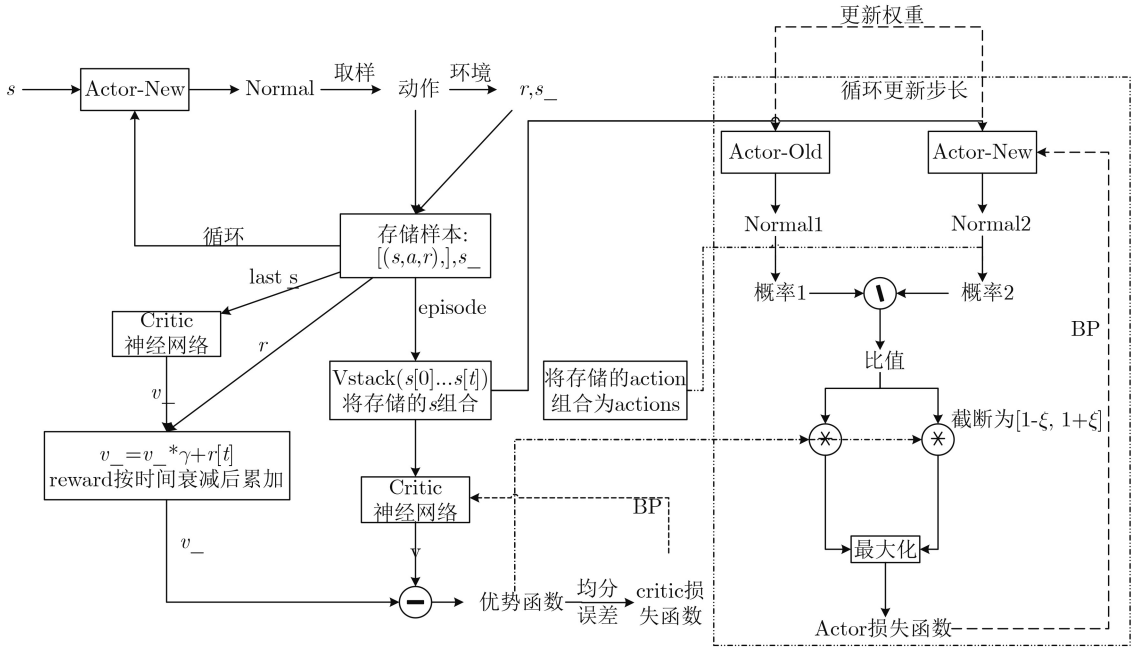

图 4 PPO算法框图

Actor-old网络中，其权重参数通过Actor-new网络定期地进行赋值更新，具体的学习流程如表1所示。

表 1 近端策略优化PPO训练Actor网络参数算法

算法1 近端策略优化(PPO)训练Actor网络参数算法
(1) 初始化Actor神经网络参数θ以及Critic的神经网络参数κ_v
(2) For episode $G=1,2,\cdots,1000$ do
(3) while经验池D中没有足够的元组do
(4) 随机选取一个初始状态s_0
(5) for step=1, 2, ···, n do
(6) 定义起始状态s，根据策略$\pi(s\|\theta)$选取动作a
(7) 采取动作a与无线网络环境进行交互后，观察下一状态s'并计算出奖励回报r.
(8) 通过式(15)计算出累计折扣奖励R_i，将元组(s_i,a_i,s_i',R_i)存入经验池D中
(9) end for
(10) end while
(11) $\theta^{\text{old}}\leftarrow\theta$
(12) for 每次更新回合 do
(13) 从经验池D中随机采样mini-batch样本
(14) 对于Critic网络而言：通过最小Critic网络中的损失函数来更新Critic的参数κ_v
(15) 对于Actor网络而言：
(16) 根据状态s_i，利用式(17)计算优势函数A_i，通过最大化actor网络的损失函数来更新Actor的参数θ
(17) end for
(18) End For

通过算法1将PPO模型训练好后，可以获取Actor神经网络的最优权重参数。利用上述参数，基站可以获得最优的策略来进行用户关联、RB分配以及功率分配，并且取得最大的能效性能。

4 仿真与讨论

在这一节中，通过与深度Q学习算法[14]和TRPO算法[13]的对比研究，详细地分析所提算法的性能。

4.1 参数设置

本文设置的网络拓扑大小为800×800 m^2, 1个HPN放置在网络中心位置，10个RRH均匀分布在网络中，HUE用户数为4, RUE用户数为35，且均匀地分布在HPN和RRH上。在仿真中，系统的时隙长度τ为10 ms，总带宽为10 MHz，子载波数目设置为32，无线信道被建模为瑞利信道，噪声功率密度为−174 dBm/Hz, HPN的路径损耗模型为$31.5+40.0\lg(d)$、RRH的路径损失模型为$31.5+35.0\lg(d)$。HPN的最大发射功率$P_{\text{HPN}}^{\max}$为43 dBm, RRH的最大发射功率$P_{\text{RRH}}^{\max}$为29 dBm, RRH和HPN的静态功率消耗分别为3.5 W和84 W。由于本文采用基于连续性的深度强化学习的算法来解决H-CRAN资源分配问题，还需要对神经网络中的参数进行训练，经验回放池的大小设置为5000, batch的大小为32。

4.2 性能分析

本节通过PPO算法的训练讨论了batch大小和损失函数对无线网络性能的影响。如图5所示，不同batch大小会使得系统的能效性能表现出巨大的差异，在batch较小的情况下，网络有可能会陷入局部最优解，并且算法的收敛速度较为缓慢。因此，合适的batch大小是DL的训练非常重要，本文将batch大小选为32。

图6展示了不同到达率对用户的平均队列长度的影响，随着仿真时隙的增加，平均队列长度起始迅速增加，随后趋于稳定。这也说明了所提的PPO算法可以有效地保证系统队列稳定性。以外，在不同到达率的条件下，平均队列长度会有所不同，随着到达率的增加，平均队列长度会越来越大。

如图7展示了不同算法下用户数对网络能效的影响，随着用户的增加，网络的吞吐量将占主导地位，网络能效越来越好。此外，由于PPO算法既解决了DQN算法无法应用于连续性以及高维动作空间的问题，又大大降低了TRPO算法的计算复杂度，因此，PPO算法对无线网络产生能效优势远远好于TRPO和DQN算法。如图8所示，PPO算法

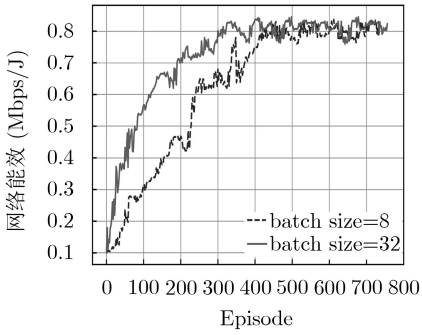

图 5 PPO算法下不同batch的网络能效

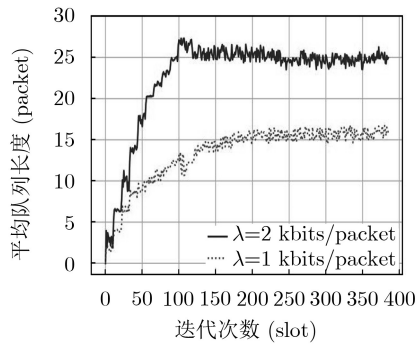

图 6 不同到达率的平均队列长度

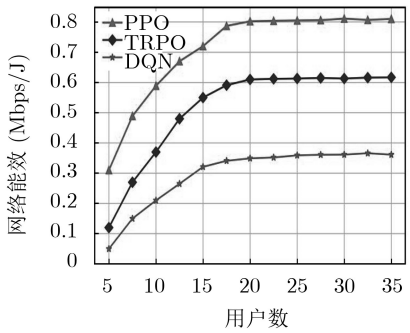

图 7 不同算法下的网络能效

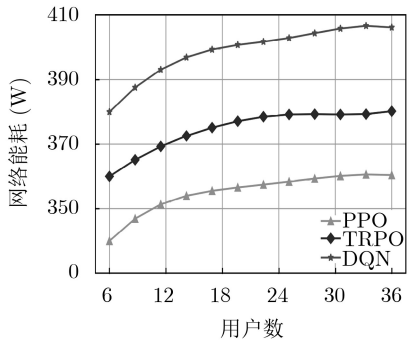

图 8 不同算法下的网络能耗

较TRPO算法而言，计算复杂度更低，从而可以更加快速、合理地获得最优的资源分配策略，避免不必要的能耗浪费。PPO算法较DQN算法而言，完美地解决了DQN在连续型环境下需要离散化的问题，使得神经网络的训练可以获得更加完善的状态信息，进而更合理地分配无线资源。

5　结论

本文在H-CRAN下行传输场景下，以队列稳定和前传链路为约束，联合优化用户关联、RB分配和功率分配，构建用户公平和网络能效的随机优化问题。将随机优化问题转化为置信域策略优化问题，通过自学习的方法求解最佳策略。此外，针对TRPO算法的标准解法产生的计算量较为庞大，采用PPO算法进行优化求解。仿真结果表明，本文所提算法在保证队列稳定约束下，进一步提高了网络的能效性能。

参 考 文 献

[1] 张广驰, 曾志超, 崔苗, 等. 无线供电混合多址接入网络的资源分配[J]. 电子与信息学报, 2018, 40(12): 3013–3019. doi: 10.11999/JEIT180219.
ZHANG Guangchi, ZENG Zhichao, CUI Miao, *et al*. Resource allocation for wireless powered hybrid multiple access networks[J]. *Journal of Electronics & Information Technology*, 2018, 40(12): 3013–3019. doi: 10.11999/JEIT180219.

[2] MOKDAD A, AZMI P, MOKARI N, *et al*. Cross-layer energy efficient resource allocation in PD-NOMA based H-CRANs: Implementation via GPU[J]. *IEEE Transactions on Mobile Computing*, 2019, 18(6): 1246–1259. doi: 10.1109/TMC.2018.2860985.

[3] ZHANG Yizhong, WU Gang, DENG Lijun, *et al*. Arrival rate-based average energy-efficient resource allocation for 5G heterogeneous cloud RAN[J]. *IEEE Access*, 2019, 7: 136332–136342. doi: 10.1109/ACCESS.2019.2939348.

[4] 陈前斌, 管令进, 李子煜, 等. 基于深度强化学习的异构云无线接入网自适应无线资源分配算法[J]. 电子与信息学报, 2020, 42(6): 1468–1477. doi: 10.11999/JEIT190511.
CHEN Qianbin, GUAN Lingjin, LI Ziyu, *et al*. Deep reinforcement learning-based adaptive wireless resource allocation algorithm for heterogeneous cloud wireless access network[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1468–1477. doi: 10.11999/JEIT190511.

[5] PENG Mugen, LI Yong, ZHAO Zhongyuan, *et al*. System architecture and key technologies for 5G heterogeneous cloud radio access networks[J]. *IEEE Network*, 2015, 29(2): 6–14. doi: 10.1109/MNET.2015.7064897.

[6] HUNG S, HSU H, CHENG S, *et al*. Delay guaranteed network association for mobile machines in heterogeneous cloud radio access network[J]. *IEEE Transactions on Mobile Computing*, 2018, 17(12): 2744–2760. doi: 10.1109/TMC.2018.2815702.

[7] TAN Zhongwei, YANG Chuanchuan, and WANG Ziyu. Energy evaluation for cloud RAN employing TDM-PON as front-haul based on a new network traffic modeling[J]. *Journal of Lightwave Technology*, 2017, 35(13): 2669–2677. doi: 10.1109/JLT.2016.2613095.

[8] DHAINI A R, HO P H, SHEN Gangxiang, *et al*. Energy efficiency in TDMA-based next-generation passive optical access networks[J]. *IEEE/ACM Transactions on Networking*, 2014, 22(3): 850–863. doi: 10.1109/TNET.2013.2259596.

[9] WANG Kaiwei, ZHOU Wuyang, and MAO Shiwen. Energy efficient joint resource scheduling for delay-aware traffic in cloud-RAN[C]. 2016 IEEE Global Communications Conference, Washington, USA, 2016: 1–6. doi: 10.1109/GLOCOM.2016.7841793.

[10] SABELLA D, DE DOMENICO A, KATRANARAS E, *et al*. Energy efficiency benefits of RAN-as-a-service concept for a cloud-based 5G mobile network infrastructure[J]. *IEEE Access*, 2014, 2: 1586–1597. doi: 10.1109/ACCESS.2014.2381215.

[11] NEELY M J. Stochastic Network Optimization with Application to Communication and Queueing Systems[M].

Morgan & Claypool, 2010: 1–211. doi: 10.2200/S00271ED1V01Y201006CNT007.

[12] NGUYEN K K, DUONG T Q, VIEN N A, *et al*. Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach[J]. *IEEE Access*, 2019, 7: 100480–100490. doi: 10.1109/ACCESS.2019.2930115.

[13] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, *et al*. Deep reinforcement learning: A brief survey[J]. *IEEE Signal Processing Magazine*, 2017, 34(6): 26–38. doi: 10.1109/MSP.2017.2743240.

[14] NASIR Y S and GUO Dongning. Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks[J]. *IEEE Journal on Selected Areas in Communications*, 2019, 37(10): 2239–2250. doi: 10.1109/JSAC.2019.2933973.

唐 伦：男，1973年生，教授、博士生导师，研究方向为新一代无线通信网络、异构蜂窝网络、软件定义无线网络等.

李子煜：女，1995年生，硕士生，研究方向为资源分配、机器学习.

管令进：男，1995年生，硕士生，研究方向为网络功能虚拟化、无线资源分配、机器学习.

陈前斌：男，1967年生，教授、博士生导师，研究方向为个人通信、多媒体信息处理与传输、下一代移动通信网络等.

责任编辑：余 蓉