# 基于多智能体强化学习的异构网络 CRE 偏置动态优化算法

张铖[1,2]，朱家烨[1]，刘泽宁[2]，黄永明[1,2]
（1. 东南大学移动通信全国重点实验室，江苏 南京 211111；
2. 网络通信与安全紫金山实验室，江苏 南京 211111）

**摘 要：**为应对无线网络用户激增导致的高吞吐量需求，针对宏微异构网络干扰场景，提出一种基于多智能体强化学习的小区范围扩展（CRE）偏置动态优化算法。基于协作多智能体强化学习的值分解网络框架，通过合理利用并在微微基站间交互系统内用户分布及其所受干扰水平，实现所有微微基站的个性化 CRE 偏置值在线本地化决策。仿真结果表明，与 CRE=5 dB、分布式 Q-Learning 算法相比，所提算法在提高系统吞吐量、均衡各基站吞吐量及改善边缘用户吞吐量方面具有明显优势。

**关键词：**异构网络；小区范围扩展；多智能体强化学习；值分解网络算法
**中图分类号：**TN92
**文献标志码：**A
**DOI:** 10.11959/j.issn.1000−436x.2023235

# Multi-agent reinforcement learning based dynamic optimization algorithm of CRE offset for heterogeneous networks

ZHANG Cheng[1,2], ZHU Jiaye[1], LIU Zening[2], HUANG Yongming[1,2]
1. National Mobile Communication Research Laboratory, Southeast University, Nanjing 211111, China
2. Purple Mountain Laboratories: Networking, Communications and Security, Nanjing 211111, China

**Abstract:** To cope with the high throughput demand caused by the proliferation of wireless network users, a multi-agent reinforcement learning based dynamic optimization algorithm of cell range expansion (CRE) offset was proposed for interference scenarios in macro-pico heterogeneous networks. Based on the value decomposition network framework of collaborative multi-agent reinforcement learning, a personalized online local decision of CRE offset for all pico-base stations was achieved by reasonably utilizing and interacting the intra-system user distribution and their interference levels among pico-base stations. Simulation results show that the proposed algorithm has significant advantages in increasing system throughput, balancing the throughput of each base station and improving edge-user throughput, compared to CRE=5 dB and distributed Q-learning algorithms.

**Keywords:** heterogeneous network, cell range expansion, multi-agent reinforcement learning, value decomposition network algorithm

## 0 引言

2022 年，全球移动数据流量每月达到 90 EB。据估计，到 2028 年，全球移动数据流量每月将增长 260%，达到 324 EB，每部智能手机的数据流量将增加 3 倍以上[1]。为了有效应对移动用户的大量

**收稿日期：**2023−08−15；**修回日期：**2023−11−13
**通信作者：**黄永明，huangym@seu.edu.cn
**基金项目：**国家自然科学基金资助项目（No.62225107, No.62271140）；江苏省前沿引领技术基础研究重大基金资助项目（No.BK20222001）；江苏省创新创业人才计划基金资助项目（No.JSSCBS20211332）
**Foundation Items:** The National Natural Science Foundation of China (No.62225107, No.62271140), The Natural Science Foundation on Frontier Leading Technology Basic Research Project of Jiangsu (No.BK20222001), The Jiangsu Innovative and Entrepreneurial Talent Program (No.JSSCBS20211332)

增长及其对高通信服务质量的巨大需求，异构网络、无蜂窝网络等新型网络架构应运而生[2-3]。3GPP 在先进长期演进技术（LTE-A, long-term evolution-advanced）的第一个版本 Release 10 中引入了由宏基站（MBS, macro base station）、微微基站（PBS, pico base station）、毫微微基站（FBS, femto base station）等组成的异构网络，即在 MBS 覆盖范围内部署 PBS、FBS 等低功率节点（LPN, low power node），构成多层次、多接入的混合网络形态[4]。

相比于传统同构网络，异构网络具有低成本、低功耗、广覆盖、高容量等优势[5-7]，但也面临着诸多亟待解决的问题与挑战。

1) 网络负载不均衡。由于 MBS 与 LPN 的发射功率差异巨大，在传统基站接入方式下，用户首选参考信号接收功率（RSRP, reference signal receiving power）最强的基站，从而导致大量用户接入 MBS，而仅有少部分用户接入 LPN。负载的不均匀分布，使 LPN 的可用资源无法得到充分利用，而 MBS 覆盖范围内的资源竞争仍非常激烈，未能充分发挥异构网络的分流优势。

2) 小区间干扰严重。异构网络在 MBS 覆盖范围内部署 PBS 等 LPN 后，网络中不仅存在相同或相似发射功率接入点之间形成的小区同层干扰，如 MBS 之间或 PBS 之间的干扰，还引入了不同发射功率接入点之间形成的小区跨层干扰，如 MBS 对 PBS 的干扰。与同层干扰相比，由于 MBS 和 LPN 的发射功率相差较大，跨层干扰对不同服务小区覆盖范围重合区域内的用户影响更严重。

为了有效解决异构网络中负载不均衡及小区间干扰严重的问题，3GPP Release 10 版本提出了增强型小区间干扰协调（eICIC, enhanced inter-cell interference coordination）技术，主要分为频率域、功率域及时域 3 个类别[8-10]。时域 eICIC 中包括小区扩展范围（CRE, cell range expansion）技术和几乎空白子帧（ABS, almost blank subframe）技术。

CRE 技术是一种基于正向偏置值的用户基站选择策略，其核心原理是在最大参考信号接收功率（Max-RSRP, maximum reference signal receiving power）接入方式的基础上，为 PBS 等 LPN 添加一个大于零的偏移量，称为 CRE 偏置值。正向 CRE 偏置值的存在，允许更多用户接入 LPN。然而，此部分用户将遭受来自 MBS 的强烈干扰。为减轻 CRE 技术下边缘用户所受干扰，现有大多数工作考虑联合使用 ABS 技术[11]。ABS 技术要求 MBS 在一个或多个子帧内不发送业务数据，期间，LPN 为小区边缘用户提供服务，从而避免来自 MBS 的主要干扰。

对于 PBS 等 LPN 而言，如果 CRE 偏置值设置过小，其覆盖范围无法充分扩大，导致系统内负载均衡效果不佳；反之，如果 CRE 偏置值设置过大，其覆盖范围将会过度扩展，使本应更适合关联 MBS 的用户被迫关联 LPN。这部分用户将遭受源自 MBS 的强烈干扰，造成通信性能损失。因此，最优 CRE 偏置值的设置成为相关研究的重点关注问题。文献[12]考虑到系统中 PBS 的位置，提出一种以最大和速率为目标的 CRE 偏置值优化算法，但仅设置一个可调整 CRE 偏置值的 PBS，忽略了多个 PBS 下 CRE 偏置值精准个性化设置的情况。文献[13]提出一种基于 PBS 用户信干噪比（SINR, signal to interference and noise ratio）的自适应 CRE 偏置值调整算法，以提高系统吞吐量；但使用该算法的 PBS 只关心自身用户干扰情况，忽略了与周围其他基站的相互影响，且只设定高、低 2 个 CRE 偏置值，动态调整 CRE 偏置值的精细度不高。

近年来，随着人工智能技术的蓬勃发展，机器学习与智能通信紧密结合[14-16]。其中，强化学习也多用于 CRE 偏置值优化，使为每个 PBS 精细设置 CRE 偏置值成为可能。文献[17]提出一种基于协作多智能体在线强化学习的 CRE 偏置值控制方案，以最大化满足时延服务质量（QoS, quality of service）要求的用户数量。文献[18]采用集中式深度学习框架，选择系统吞吐量作为奖励函数，由一个中央控制器作为智能体，为系统内所有基站设定不同的 CRE 偏置值。但在该集中式框架下，基站数量的增加会导致偏置值空间尺寸呈指数级增长，优化难度和复杂性也显著增加。

Kudo 等[19]旨在最小化用户中断次数，通过使用分布式 Q-Learning 算法为每个用户确定不同的 CRE 偏置值。文献[20-21]也采用分布式学习框架，分别基于 Q-Learning、状态−动作−奖励−状态−动作（SARSA）算法和神经网络，将全局最优拆解至单个最优，由单个基站独立自主使用智能算法寻找自身最优 CRE 偏置值。该类分布式方法的优点在于降低了优化算法的复杂性，实现简单；缺点在于上

述完全分布式框架中各智能体相互独立，即各智能体将其他智能体视作局部可观测环境的一部分，存在非平稳的问题，算法容易陷入局部最优[22]。非平稳是指每个智能体所面临的环境状态受其他智能体决策影响而动态变化，此时环境状态转移不再具备连续性，状态转移概率等不再稳定。

综上所述，利用强化学习算法感知系统内环境变化，为每个 PBS 动态精细设置 CRE 偏置值具有可实现性，但已有方案存在动作空间膨胀、优化难度高或者全局信息交互不足、难以找到全局最优解等问题。为解决上述问题，借鉴具有集中式训练、分布式执行特点的值分解网络（VDN, value decomposition network）[23]。本文针对 Macro-Pico 双层异构通信网络，考虑联合使用 CRE 技术和 ABS 技术，提出一种基于协作多智能体强化学习框架 VDN 的 CRE 偏置值在线分布式动态优化算法 VDN-CRE，以少量基站间通信开销为代价取得系统吞吐量的进一步提升。本文主要研究工作如下。

1) 建立了 Macro-Pico 双层异构网络下的干扰模型，考虑联合使用 CRE 技术和 ABS 技术，提出动态优化 CRE 偏置值以最大化系统吞吐量的多变量非凸优化问题。

2) 综合考虑用户分布及小区间干扰测量结果，基于协作多智能体强化学习 VDN 框架，设计一种集中式训练、分布式执行的 CRE 偏置值动态优化算法 VDN-CRE。将系统内 PBS 划分为多个智能体，每个智能体视作一个近似的“深度 Q 网络（DQN, deep Q-network）”结构，在考虑基站间协作关系的基础上，单独调整各 PBS 的 CRE 偏置值，实现所有 PBS 的 CRE 偏置值在线本地化决策。

3) 仿真结果表明，与已有 CRE=5 dB、分布式 Q-Learning 算法相比，本文提出的 VDN-CRE 算法能够有效提高系统吞吐量、均衡各基站吞吐量及改善边缘用户吞吐量。

## 1 系统模型

本文选择 Macro-Pico 异构网络干扰场景（如图 1 所示）作为主要研究对象，选择 PBS 作为异构网络中的 LPN，考虑由 MBS 和 PBS 构成的双层异构网络下行传输链路。假设系统中有一个 MBS，其覆盖范围内随机部署 $N_{\mathrm{P}}$ 个 PBS 和 $N_{\mathrm{U}}$ 个用户，其中，每个 PBS 的覆盖范围内包含 $N_{\mathrm{UP}}$ 个用户。

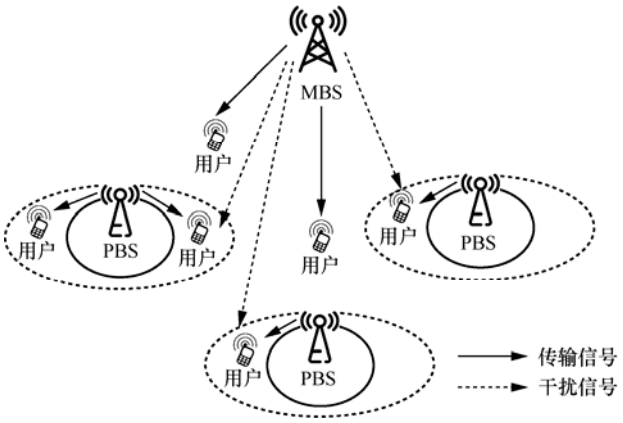


图 1 异构网络干扰场景

### 1.1 小区范围扩展技术

在系统模型中，用户使用 Max-RSRP 策略来选择关联基站，即每个用户与观测到的能提供最高 RSRP 的基站相关联。CRE 技术的存在，为每个 PBS 添加一个正向偏置值。此时，用户的关联基站为

$$b_u^* = \arg\max_{0\leqslant i\leqslant N_{\mathrm{P}}} \left\{ \mathrm{RSRP}_{iu} + \sigma_i \right\} \tag{1}$$

其中，$i$ 和 $u$ 分别表示第 $i$ 个基站和第 $u$ 个用户，$i=0$ 指代 MBS，$1\leqslant i\leqslant N_{\mathrm{P}}$ 指代 PBS；$\sigma_i$ 表示第 $i$ 个基站的 CRE 偏置值，由于 MBS 不设置 CRE 偏置值，因此 $\sigma_0=0$；$\mathrm{RSRP}_{iu}$ 代表第 $u$ 个用户接收到的来自第 $i$ 个基站的 RSRP，计算式为

$$\mathrm{RSRP}_{iu} = P_i g_{iu} \tag{2}$$

$$P_i = \begin{cases} P_{\mathrm{M}} & , i=0 \\ P_{\mathrm{P}} & , i\neq 0 \end{cases} \tag{3}$$

其中，$P_i$ 表示第 $i$ 个基站的发射功率，$P_{\mathrm{M}}$ 表示 MBS 的发射功率，$P_{\mathrm{P}}$ 表示 PBS 的发射功率；$g_{iu}$ 表示第 $u$ 个用户与第 $i$ 个基站之间的信道增益。

根据关联基站的不同，系统内用户可分为与 MBS 关联的宏小区用户（MUE, macro-cell user equipment）及与 PBS 关联的微微小区用户（PUE, pico-cell user equipment）。PUE 又可进一步划分为 2 种类型：由于正向 CRE 偏置值存在而关联 PBS 的微微小区扩展范围用户（PEUE, pico-cell expansion user equipment）和即使不添加正向 CRE 偏置值也会关联 PBS 的微微小区中心用户（PCUE, pico-cell centre user equipment）。

### 1.2 几乎空白子帧技术

本文联合使用时域 eICIC 中的 ABS 技术，以减轻 CRE 技术对边缘用户的干扰，如图 2 所示。

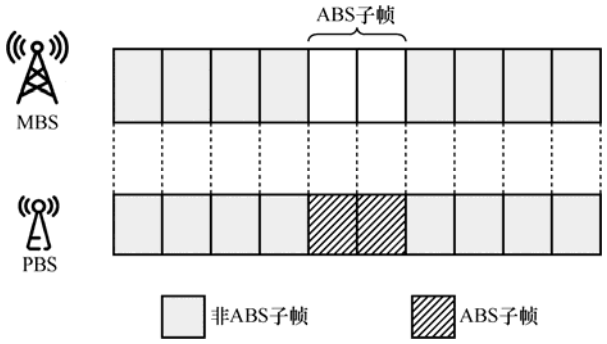


图 2　ABS 技术帧结构

ABS 技术的特点在于将子帧划分为 2 种类型：非 ABS 子帧和 ABS 子帧。ABS 子帧期间，MBS 不传输任何有效数据，PBS 调度 PCUE 和受到 MBS 严重干扰的 PEUE；非 ABS 子帧期间，MBS 正常服务 MUE，PBS 只调度 PCUE。定义第 $i$ 个基站在 ABS 子帧期间的服务用户数量为 $N_i^{\mathrm{ABS}}$，在非 ABS 子帧期间的服务用户数量为 $N_i^{\mathrm{nABS}}$。

鉴于用户在 ABS 子帧与非 ABS 子帧期间受到不同程度的小区间干扰，由 PBS 提供服务的第 $u$ 个用户处的 SINR 在 ABS 子帧与非 ABS 子帧期间可分别表示为

$$\mathrm{SINR}_{\mathrm{p},u}^{\mathrm{ABS}}=\frac{P_{b_u^*}g_{b_u^*u}}{\sum\limits_{i=1,i\neq b_u^*}^{N_\mathrm{P}}P_ig_{iu}+N_0} \tag{4}$$

$$\mathrm{SINR}_{\mathrm{p},u}^{\mathrm{nABS}}=\frac{P_{b_u^*}g_{b_u^*u}}{\underbrace{P_0g_{0u}+\sum\limits_{i=1,i\neq b_u^*}^{N_\mathrm{P}}P_ig_{iu}}_{\text{小区间干扰}}+N_0} \tag{5}$$

其中，$b_u^*$ 表示式(1)中的关联基站序号，$N_0$ 表示噪声功率。

同理，由 MBS 提供服务的第 $u$ 个用户处的 SINR 在 ABS 子帧与非 ABS 子帧期间可分别表示为

$$\mathrm{SINR}_{\mathrm{m},u}^{\mathrm{ABS}}=0 \tag{6}$$

$$\mathrm{SINR}_{\mathrm{m},u}^{\mathrm{nABS}}=\frac{P_0g_{0u}}{\sum\limits_{i=1}^{N_\mathrm{P}}P_ig_{iu}+N_0} \tag{7}$$

由于 MBS 在 ABS 子帧期间不传输任何有效数据，因此 SINR 表示为零。

定义 ABS 比例 $\beta$ 为一帧内 ABS 子帧数占总子帧数的比值，$0<\beta<1$。ABS 技术主要利用时域资源的分段与协调，来调度受到较大干扰的 PEUE，从而减轻 PEUE 受到的来自 MBS 的干扰。虽然基于 ABS 的时域干扰协调能够有效降低基站与用户特别是 MBS 与 PEUE 之间的干扰，提高系统吞吐量，但其本质是通过牺牲 MUE 的通信性能以换取系统吞吐量的提升。如果 ABS 比例设置过低，将导致 PUE 特别是其中 PEUE 的平均数据速率较低；如果 ABS 比例设置过高，将导致 MBS 性能损耗过大。

本文将系统内 $N_\mathrm{P}$ 个 PBS 覆盖范围内随机分布的 $N_\mathrm{P}\times N_\mathrm{UP}$ 个用户中高干扰用户数占比作为 ABS 比例。具体而言，将 PBS 覆盖范围内用户与最近 PBS 关联时的 SINR 与 SINR 阈值相比，小于 SINR 阈值时标记为高干扰用户，大于 SINR 阈值时标记为低干扰用户，将高干扰用户数量占总用户数量的比值向下舍入保留一位小数后作为系统的 ABS 比例，计算式为

$$\eta_{i,u}=\begin{cases}0, & \mathrm{SINR}_{i,u}\geqslant\mathrm{SINR}_{\mathrm{th}}\\ 1, & \mathrm{SINR}_{i,u}<\mathrm{SINR}_{\mathrm{th}}\end{cases} \tag{8}$$

$$N_{\mathrm{Hint},i}=\sum_{u=1}^{N_\mathrm{UP}}\eta_{i,u} \tag{9}$$

$$\beta=\mathrm{rounddown}\left(\frac{\sum\limits_{i=1}^{N_\mathrm{P}}N_{\mathrm{Hint},i}}{N_\mathrm{U}}\right) \tag{10}$$

其中，$\eta_{i,u}\left(1\leqslant i\leqslant N_\mathrm{P}\right)$ 表示第 $i$ 个 PBS 覆盖范围内第 $u$ 个用户的干扰指示，高干扰用户取值为 1，低干扰用户取值为 0；$\mathrm{SINR}_{i,u}$ 表示第 $i$ 个 PBS 覆盖范围内随机分布的第 $u$ 个用户与该 PBS 关联时的 SINR，具体计算式可参考式(5)；$\mathrm{SINR}_{\mathrm{th}}$ 表示 SINR 阈值；$N_{\mathrm{Hint},i}$ 表示第 $i$ 个 PBS 覆盖范围内的高干扰用户数；rounddown(·) 表示向下舍入计算，此处保留一位小数。

### 1.3　问题描述

假设系统内基站采用比例公平方式调度用户，每个用户获得几乎相等的资源量。此时，可根据香农公式计算第 $u$ 个用户的吞吐量为

$$C_u=\begin{cases}\beta\dfrac{W}{N_{b_u^*}^{\mathrm{ABS}}}\mathrm{lb}\left(1+\mathrm{SINR}_u^{\mathrm{ABS}}\right)\\ (1-\beta)\dfrac{W}{N_{b_u^*}^{\mathrm{nABS}}}\mathrm{lb}\left(1+\mathrm{SINR}_u^{\mathrm{nABS}}\right)\end{cases} \tag{11}$$

其中，$W$ 表示系统带宽，$N_{b_u^*}^{\mathrm{ABS}}$ 和 $N_{b_u^*}^{\mathrm{nABS}}$ 分别表示第 $u$ 个用户所关联的第 $b_u^*$ 个基站在 ABS 子帧和非

ABS 子帧期间的服务用户数量。ABS 子帧期间，当用户关联 PBS 时，$\mathrm{SINR}_u^{\mathrm{ABS}}=\mathrm{SINR}_{\mathrm{p},u}^{\mathrm{ABS}}$；当用户关联 MBS 时，$\mathrm{SINR}_u^{\mathrm{ABS}}=\mathrm{SINR}_{\mathrm{m},u}^{\mathrm{ABS}}$。同理，非 ABS 子帧期间，当用户关联 PBS 时，$\mathrm{SINR}_u^{\mathrm{nABS}}=\mathrm{SINR}_{\mathrm{p},u}^{\mathrm{nABS}}$；当用户关联 MBS 时，$\mathrm{SINR}_u^{\mathrm{nABS}}=\mathrm{SINR}_{\mathrm{m},u}^{\mathrm{nABS}}$，具体计算式可见式(4)～式(7)。

关联到第 $i$ 个基站的所有用户吞吐量之和计算式为

$$C_i=C_i^{\mathrm{ABS}}+C_i^{\mathrm{nABS}} \tag{12}$$

$$C_i^{\mathrm{ABS}}=\frac{\beta W}{N_i^{\mathrm{ABS}}}\sum_{u=1}^{N_i^{\mathrm{ABS}}}\mathrm{lb}\left(1+\mathrm{SINR}_u^{\mathrm{ABS}}\right) \tag{13}$$

$$C_i^{\mathrm{nABS}}=\frac{(1-\beta)W}{N_i^{\mathrm{nABS}}}\sum_{u=1}^{N_i^{\mathrm{nABS}}}\mathrm{lb}\left(1+\mathrm{SINR}_u^{\mathrm{nABS}}\right) \tag{14}$$

其中，$C_i^{\mathrm{ABS}}$ 表示 ABS 子帧期间第 $i$ 个基站内所有用户吞吐量之和，$C_i^{\mathrm{nABS}}$ 表示非 ABS 子帧期间第 $i$ 个基站内所有用户吞吐量之和。

系统吞吐量是系统内所有小区吞吐量之和，也是系统中所有用户吞吐量之和。本文的目标是通过联合优化系统内所有 PBS 的 CRE 偏置值，最大化系统吞吐量

$$\left\{\sigma_1^*,\sigma_1^*,\cdots,\sigma_{N_\mathrm{P}}^*\right\}=\underset{\sigma_1,\sigma_2,\cdots,\sigma_{N_\mathrm{P}}}{\arg\max}\sum_{i=0}^{N_\mathrm{P}}C_i \tag{15}$$

## 2　算法设计

针对上述提出的 Macro-Pico 异构网络干扰场景下最大化系统吞吐量问题，本节综合考虑系统内用户的分布及其所受小区间干扰情况，利用强化学习中的多智能体 VDN 算法，提出一种针对系统内单个 PBS 动态优化 CRE 偏置值的算法。CRE 偏置动态优化方案中，每个 PBS 都可通过实时信息采集，获得关联至本 PBS 及系统内其他 PBS 的 PEUE 数量、PEUE 所受小区间干扰情况，将其作为状态信息，利用多智能体 VDN 算法，不断寻找异构网络中最大化系统吞吐量的最优 CRE 偏置值设定。为降低算法运行与实际环境的交互开销并提高其响应速度，可借鉴基于数字孪生的辅助优化模型[24]。考虑到数据采集的非理想性等因素，数字孪生辅助优化模型与真实物理世界间可能存在偏差，对数字孪生辅助优化效果产生负面影响。后续工作将聚焦解决上述问题，如利用数据增广技术引入更多数据进行训练，提高模型的精度和准确性；引入实时数据采集和反馈机制，使数字孪生模型能够及时调整以适应环境的动态性；考虑清洗、校正数据，确保模型训练和优化所使用的数据集具有高质量和代表性。

以下将对该算法的描述、设计思路及步骤流程进行详细阐述。

### 2.1　值分解网络算法

若为每个 PBS 单独设置最优 CRE 偏置值，势必会造成组合空间的极速膨胀。例如，每个 PBS 的 CRE 偏置值有 10 种可能选择，则 5 个 PBS 构成的 CRE 偏置值组合高达 $10^5$ 种。此时，单智能体强化学习算法将不再适用于解决上述庞大的组合优化问题，需要引入多智能体强化学习算法。在异构网络干扰场景中，各个 PBS 之间既相互影响又属于合作关系，共同作用于系统内所有用户的通信质量。因此，本文选择协作多智能体强化学习中的 VDN 算法并进行针对性设计。

VDN 算法属于多智能体强化学习中的值函数分解法，采用集中式训练分布式执行的架构，通过增加值分解层来表示多智能体问题中的联合动作价值函数[25-26]。VDN 算法架构如图 3 所示。

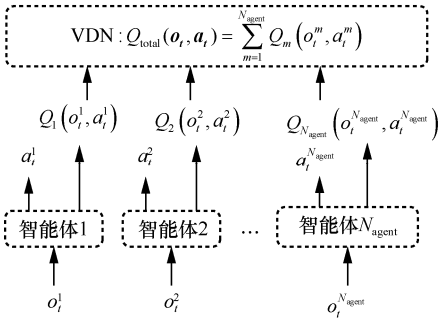


图 3　VDN 算法架构

首先，每个智能体基于各自的状态信息 $o_t^m$ 求解自己的单体动作价值函数 $Q_m\left(o_t^m,a_t^m\right)$。然后，将所有智能体的单体动作价值函数 $Q_m\left(o_t^m,a_t^m\right)$ 融合成一个面向全部智能体的联合动作价值函数 $Q_{\mathrm{total}}\left(\boldsymbol{o}_t,\boldsymbol{a}_t\right)$，其中，$\boldsymbol{o}_t=\left\{o_t^1,\cdots,o_t^{N_{\mathrm{agent}}}\right\}$ 表示所有智能体的联合状态信息，$N_{\mathrm{agent}}$ 为智能体数量，$\boldsymbol{a}_t=\left\{a_t^1,\cdots,a_t^{N_{\mathrm{agent}}}\right\}$ 表示所有智能体的联合动作。最后，基于该联合动作价值函数和全局奖励进行学习。VDN 算法将联合动作价值函数 $Q_{\mathrm{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t)$ 与每个智能体的单体动作价值函数 $Q_m(o_t^m,a_t^m)$ 之间的融

合关系视作线性加和，即利用 $Q_m\left(o_t^m,a_t^m\right)$ 的和表示 $Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t)$。此时，联合动作价值函数表达式为

$$Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t)=\sum_{m=1}^{N_{\text{agent}}}Q_m\left(o_t^m,a_t^m\right) \tag{16}$$

VDN 中单个智能体采用与 DQN 算法相似的结构，由预测神经网络和目标神经网络双重神经网络进行学习。神经网络的输入为单个智能体的状态信息 $o_t^m$，输出为单体动作价值函数，神经网络的参数从全局奖励中学习。本文采用全连接神经网络，由输入层、隐藏层和输出层三部分组成，层与层之间引入线性整流函数（ReLU, rectified linear unit）。输入层的神经元个数等于状态信息 $o_t^m$ 的个数，输出层的神经元个数等于 $a_t^m$ 的个数。$t$ 时刻，VDN 算法基于时序差分（TD, temporal difference）的可学习目标值 $y_t$ 定义为

$$y_t=r_t+r_d\max_{\boldsymbol{a}}Q_{\text{total}}^{-}\left(\boldsymbol{o}_{t+1},\boldsymbol{a};\boldsymbol{\theta}^{-}\right) \tag{17}$$

神经网络损失函数 $L(\boldsymbol{\theta})$ 定义为

$$L(\boldsymbol{\theta})=\mathbb{E}\left[\left(r_t+r_d\max_{\boldsymbol{a}}Q_{\text{total}}^{-}\left(\boldsymbol{o}_{t+1},\boldsymbol{a};\boldsymbol{\theta}^{-}\right)-Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta})\right)^2\right] \tag{18}$$

其中，$r_t$ 表示 $t$ 时刻的全局奖励，$r_d$ 表示折扣因子；$Q_{\text{total}}^{-}\left(\boldsymbol{o}_{t+1},\boldsymbol{a};\boldsymbol{\theta}^{-}\right)$ 表示目标神经网络输出的联合动作价值函数，$\boldsymbol{\theta}^{-}$ 表示与联合动作价值函数相关的所有目标神经网络的参数，$\boldsymbol{\theta}^{-}=\left\{\theta_1^{-},\cdots,\theta_{N_{\text{agent}}}^{-}\right\}$；$Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta})$ 表示预测神经网络输出的联合动作价值函数，$\boldsymbol{\theta}$ 表示与联合动作价值函数相关的所有预测神经网络的参数，$\boldsymbol{\theta}=\left\{\theta_1,\cdots,\theta_{N_{\text{agent}}}\right\}$。$Q_{\text{total}}^{-}\left(\boldsymbol{o}_{t+1},\boldsymbol{a};\boldsymbol{\theta}^{-}\right)$ 与 $Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta})$ 可分别由目标神经网络及预测神经网络输出的单体动作价值函数通过融合函数融合而成

$$Q_{\text{total}}^{-}\left(\boldsymbol{o}_{t+1},\boldsymbol{a};\boldsymbol{\theta}^{-}\right)=\sum_{m=1}^{N_{\text{agent}}}Q_m^{-}\left(o_{t+1}^m,a^m;\theta_m^{-}\right) \tag{19}$$

$$Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta})=\sum_{m=1}^{N_{\text{agent}}}Q_m\left(o_t^m,a_t^m;\theta_m\right) \tag{20}$$

其中，$\theta_m^{-}$ 表示第 $m$ 个智能体的目标神经网络参数，$\theta_m$ 表示第 $m$ 个智能体的预测神经网络参数。

使用小批量梯度下降（MBGD, mini-batch gradient descent）法更新神经网络参数

$$\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}-r_a\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta}) \tag{21}$$

$$\nabla_{\boldsymbol{\theta}}L(\boldsymbol{\theta})\approx\frac{1}{N_{\text{batch}}}\sum_{l}^{N_{\text{batch}}}(y_t)\nabla Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta}) \tag{22}$$

其中，$r_a$ 为学习率，$N_{\text{batch}}$ 为小批量样本数量。

VDN 算法采用集中式训练、分布式执行的方式。在训练阶段，使用损失函数 $L(\boldsymbol{\theta})$ 进行更新时，相当于在不断优化联合动作价值函数 $Q_{\text{total}}(\boldsymbol{o}_t,\boldsymbol{a}_t;\boldsymbol{\theta})$ 的参数 $\boldsymbol{\theta}$，而 $\boldsymbol{\theta}$ 又由各个单体动作价值函数 $Q_m\left(o_t^m,a_t^m;\theta_m\right)$ 中的参数 $\theta_m$ 组成，因此单体动作价值函数的参数 $\theta_m$ 会随着损失函数的优化而实现端到端的优化。

### 2.2　动态优化 CRE 偏置值算法

为解决异构网络中 PBS 的 CRE 偏置值优化问题，提高系统吞吐量，本文采用协作多智能体强化学习 VDN 算法，集中训练、分布执行，系统内的每个 PBS 可根据自己的状态信息完成动作决策。据此，定义以下 VDN 算法中的 4 个关键元素。

1) 智能体 Agent

对于协作多智能体强化学习 VDN 算法，每一个 PBS 都可作为一个智能体。为使算法更具有普适性，本文将系统内所有 PBS 划分为 $K$ 个不同的区域，第 $k(1\leqslant k\leqslant K)$ 个区域内的 $N_k$ 个 PBS 被视为一个智能体，共享同一个 CRE 偏置值。此时，智能体 Agent 表示为

$$\left\{\text{agent}_1,\cdots,\text{agent}_K\right\} \tag{23}$$

其中，$\text{agent}_k=\left\{\text{PBS}_{k,1},\cdots,\text{PBS}_{k,N_k}\right\},k=1,\cdots,K$。很明显，当 $K=1$、$N_1=N_{\text{P}}$ 时，系统内所有 PBS 共享同一个动态调整的 CRE 偏置值；当 $K=N_{\text{P}}$、$N_1=\cdots=N_K=1$ 时，系统内每个 PBS 单独拥有一个个性化调整的 CRE 偏置值。

2) 状态 State

当 PBS 调整 CRE 偏置值时，不仅会对其本身覆盖范围内用户产生影响，使关联到该 PBS 的用户数量和用户群体发生变化，同时也会影响周围其他基站。VDN 算法选择将关联至 PBS 的 PEUE 数量及 PEUE 所受小区间干扰情况作为状态信息。对于第 $k$ 个智能体，其状态信息 $o^k$ 由局部信息和全局信息两部分组成。

第一部分：局部信息，第 $k$ 个智能体中各个 PBS 小区扩展范围内的 PEUE 数量及 PEUE 所受干扰情况，表示为

$$\left\{N_{\mathrm{PEUE}_{k,1}},\cdots,N_{\mathrm{PEUE}_{k,N_k}},\overline{I}_{\mathrm{PEUE}_{k,1}},\cdots,\overline{I}_{\mathrm{PEUE}_{k,N_k}}\right\} \quad (24)$$

其中，$N_{\mathrm{PEUE}_{k,i}}(1\leqslant i\leqslant N_k)$ 表示由于 CRE 偏置值存在，关联到第 $k$ 个智能体中第 $i$ 个 PBS 的 PEUE 数量；$\overline{I}_{\mathrm{PEUE}_{k,i}}(1\leqslant i\leqslant N_k)$ 表示关联到第 $k$ 个智能体中第 $i$ 个 PBS 的 PEUE 所受小区间干扰强度均值，计算式为

$$\overline{I}_{\mathrm{PEUE}_{k,i}}=\frac{1}{N_{\mathrm{PEUE}_{k,i}}}\sum_{u=1}^{N_{\mathrm{PEUE}_{k,i}}}I_{\mathrm{PEUE}_{k,i,u}} \quad (25)$$

其中，$I_{\mathrm{PEUE}_{k,i,u}}$ 表示关联到第 $k$ 个智能体中第 $i$ 个 PBS 的第 $u$ 个 PEUE 所受小区间干扰，具体计算式如式(5)所示。

第二部分：全局信息，其他 $K-1$ 个智能体中所有 PBS 小区扩展范围内的 PEUE 数量及 PEUE 所受干扰情况，表示为

$$\left\{N_{\mathrm{PEUE}_1},\cdots,N_{\mathrm{PEUE}_{l(l\neq k)}},\cdots,N_{\mathrm{PEUE}_K},\right.$$
$$\left.\overline{I}_{\mathrm{PEUE}_1},\cdots,\ \overline{I}_{\mathrm{PEUE}_{l(l\neq k)}},\cdots,\overline{I}_{\mathrm{PEUE}_K}\right\} \quad (26)$$

其中，$N_{\mathrm{PEUE}_{l(l\neq k)}}$ 表示由于 CRE 偏置值存在，关联到第 $l$ 个智能体中 PBS 小区扩展范围内的所有 PEUE 数量，计算式为

$$N_{\mathrm{PEUE}_l}=\sum_{i=1}^{N_l}N_{\mathrm{PEUE}_{l,i}} \quad (27)$$

其中，$N_{\mathrm{PEUE}_{l,i}}$ 表示第 $l$ 个智能体中第 $i$ 个 PBS 的 PEUE 数量。$\overline{I}_{\mathrm{PEUE}_{l(l\neq k)}}$ 表示关联到第 $l$ 个智能体中 PBS 的所有 PEUE 所受小区间干扰强度均值，计算式为

$$\overline{I}_{\mathrm{PEUE}_l}=\frac{1}{N_{\mathrm{PEUE}_l}}\sum_{i=1}^{N_l}\sum_{u=1}^{N_{\mathrm{PEUE}_{l,i}}}I_{\mathrm{PEUE}_{l,i,u}} \quad (28)$$

其中，$I_{\mathrm{PEUE}_{l,i,u}}$ 表示关联到第 $l$ 个智能体中第 $i$ 个 PBS 的第 $u$ 个 PEUE 所受到的干扰强度。

此时，强化学习系统中所有智能体的状态信息构成联合状态信息 $\boldsymbol{o}$，表示为

$$\boldsymbol{o}=\left\{o^1,\cdots,o^K\right\} \quad (29)$$

3) 动作 Action

动作是智能体的输出，是智能体与环境交互时做出的行为动作。VDN 算法中，智能体的动作就是 PBS 的 CRE 偏置值，因此，第 $k$ 个智能体的动作 $a^k$ 为

$$a^k=\left\{\sigma^k\right\},\sigma^k\in\mathcal{A}_k \quad (30)$$

其中，$\sigma^k$ 表示 CRE 偏置值，所有 PBS 的 CRE 偏置值构成集合 $\boldsymbol{\sigma}=\left\{\sigma^1,\cdots,\sigma^{N_{\mathrm{P}}}\right\}$；$\mathcal{A}_k$ 为第 $k$ 个智能体的动作空间，表示所有可能 CRE 偏置值的集合。多智能体系统内所有智能体的动作构成联合动作 $\boldsymbol{a}$，表示为

$$\boldsymbol{a}=\left\{a^1,\cdots,a^K\right\} \quad (31)$$

4) 奖励 Reward

本文利用 VDN 算法动态优化系统内 PBS 的 CRE 偏置值，目的在于最大化系统吞吐量，因此奖励可设定为基于 VDN-CRE 算法动态优化 CRE 偏置值时系统吞吐量与不使用 CRE 技术时系统吞吐量之差的相关函数。对于第 $k$ 个智能体，奖励 $r^k$ 定义为

$$r^k=\left(\sum_{i=1}^{N_k}C_{k,i}-\sum_{i=1}^{N_k}C_{k,i,0}\right)+\tau_k(C_0-C_{0,0})-\mu \quad (32)$$

其中，$C_{k,i}$ 表示基于 VDN-CRE 算法下第 $k$ 个智能体中第 $i$ 个基站的吞吐量，$C_{k,i,0}$ 表示不使用 CRE 技术时第 $k$ 个智能体中第 $i$ 个基站的吞吐量；$\sum_{i=1}^{N_k}C_i-\sum_{i=1}^{N_k}C_{i,0}$ 表示基于 VDN-CRE 算法动态优化 CRE 偏置值与不使用 CRE 技术，第 $k$ 个智能体中所有 PBS 的吞吐量总和之差；$C_0-C_{0,0}$ 表示基于 VDN-CRE 算法动态优化 CRE 偏置值与不使用 CRE 技术时系统内 MBS 的吞吐量之差；$\mu$ 表示奖励调节因子；$\tau_k$ 表示第 $k$ 个智能体对 MBS 吞吐量变化的贡献因子，定义为第 $k$ 个智能体中 PEUE 数量与系统内总 PEUE 数量之比，计算式为

$$\tau_k=\frac{N_{\mathrm{PEUE}_k}}{\sum_{l=1}^{K}N_{\mathrm{PEUE}_l}} \quad (33)$$

所有智能体的奖励构成全局奖励 $r$ 为

$$r=\sum_{k=1}^{K}r^k \quad (34)$$

本文提出的 VDN-CRE 算法具体包括如下步骤。

1) 多智能体与环境交互

首先，单个智能体在 $t$ 时刻观测得到状态信息 $o_t^k$，包括式(24)的局部信息和式(26)的全局信息。为加快神经网络的学习速度，本文引入参数共享技巧，即所有智能体使用相同的神经网络逼近单体动作价值函数。因此，VDN-CRE 算法中各个智能体的预测神经网络的参数 $\boldsymbol{\theta}$ 相同，目标神经网络的参数 $\boldsymbol{\theta}^-$ 也相同。为了进一步区分参数共享下的多智能

体，将智能体编号也作为神经网络输入的一部分。由此，将单个智能体状态信息 $o_t^k$ 与智能体编号共同输入预测神经网络，获取动作空间内不同 CRE 偏置值的单体动作价值函数值，并使用 $\varepsilon$-贪婪策略选择动作 $a_t^k$。单个智能体执行动作 $a_t^k$，设定新的 CRE 偏置值后，获取下一时刻状态信息 $o_{t+1}^k$ 及环境反馈的奖励 $r_t^k$。$\varepsilon$-贪婪策略更新式为

$$\varepsilon = \frac{\varepsilon_0}{n_{\text{episode}} + 1} \tag{35}$$

其中，$n_{\text{episode}}$ 表示强化学习轮次数。在全部智能体完成上述过程后，根据式(34)计算全局奖励 $r_t$，并且获得联合动作 $\boldsymbol{a}_t$、联合状态信息 $\boldsymbol{o}_t$ 和下一个时刻的联合状态信息 $\boldsymbol{o}_{t+1}$。

至此，多智能体与环境完成一次交互，将上述交互信息作为样本数据，以 $(\boldsymbol{o}_t, r_t, \boldsymbol{a}_t, \boldsymbol{o}_{t+1})$ 的形式存入经验池 $\mathcal{D}$。

2) 预测网络更新

从经验池 $\mathcal{D}$ 中随机取出一小批量（$N_{\text{batch}}$ 个）经验样本数据，用于训练预测神经网络。首先根据式(17)计算目标值，然后根据式(18)构建神经网络损失函数，接着利用式(21)、式(22)中 MBGD 法更新网络参数。

3) 目标网络更新

设定目标神经网络更新频率为 $N_{\text{update}}$。经过 $N_{\text{update}}$ 轮次迭代后，将预测网络参数 $\boldsymbol{\theta}$ 赋予目标网络参数 $\boldsymbol{\theta}^-$，用于目标网络权重参数的更新。

综上，本文提出的 VDN-CRE 算法如算法 1 所示。

**算法 1**　VDN-CRE 算法

**初始化**　最大训练轮次数 $N_{\text{episode}}$，每轮最大训练步骤数 $N_{\text{step}}$，每轮最大相同步骤数 $N_{\text{samestep}}$，目标神经网络更新频率 $N_{\text{update}}$，学习率 $r_a$，折扣因子 $r_d$，贪婪策略初始值 $\varepsilon_0$，经验池 $\mathcal{D}$，预测神经网络参数 $\boldsymbol{\theta}$，目标神经网络参数 $\boldsymbol{\theta}^-$，CRE 偏置值 $\boldsymbol{\sigma}_0$

for $n_{\text{episode}} = 0$ to $N_{\text{episode}}$ do
　设置时间步 $n_{\text{step}} = 0$，相同步数 $n_{\text{samestep}} = 0$，最近吞吐量 $C_{\text{last}} = 0$；
　设置 CRE 偏置值 $\boldsymbol{\sigma} = \boldsymbol{\sigma}_0$；
　根据式(35)更新贪婪策略；
　while $n_{\text{samestep}} \leqslant N_{\text{samestep}}$ and $n_{\text{step}} \leqslant N_{\text{step}}$ do
　　for $k = 1$ to $K$ do
　　获取状态信息 $o_t^k$；
　　根据 $\varepsilon$-贪婪策略选择动作 $a_t^k$；
　　执行动作 $a_t^k$，获取下一个状态信息 $o_{t+1}^k$ 和奖励 $r_t^k$；
　　end for
　　根据式(34)计算全局奖励 $r_t$；
　　获取联合动作 $\boldsymbol{a}_t$，联合状态信息 $\boldsymbol{o}_t$ 和下一个联合状态信息 $\boldsymbol{o}_{t+1}$；
　　将经验样本数据 $(\boldsymbol{o}_t, r_t, \boldsymbol{a}_t, \boldsymbol{o}_{t+1})$ 存入经验池 $\mathcal{D}$；
　　if 经验池 $\mathcal{D}$ 中样本数量大于小批量样本数量 $N_{\text{batch}}$ then
　　从经验池 $\mathcal{D}$ 中随机抽取数量为 $N_{\text{batch}}$ 个经验样本数据；
　　根据式(17)计算目标值；
　　根据式(18)～式(22)更新预测神经网络参数 $\boldsymbol{\theta}$；
　　end if
　　更新时间步：$n_{\text{step}} = n_{\text{step}} + 1$；
　　计算系统吞吐量：$C_t = \sum_{i=0}^{N_{\text{P}}} C_i$；
　　if $C_t = C_{\text{last}}$ then
　　$n_{\text{samestep}} = n_{\text{samestep}} + 1$；
　　else
　　$n_{\text{samestep}} = 0$，$C_{\text{last}} = C_t$；
　　end if
　　每隔 $N_{\text{update}}$ 更新目标神经网络参数：$\boldsymbol{\theta}^- = \boldsymbol{\theta}$；
　end while
end for

相较于分布式 Q-Learning 算法完全分布式训练、分布式执行的架构，所提 VDN-CRE 算法利用集中式训练、分布式执行，寻找面向全智能体的联合动作价值函数与单体动作价值函数的关系，利用全局信息进行训练，有效应对了完全分布式的“非平稳”问题[27-28]。相比分布式 Q-Learning 算法，VDN-CRE 算法的复杂度有所提升，具体表现在以下几点。

1) 模型结构更复杂。一方面，VDN-CRE 算法

使用神经网络代替 Q 表，由双重神经网络进行学习。另一方面，为了实现集中式训练与分布式执行，VDN-CRE 算法中增添了将所有智能体的单体动作价值函数融合成面向全智能体的联合动作价值函数的步骤。

2) 数据量更多。VDN-CRE 算法的状态数据包括局部信息和全局信息两部分，奖励数据包括基于动态优化前后系统内 PBS 的吞吐量之差与 MBS 的吞吐量之差两部分。

3) 数据交互更频繁。由于 VDN-CRE 算法训练过程中状态信息包括局部信息和全局信息，因此作为智能体的 PBS 需要与 MBS、其他 PBS 多次交互相关信息。

下面，总结上述优化策略下发至真实物理世界的执行过程。在真实物理世界执行过程中，处于 PBS 扩展范围内的 PEUE 将所受小区间干扰强度大小报告服务 PBS，PBS 获取服务的 PEUE 数量及 PEUE 所受小区间干扰强度均值，同时与其他基站智能体进行交互，获取其他 PBS 的 PEUE 数量及 PEUE 所受小区间干扰，由此组成本地状态信息。各 PBS 根据本地状态信息，将其输入本地神经网络后即得到其动作决策 CRE 偏置值大小，并广播至覆盖范围内的用户。用户根据附近各个基站（包括 MBS 与 PBS）的参考信号接收功率强度大小及 CRE 偏置值等相关配置参数，决定关联基站序号。

## 3 仿真结果及分析

本节通过仿真实验评估基于多智能体强化学习的 CRE 偏置值动态优化算法 VDN-CRE 的有效性与优势。如图 4 所示，假设 Macro-Pico 异构网络干扰场景下 MBS 位置位于坐标原点，PBS 与用户均在其小区半径内随机分布。PBS 与 MBS、用户与 MBS、用户与 PBS 的距离均设置一定约束。

具体考虑一个 MBS、4 个 PBS 与 200 个用户，干扰场景仿真参数和强化学习网络参数分别如表 1 和表 2 所示。为了更好地体现 VDN-CRE 算法的适用性，本文针对每个不同 PBS 覆盖范围内用户数量 $N_{\mathrm{UP}}$，分别产生 100 组用户随机分布，并计算对应结果的统计平均值。将系统内所有 PBS 划分成 $K=4$ 个区域、智能体数量为 $N_{\mathrm{agent}}=4$，此时每个 PBS 单独拥有一个根据自身状态信息动态优化的 CRE 偏置值。

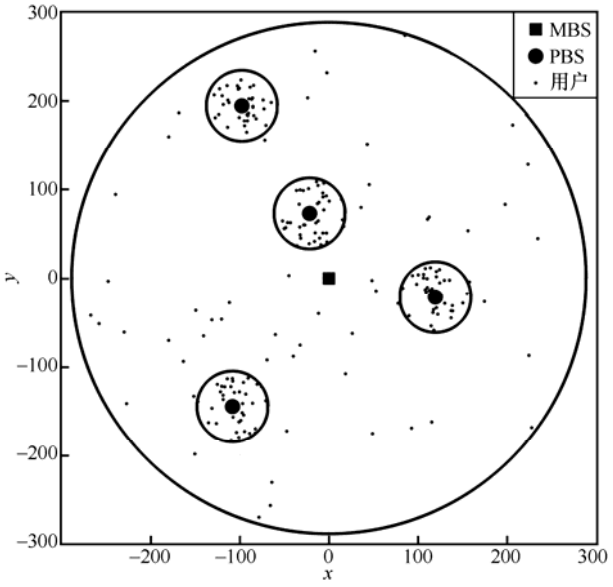


图 4 Macro-Pico 异构网络干扰场景

**表 1 干扰场景仿真参数**

| 参数 | 取值 | |
|---|---|---|
| | MBS | PBS |
| 小区半径/m | 289 | 40 |
| 发射功率/dBm | 46 | 30 |
| 距离相关路径损耗/dB | 140.7+36.7lg$R$ | 128.1+36.7lg$R$ |
| 载波频率/GHz | 2 | |
| 带宽/MHz | 20 | |
| 热噪声密度/(dBm·Hz$^{-1}$) | −174 | |
| 信道衰落 | 瑞利衰落 | |
| 偏置值范围/dB | {1,2,3,4,5,6,7,8,9} | |
| SINR 阈值/dB | 15 | |
| MBS 与 PBS 之间最小距离/m | 75 | |
| PBS 与 PBS 之间最小距离/m | 40 | |
| MBS 与用户之间最小距离/m | 35 | |
| PBS 与用户之间最小距离/m | 10 | |

**表 2 强化学习网络参数**

| 参数 | 取值 |
|---|---|
| 学习率 $r_a$ | 0.000 5 |
| 折扣因子 $r_d$ | 0.88 |
| 奖励调节因子 $\mu$ | 65 |
| 贪婪策略初始值 $\varepsilon_0$ | 0.5 |
| 每轮最大相同步骤数 | 30 |
| 最大训练轮数 | 500 |
| 最大训练步骤数 | 100 |
| 经验池容量 | 10 000 |
| 小批量数据 | 32 |
| 隐藏层 1 神经元数量 | 32 |
| 隐藏层 2 神经元数量 | 32 |

首先，对 VDN-CRE 算法的收敛性能进行仿真验证分析。如图 5 所示，强化学习神经网络损失函数值随着训练步数的增加而逐渐降低。如图 6 所示，对于从 VDN-CRE 算法中学习到的策略，每轮次平均奖励随着训练轮次的增加而不断增长，约在 50 个轮次后开始收敛，并在 180 个轮次后稳定。

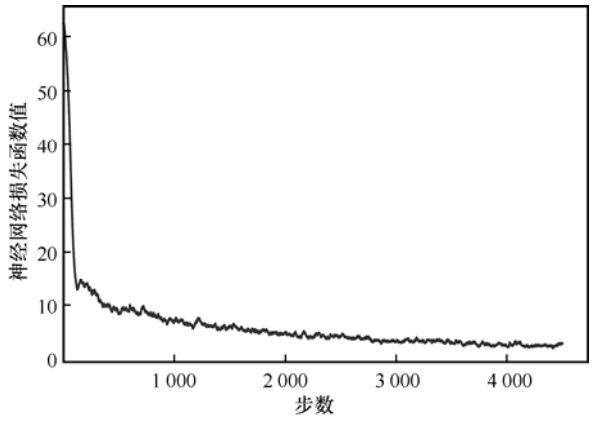


图 5　神经网络损失函数值

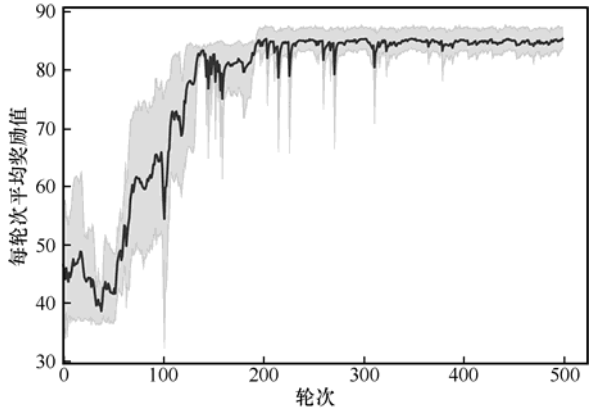


图 6　每轮次平均奖励值

其次，为了更好地比较 VDN-CRE 算法在提高系统吞吐量、改善边缘用户吞吐量方面的优势，将其与以下 3 种已有算法进行比较。

1) 不使用 CRE 技术（简称为无 CRE）。系统内用户基于 Max-RSRP 策略直接选择关联基站。

2) 固定 CRE 偏置值为 5 dB（简称为 CRE=5 dB）。系统内所有 PBS 的 CRE 偏置值设定为 5 dB。

3) 基于分布式 Q-Learning 的动态 CRE 偏置设定（简称为分布式 Q-Learning）[19]。利用分布式 Q-Learning 算法动态优化 CRE 偏置值。其中，最小 CRE 偏置值设定为 1 dB，最大 CRE 偏置值设定为 9 dB，调整步频为 1 dB。

图 7 展示了在不同 CRE 设定方式下系统内不同类型用户数量。由图 7 可知，与 CRE=5 dB 相比，经过分布式 Q-Learning 算法与 VDN-CRE 算法动态优化 CRE 偏置值后，PEUE 数量有所下降，且 VDN-CRE 算法下降数量更多。这部分减少的 PEUE 成为 MUE，从而避免了强行关联 PBS 而遭受的来自 MBS 的强烈干扰。随着 $N_{UP}$ 的增大，PBS 覆盖范围内用户数量增加，此时，若继续保持 CRE 偏置值不变，将导致大量用户关联至 PBS，造成 PBS 资源挤兑。因此，应适当降低 CRE 偏置值，提高接入 PBS 的门槛，使一部分本就遭受 MBS 强干扰的 PEUE 切换至 MBS，在减轻用户所受小区间干扰的同时降低 PBS 负载压力，以此提高吞吐量。

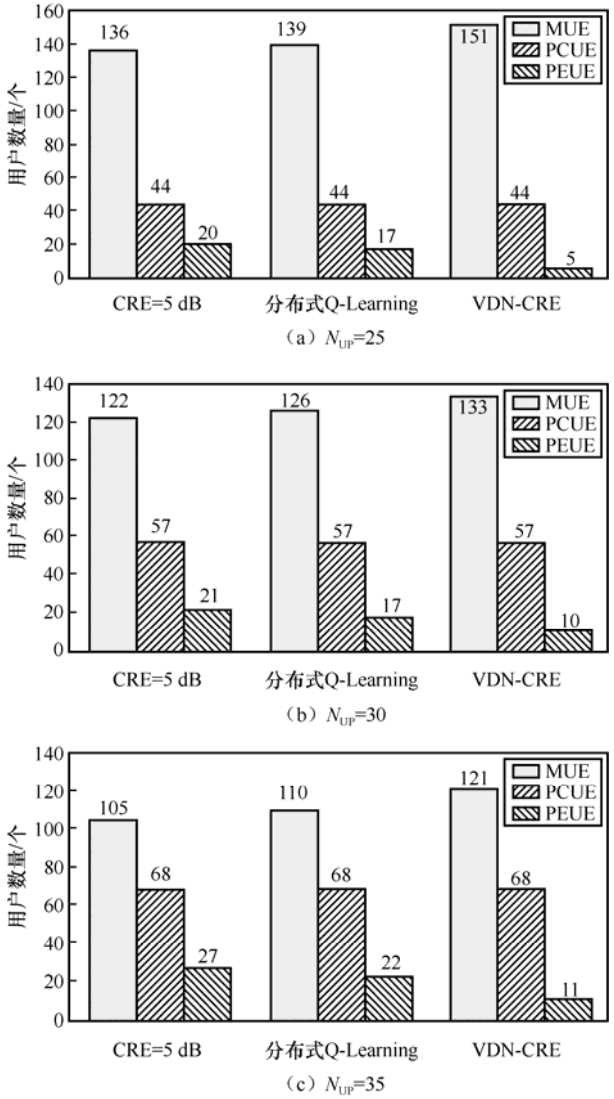


图 7　在不同 CRE 设定方式下系统内不同类型用户数量

图 8 展示了在不同 CRE 设定方式下系统吞吐量。其中，$x$ 轴表示在总用户数量不变的情况下，每个 PBS 覆盖范围内随机分布的用户数量 $N_{\mathrm{UP}}$；左侧 $y$ 轴对应的是柱状图，表示不同 CRE 设定方式下系统吞吐量与无 CRE 相比增加的系统吞吐量；右侧 $y$ 轴对应的是折线图，表示不同 CRE 设定方式下系统吞吐量大小及其随 $N_{\mathrm{UP}}$ 增大的变化情况。由图 8 可知，使用 CRE 技术能明显增加系统吞吐量，且分布式 Q-Learning 算法的效果好于 CRE=5 dB 的效果，VDN-CRE 算法的效果进一步好于分布式 Q-Learning 算法的效果。

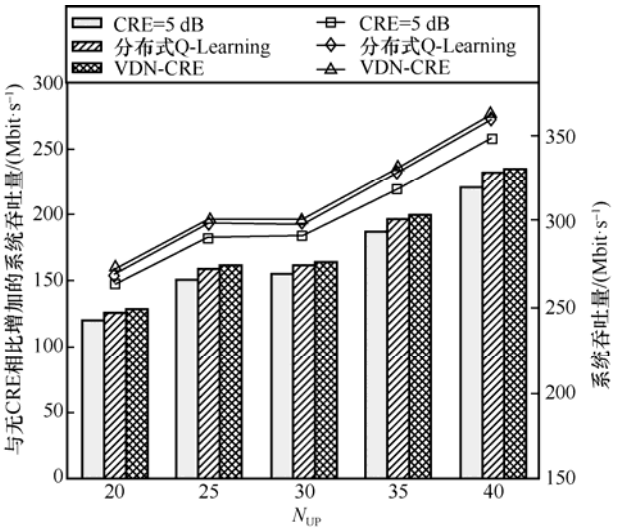


图 8　在不同 CRE 设定方式下系统吞吐量

图 9 展示了在不同 CRE 设定方式下系统内不同基站（MBS 或 PBS）所关联用户的吞吐量之和。由图 9 可知，与 CRE=5 dB 相比，分布式 Q-Learning 算法与 VDN-CRE 算法使关联至 MBS 的用户吞吐量之和下降，而关联至各个 PBS 的用户吞吐量之和上升。可见，通过动态调整各 PBS 的个性化 CRE 偏置值，牺牲一部分 MBS 用户的吞吐量以换取其他 PBS 用户吞吐量的提升，以此提升整个系统的吞吐量。VDN-CRE 算法的提升效果优于分布式 Q-Learning 算法。

图 10 展示了在不同 CRE 设定方式下系统内用户吞吐量的累积分布函数（CDF, cumulative distribution function）。为了获得更好的显示效果，图 10 纵坐标使用不同的间隔表示。由图 10 可知，对于每个 PBS 覆盖范围内用户数量 $N_{\mathrm{UP}}$ 取值为 25、30、35 的 3 种情况，使用 CRE 技术的用户吞吐量分布均明显优于不使用 CRE 技术的用户吞吐量，低吞吐量用户占比明显降低；动态优化 CRE 偏置值算法的用户吞吐量分布优于 CRE=5 dB 的用户吞吐量，且本文所提出的 VDN-CRE 算法能够明显改善用户吞吐量在 0～10 Mbit/s 范围内的占比，从而改善用户通信性能。

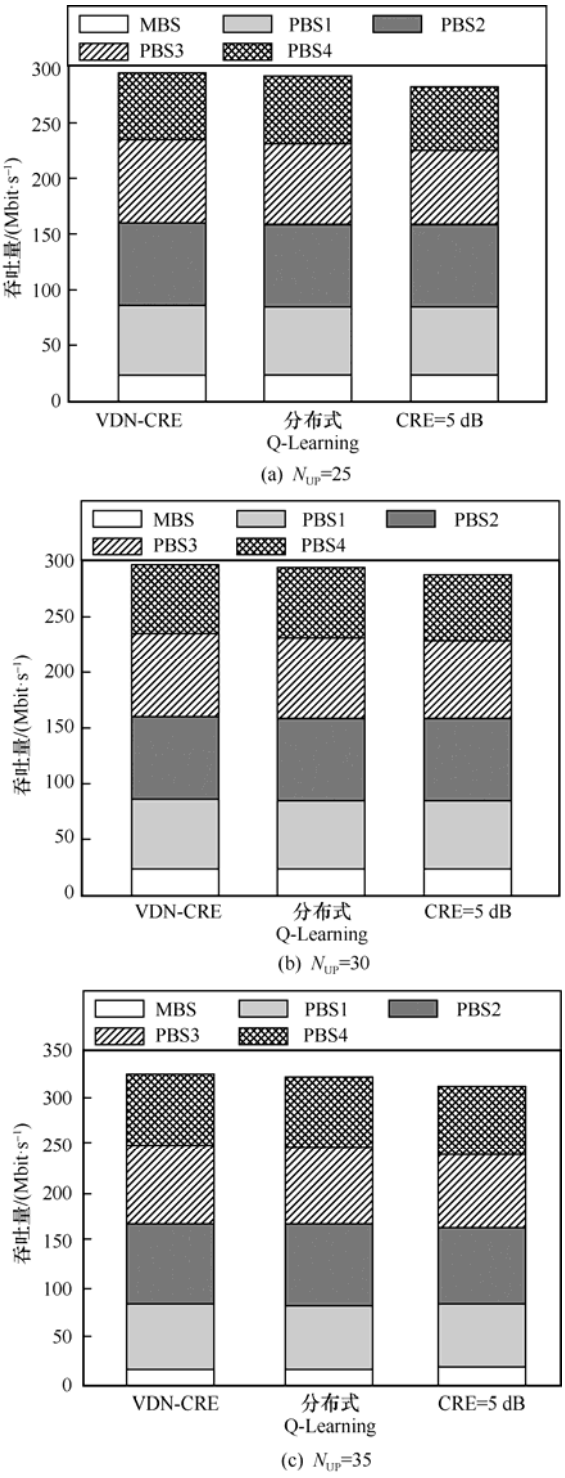


图 9　在不同 CRE 设定方式下系统内不同基站（MBS 或 PBS）所关联用户的吞吐量之和

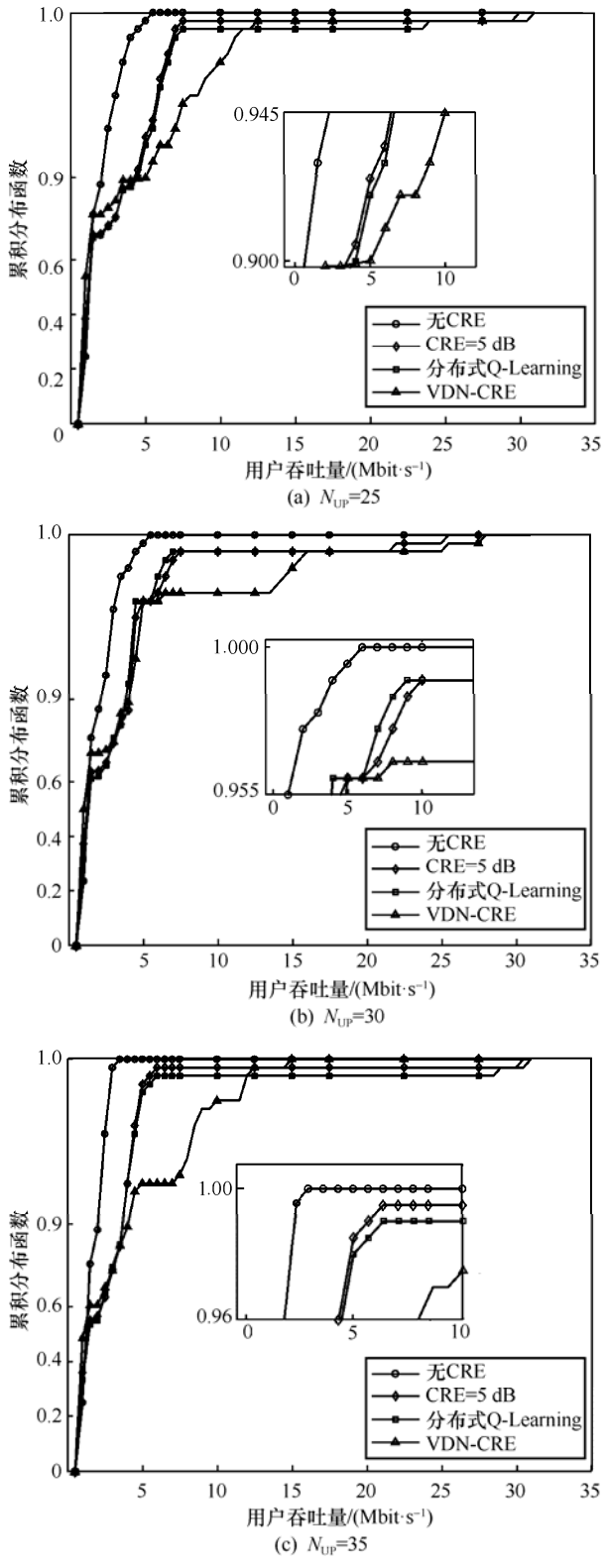


图 10　在不同 CRE 设定方式下系统内用户吞吐量的累积分布函数

## 4　结束语

本文针对 Macro-Pico 双层异构网络中 CRE 偏置值动态优化策略进行研究，以提高系统吞吐量。联合使用 eICIC 中 CRE 与 ABS 技术，针对现有基于强化学习的 CRE 偏置值动态优化算法中，集中式框架存在动作空间随 PBS 数量呈指数级增长、优化复杂度上升，完全分布式框架存在全局信息交互不足、陷入局部最优的问题，基于协作多智能体算法中集中训练、分布执行的 VDN 框架，根据 PBS 覆盖范围内 PEUE 数量及 PEUE 所受干扰情况，动态优化每个 PBS 的 CRE 偏置值。仿真结果表明，与 CRE=5 dB、分布式 Q-Learning 算法相比，VDN-CRE 算法能够有效提升系统吞吐量与边缘用户通信质量。未来可研究如何进一步降低 VDN-CRE 算法所需的通信开销问题。

## 参考文献：

[1] CHUANG K, YEKTAII H, OUTALEB N, et al. Towards sustainable networks: attacking energy consumption in wireless infrastructure with novel technologies[J]. IEEE Microwave Magazine, 2023, 24(12): 44-59.

[2] ELHOUSHY S, IBRAHIM M, HAMOUDA W. Cell-free massive MIMO: a survey[J]. IEEE Communications Surveys & Tutorials, 2022, 24(1): 492-523.

[3] XU Y J, GUI G, GACANIN H, et al. A survey on resource allocation for 5G heterogeneous networks: current research, future trends, and challenges[J]. IEEE Communications Surveys & Tutorials, 2021, 23(2): 668-695.

[4] 3GPP. Requirements for further advancements for evolved universal terrestrial radio access (E-UTRA) (LTE-advanced): TR 36.913[S]. 2011.

[5] BIANZINO A P, CHAUDET C, ROSSI D, et al. A survey of green networking research[J]. IEEE Communications Surveys & Tutorials, 2012, 14(1): 3-20.

[6] JAMIL S, ABBAS M S, UMAIR M, et al. A review of techniques and challenges in green communication[C]//Proceedings of International Conference on Information Science and Communication Technology (ICISCT). Piscataway: IEEE Press, 2020: 1-6.

[7] DAMNJANOVIC A, MONTOJO J, WEI Y B, et al. A survey on 3GPP heterogeneous networks[J]. IEEE Wireless Communications, 2011, 18(3): 10-21.

[8] ABBAS Z H, HAROON M S, MUHAMMAD F, et al. Enabling soft frequency reuse and stienen's cell partition in two-tier heterogeneous networks: cell deployment and coverage analysis[J]. IEEE Transactions on Vehicular Technology, 2021, 70(1): 613-626.

[9] LI J, WANG X M, LI Z Q, et al. Energy efficiency optimization based on eICIC for wireless heterogeneous networks[J]. IEEE Internet of Things Journal, 2019, 6(6): 10166-10176.

[10] MICHEL D D E, ROGER F B A, GUTENBERT K W J. Performance evaluation of the eICIC technique applied to a heterogeneous 4G mobile network[J]. European Journal of Applied Sciences, 2022, 10(2): 540-560.

[11] TORRES-CRUZ N, VILLORDO-JIMENEZ I, MONTIEL-SAAVEDRA A. Analysis of the geographical-information impact on the performance of ABS-CRE HetNets[J]. IEEE Latin America Transactions, 2020, 18(3): 613-622.

[12] JUNG T, SONG I, LEE S, et al. Cell range expansion with geometric information of pico-cell in heterogeneous networks[C]//Proceedings of IEEE 87th Vehicular Technology Conference (VTC Spring). Piscataway: IEEE Press, 2018: 1-5.

[13] LEE C N, LIN J H, WU C F, et al. A dynamic CRE and ABS scheme for enhancing network capacity in LTE-advanced heterogeneous networks[J]. Wireless Networks, 2019, 25(6): 3307-3322.

[14] 成思玥, 李浩然, 白卫岗, 等. 基于多智能体深度强化学习的测运控一体化资源调度方法[J]. 天地一体化信息网络, 2023, 4(1): 12-22.
CHENG S Y, LI H R, BAI W G, et al. Resource scheduling method for integration of TT&C and observation based on multi-agent deep reinforcement learning[J]. Space-Integrated-Ground Information Networks, 2023, 4(1): 12-22.

[15] 张彪, 汪西明, 徐逸凡, 等. 基于多智能体深度强化学习的多域协同抗干扰方法研究[J]. 物联网学报, 2022, 6(4): 104-116.
ZHANG B, WANG X M, XU Y F, et al. Multi-domain collaborative anti-jamming based on multi-agent deep reinforcement learning[J]. Chinese Journal on Internet of Things, 2022, 6(4): 104-116.

[16] 丁雨, 李晨凯, 韩会梅, 等. 基于 5G 无人机通信的多智能体异构网络选择方法[J]. 电信科学, 2022, 38(8): 28-36.
DING Y, LI C K, HAN H M, et al. Multi-agent heterogeneous network selection method based on 5G UAV communication[J]. Telecommunications Science, 2022, 38(8): 28-36.

[17] CHOI H, KIM T, PARK H S, et al. A cooperative online learning-based load balancing scheme for maximizing QoS satisfaction in dense HetNets[J]. IEEE Access, 2021, 9: 92345-92357.

[18] ALSUHLI G, BANAWAN K, ATTIAH K, et al. Mobility load management in cellular networks: a deep reinforcement learning approach[J]. IEEE Transactions on Mobile Computing, 2023, 22(3): 1581-1598.

[19] KUDO T, OHTSUKI T. Cell range expansion using distributed Q-learning in heterogeneous networks[J]. EURASIP Journal on Wireless Communications and Networking, 2013(1): 1-10.

[20] ASGHARI M Z, OZTURK M, HAMALAINEN J. Reinforcement learning based mobility load balancing with the cell individual offset[C]//Proceedings of IEEE 93rd Vehicular Technology Conference (VTC2021-Spring). Piscataway: IEEE Press, 2021: 1-5.

[21] TABUCHI S, MAKINO I, MIKI N. Combined usage of convex optimization and neural network for resource allocation[C]//Proceedings of 14th International Conference on Signal Processing and Communication Systems (ICSPCS). Piscataway: IEEE Press, 2020: 1-6.

[22] MATIGNON L, LAURENT G J, LE FORT-PIAT N. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems[J]. The Knowledge Engineering Review, 2012, 27(1): 1-31.

[23] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[C]//Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. New York: ACM Press, 2018: 2085-2087.

[24] DAI Y Y, ZHANG K, MAHARJAN S, et al. Deep reinforcement learning for stochastic computation offloading in digital twin networks[J]. IEEE Transactions on Industrial Informatics, 2021, 17(7): 4968-4977.

[25] FERIANI A, HOSSAIN E. Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: a tutorial[J]. IEEE Communications Surveys & Tutorials, 2021, 23(2): 1226-1252.

[26] WANG H N, LIU N, ZHANG Y Y,et al. Deep reinforcement learning: a survey[J]. Frontiers of Information Technology & Electronic Engineering, 2020, 21(12): 1726-1744.

[27] TABISH R, MIKAYEL S, SCHROEDER D W C, et al. Monotonic value function factorisation for deep multi-agent reinforcement learning[J]. Journal of Machine Learning Research, 2020, 21(1): 7234-7284.

[28] CASTELLINI J, OLIEHOEK F A, SAVANI R, et al. The representational capacity of action-value networks for multi-agent reinforcement learning[C]//Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. New York: ACM Press, 2019: 1862-1864.

[作者简介]

**张铖**（1988− ），男，安徽望江人，博士，东南大学副教授、博士生导师，主要研究方向为无线通信系统的空时信号处理、机器学习辅助的无线通信智能优化技术等。

**朱家烨**（1998− ），女，江苏无锡人，东南大学硕士生，主要研究方向为无线通信网络中的多小区干扰协调。

**刘泽宁**（1993− ），男，江苏淮安人，博士，网络通信与安全紫金山实验室研究员，主要研究方向为边缘计算、智能干扰优化和资源分配技术。

**黄永明**（1977− ），男，江苏吴江人，博士，网络通信与安全紫金山实验室研究员，东南大学博士生导师，主要研究方向为智能 5G/6G 移动通信、毫米波无线通信等。