引用格式：王越, 刘如意, 杨蓓, 等. 基于多智体学习的多小区 NOMA 协作波束训练 [J]. 电子科技大学学报, 2025, 54(6): 866-874.
WANG Y, LIU R Y, YANG B, et al. Multi-cell NOMA cooperative beam training based on multi-agent learning[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(6): 866-874.

基于多智体学习的多小区 NOMA 协作波束训练

王 越[1]，刘如意[2]，杨 蓓[1]，王建秀[1]，冯 钢[2*]

（1. 中国电信股份有限公司北京研究院，100083；2. 电子科技大学 通信抗干扰全国重点实验室，成都 611731）

摘要：该文主要研究毫米波网络中协作非正交多址（non-orthogonal multiple access，NOMA）下的多小区的波束赋形优化问题。为了最大化系统吞吐量，并且考虑用户位置及信道信息，将基站的波束配置问题建模为马尔竞争博弈问题，并采用强化学习算法多智体深度确定性策略梯度（multi-agent deep deterministic policy gradient，MADDPG）对其求解，设计了一种多智能体强化学习的多小区协作 NOMA 波束赋形训练算法，以合理分配多基站体系中的波束、功率等资源，并提高系统的吞吐量。仿真结果表明，提出的 MADDPG 算法能达到更好的系统吞吐量及用户覆盖率。

关键词：波束管理；波束训练；多小区 NOMA；深度强化学习；多智能体学习
中图分类号：TN929.5 **文献标志码：**A DOI: 10.12178/1001-0548.2024207

Multi-cell NOMA cooperative beam training based on multi-agent learning

WANG Yue[1], LIU Ruyi[2], YANG Bei[1], WANG Jianxiu[1], and FENG Gang[2*]

(1. China Telecom Corporation Beijing Research Institute, Beijing 100083, China;
2. National Key Laboratory of Wireless Communications, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: This paper mainly focuses on the beamforming training problem in cooperative non-orthogonal multiple access (NOMA) scenarios in millimeter-wave communication, extending the work from single-cell NOMA to multi-cell NOMA scenarios. To maximize system throughput while considering user locations and channel information, the beam configuration problem at the base station is modeled as a Markov cooperative-competitive game problem. And then the problem is solved by exploiting multi-agent deep deterministic policy gradient (MADDPG) based reinforcement learning algorithm. A multi-agent reinforcement learning-based beamforming training algorithm for cooperative NOMA in multi-cell scenarios is designed to effectively allocate resources such as beams and power in multi-base station systems, thereby enhancing system throughput. Numerical simulations demonstrate that the proposed MADDPG algorithm achieves better system throughput and user coverage.

Key words: beam management; beam training; multi-cell NOMA; deep reinforcement learning; multi-agent learning

随着无线通信领域的发展，5G 网络已经在世界范围内进行部署，第六代移动通信系统（6th generation mobile communication system, 6G）的研究也正在进行中。对于 6G，研究人员期待更高的系统容量、更低的时延及更强的可靠性[1-2]。为了实现这一目标，毫米波大规模多输入输出（multiple input multiple output, MIMO）成为研究的热点[3]。在 MIMO 系统中，基站利用波束赋形技术同时生成指向不同方向的多个波束。由于这些波束具有较强的指向性，因此可以在空间域中轻松加以区分。系统为每个用户分配特定的波束或波束对，从而保证用户只接收面向自身的信号。在此基础上，可在

收稿日期：2024−08−06
基金项目：国家自然科学基金青年基金项目（62201121）；中国电信研究院合作项目（231382）
作者简介：王越，博士研究生，主要从事移动通信方面的研究。
*通信作者 E-mail: fenggang@uestc.edu.cn

单个时频资源上复用多个用户来增加系统容量。但是，这样的非正交叠加方式使得用户在时频二维空间的区分上变得困难。为了解决这一问题，常使用非正交多址（non-orthogonal multiple access, NOMA）技术引入功率域，为占用相同时频资源的用户分配不同的发射功率。在 MIMO-NOMA 系统中，除了利用空间域进行区分外，还引入了功率域复用。具体来说，对于方位接近的用户，可以通过分配不同的功率来加以区分，从而实现多个用户共享相同的时频资源。用户可根据空间位置划分为不同的用户簇，每个簇包含方位相似但信道特性不同的用户，而不同簇之间的用户方位则存在明显差异。通过波束赋形技术可以有效分离不同用户簇的信号，减轻簇间干扰；同时，在每个簇内采用 NOMA 方案，通过分配不同的发射功率进一步区分用户，降低簇内干扰。

为了实现高频谱效率，NOMA 与 MIMO 通信相结合，已扩展到多小区系统[4-5]。多小区网络存在小区间干扰（inter-cell interference, ICI），这导致小区边缘用户的数据速率降低。为了减轻在小区边缘用户处接收到的干扰，使用协同多点传输（coordinated multi-point, CoMP）促进跨多个小区的协作传输[6]。将多小区 NOMA 与 CoMP 结合，利用空间分集增益和功率复用的优势，进一步提高了 SE 并扩大了网络内的覆盖范围[7-10]。CoMP 技术可以分为：协调调度和协调波束赋形（coordinated scheduling/coordinated beamforming, CS/CB）、联合传输（joint transformation, JT）和传输点选择（transmission point selection, TPS）[11]。其中，CB-CoMP 的实现方式为基站根据用户位置和网络条件，协调各自的波束形状和方向。

在多小区 NOMA 网络中，资源分配问题比单小区场景更为复杂。进行资源分配（如波束赋形、用户配对和功率分配）时，不仅需要考虑小区间干扰，还要处理小区内部的簇内和簇间干扰，这使问题的复杂度显著提升。由于此类联合优化问题通常是非凸的，研究者常将其分解为若干子问题加以求解。机器学习方法可以嵌入一个或多个子问题中，与传统优化方法结合使用。然而，当协作场景规模扩大、变量数量急剧增加时，传统机器学习已难以应对。

近年来，深度强化学习被广泛引入，用于解决 NOMA 系统中的各类资源分配挑战。如文献 [9] 通过研究功率和子阵比率分配的连续性以及子带分配的离散性，提出了一种混合离散和连续动作的多任务深度强化学习算法，以解决长期波束成形中的带宽功率分配问题。文献 [12] 使用对决双深度 Q 网络以保证 NOMA 系统中用户的公平性。文献 [13] 采用深度 Q 网络完成子载波分配，使用深度确定性梯度完成功率分配。文献 [14] 使用多智能体深度 Q 网络完成大规模 MIMO-NOMA 中的用户分组、功率分配及波束形成。深度强化学习技术作为一种极具前景的研究方向，也是本文采取的决策算法设计方式。文献 [12-14] 的研究仅限于单小区 MIMO-NOMA 系统，由于多小区 NOMA 系统所面临的无线通信环境更为复杂和具有动态性，很难直接将上述研究的算法直接应用到多小区 NOMA 系统中。因此，多小区 NOMA 系统中的联合资源分配优化问题仍充满挑战。

基于以上考虑，本文使用 CB-CoMP 技术将单小区 NOMA 拓展到多小区 NOMA 场景，研究具有 NOMA 的多小区网络中的功率分配与波束赋形问题。为了使系统吞吐量最大化，同时实现基站侧波束配置策略的智能化，本文提出了一种基于多智体深度确定性策略梯度（multi-agent deep deterministic policy gradient, MADDPG）的智能波束赋形算法。该策略应用多智体深度强化学习算法 MADDPG，实现了连续状态动作空间下的连续动作输出，具有优秀的探索能力，将多小区协调波束训练转化为多智能体联合训练，以提升系统吞吐量。

1　单小区 NOMA 与多小区 NOMA

毫米波 MIMO 技术利用大量的天线单元，以实现高速数据传输和高频谱利用率。在 MIMO 系统中，主要使用数字波束赋形、模拟波束赋形和模拟数字（模数）波束赋形 3 种方式。传统 MIMO 系统常采用数字波束赋形，每根天线连接一条 RF 链，硬件复杂度高，且对通信系统的带宽以及信号采样率要求较高。

已有的毫米波 MIMO 研究采用模数波束赋形技术，在高维部分使用模拟波束赋形，在低维部分使用数字波束赋形，使用较少的 RF 链实现接近数字波束赋形的性能，同时在一定程度上减少了硬件复杂度和成本。因此，本文采用模数波束赋形技术。为了进一步提高频谱效率和连接密度，在毫米

波 MIMO 系统中使用 NOMA，如图 1 所示。在这种情况中，每个选定波束内可以同时服务一个以上的用户。

在此基础上，将 MIMO-NOMA 系统进行拓展，考虑多小区 NOMA 网络。在这个网络中，用户会受到两种干扰，即内部干扰和外部干扰。内部干扰来自同小区内的其他用户，外部干扰来自其他小区的用户，在多小区环境中，尤其是当小区之间的边界用户同时由多个小区基站服务时，来自其他小区用户的信号会对其产生干扰。在解码时，NOMA 使用连续干扰消除技术，在多小区环境中，每个小区内的用户首先尝试解码其他小区的信号，然后是本小区内其他用户的信号。

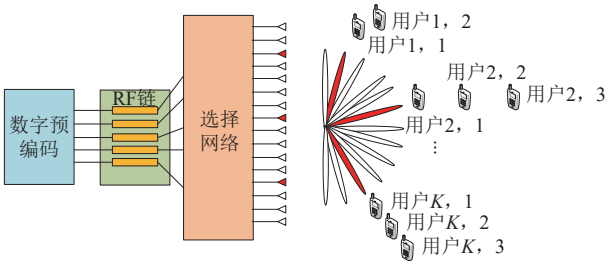

图 1 毫米波 MIMO-NOMA 系统

总体而言，在多小区 NOMA 系统中，主要的设计问题是在一定的性能指标下联合或部分地优化波束赋形、功率分配和用户聚类。在本研究中，考虑联合优化多小区 NOMA 系统中的波束赋形以及功率分配，对于用户聚类，将采用固定的原则来简化问题。

2 问题建模

2.1 系统模型

本文使用 NOMA 的下行链路多小区网络，如图 2 所示，其中存在I个 BS 和J个用户。分别用$\boldsymbol{I}=\{1,2,\cdots,I\}$和$\boldsymbol{J}=\{1,2,\cdots,J\}$表示 BS 和用户的集合。由 BS $(i\in\boldsymbol{I})$服务的唯一用户组由集合$\boldsymbol{J}_i=\{J_{i-1}+1, J_{i-1}+2,\cdots,J_i\}$表示，其中$J_0=0$，$J_I=J$，$J_i=\sum_{l=1}^{i}|\boldsymbol{J}_l|$。系统为小区之间存在干扰的下行链路。

假设 BS 上的天线可以同时生成N个窄波束。通过使用 NOMA，BS 通过在功率域中对多个用户进行拆分来为多个用户提供服务。假设每个 BS 共享相同的频谱，对每个 BS，总带宽B_{total}被等分为M个子信道，其中每个子信道的带宽为$B=B_{\text{total}}/M$。设$\boldsymbol{M}=\{1,2,\cdots,M\}$是子信道的集合。基站 i 在子信道 m 上服务的用户集合由$\boldsymbol{J}_{im}=\{J_{i(m-1)}+1, J_{i(m-1)}+2,\cdots, J_{im}\}$表示。子信道$m$上基站 i 和用户 $j\in\boldsymbol{J}_{im}$之间的信道增益用h_{ijm}表示。在不失一般性的情况下，信道排序为$|h_{i(J(m-1)+1)m}|\leqslant|h_{i(J(m-1)+2)m}|\leqslant\cdots\leqslant|h_{iJ_{im}m}|$，$\forall i\in\boldsymbol{I},m\in\boldsymbol{M}$。

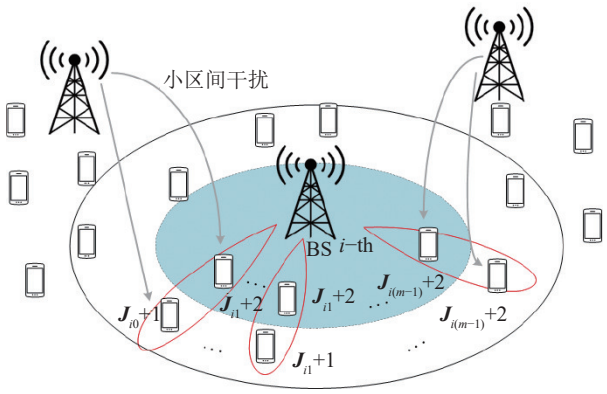

图 2 基于 NOMA 的下行链路多小区网络模型

根据 NOMA 原理，基站 i 同时向$\boldsymbol{J}_{im}$中的所有服务用户发送信号s_{im}。传输信号s_{im}可表示为：

$$s_{im}=\sum_{j=J_{i(m-1)+1}}^{J_{im}}\sqrt{p_{ijm}}s_{ijm} \tag{1}$$

式中，s_{ijm}和p_{ijm}分别是用户$j\in\boldsymbol{J}_{im}$的信号和分配功率。

用户$j\in\boldsymbol{J}_{im}$在子信道m上的观测由以下等式给出：

$$\begin{aligned}y_{ijm}=&h_{ijm}s_{im}+\sum_{k\in I\setminus\{i\}}h_{kjm}s_{km}+n_{jm}=\\&\sum_{l=J_{i(m-1)+1}}^{J_{im}}h_{ijm}\sqrt{p_{ijm}}s_{ilm}+\\&\sum_{k\in I\setminus\{i\}}\sum_{n=J_{k(m-1)+1}}^{J_{km}}h_{kjm}\sqrt{p_{knm}}s_{knm}+n_{jm}\end{aligned} \tag{2}$$

式中，h_{kjm}是子信道m上基站 k 和用户 j之间的跨信道增益；而n_{jm}表示方差为σ^2的加零性平均高斯噪声。根据文献 [15-18]，每个用户在解码自己的信息之前，应该解码同一小区中信道增益较低的其他用户信息。

用$q_{im}=\sum_{j=J_{i(m-1)+1}}^{J_{im}}p_{ijm}$表示基站 i 在子信道m上的总发射功率，用户$j\in\boldsymbol{J}_{im}$在子信道上检测用户$l\in\{J_{i(m-1)+1},\cdots,j\}$的传输速率为：

$$r_{ijlm}=B\log_2\left(1+\frac{\left|h_{ijm}\right|^2 p_{ilm}}{\left|h_{ijm}\right|^2\sum\limits_{n=l+1}^{J_{im}}p_{inm}+Z_{ijlm}}\right) \tag{3}$$

其中，

$$Z_{ijlm}=\sum_{k\in\boldsymbol{I}\backslash\{i\}}q_{km}\left|h_{kjm}\right|^2+\sigma^2 \tag{4}$$

根据式 (3)，具有高信道增益的强用户 j需要解码具有低信道增益的弱用户$l\leqslant j$的消息。为了确保 SIC 成功，用户 $j\in\boldsymbol{J}_{im}$在子信道m上的速率由下式给出：

$$r_{ijm}=\min_{l\in\{j,\cdots,J_{im}\}}B\log_2\left(1+\frac{p_{ijm}}{\sum\limits_{n=j+1}^{J_{im}}p_{inm}+\dfrac{Z_{ijlm}}{\left|h_{ijm}\right|^2}}\right)=B\log_2\left(1+\frac{p_{ijm}}{\sum\limits_{n=j+1}^{J_{im}}p_{inm}+g_{ijm}}\right) \tag{5}$$

其中，

$$g_{ijm}=\max_{l\in\{j,\cdots J_m\}}\frac{\sum\limits_{k\in\boldsymbol{I}\backslash\{i\}}q_{km}|h_{kjm}|^2+\sigma^2}{|h_{ijm}|^2} \tag{6}$$

2.2　优化问题建模

应用式 (5)，多小区 NOMA 波束赋形问题可表述为以下线性约束：

$$\min_{p\geqslant 0,q\geqslant 0}V(\boldsymbol{p},\boldsymbol{q}) \tag{7}$$

$$\text{s.t. } q_{im}=\sum_{J_{i(m-1)+1}}^{J_{im}}p_{ijm},\forall i\in\boldsymbol{I},m\in\boldsymbol{M} \tag{7a}$$

$$p_{ijm}\geqslant\left(2^{\frac{R_{ijm}}{B}}-1\right)\left(\sum_{n=j+1}^{J_{im}}p_{in}+g_{ijm}\right) \tag{7b}$$

$$\forall i\in\boldsymbol{I},m\in\boldsymbol{M},j\in\boldsymbol{J}_{im} \tag{7c}$$

$$\sum_{m=1}^{M}q_{im}\leqslant Q_i,\forall i\in\boldsymbol{I} \tag{7d}$$

式中，目标函数的表达式如式 (8) 所示，$\boldsymbol{p}=\left[p_{111},\cdots,p_{J_1M},\cdots,p_{J_1MM}\right]^{\mathrm{T}}$是传输功率矢量；$\boldsymbol{q}=\left[q_{11},\cdots,q_{1M},\cdots,q_{IM}\right]^{\mathrm{T}}$是总传输功率矢量；$g_{ijm}$在式 (6) 中定义；$Q_i$是基站 i 的最大传输功率。

约束 (7a) 即用户在子信道的发射功率之和等于基站发射功率。约束 (7b) 反映了可以满足所有用户的最小速率需求。约束 (7d) 表示用户连接基站的发射功率之和小于基站的最大传输功率。

$$V(\boldsymbol{p},\boldsymbol{q})=\sum_{i=1}^{I}\sum_{m=1}^{M}\sum_{j=J_{i(m-1)+1}}^{J_{im}}r_{ijm} \tag{8}$$

3　基于 MADDPG 的多小区 NOMA 波束训练算法

在本文提出的多小区 NOMA 系统中，用户被划分为若干用户簇，每个簇包含多个方向角相近但信道特性差异显著的用户。本文提出联合设计波束赋形和功率分配方案，以有效抑制簇内及簇间干扰，从而提升能量效率。用户分簇本质上是一个离散优化问题，当搜索空间有限时可以通过穷举法获得最优解，但随着系统用户数的增加，穷举搜索的复杂度会呈指数增长。为在性能与复杂度之间取得平衡，引入一种简化的分簇策略：先按空间位置对用户进行分类，将靠近基站的用户定义为强用户，将远离基站的用户定义为弱用户，然后从强用户组和弱用户组中各选取一名用户，组成一个双用户簇。

若将场景中的基站视为一个 agent，并将系统中其他基站视作环境，这时 agent 将面临复杂且不稳定的环境，当 agent 在优化资源分配策略时，被视作环境的其他基站也在同时进行策略的优化，此时，使用强化学习进行交互时，环境的近似概率分布是复杂多变的，这并不利于基站的学习训练。除此之外，多小区 NOMA 中，每个基站的决策是同时进行的，若采用序贯决策的方式进行训练，一个 agent 变化时，会引起其他 agent 环境的剧烈变化。在多智能体框架下，各智能体得以互相协调并行求解，避免了复杂的环境变化问题。

因此，考虑将多小区 NOMA 场景视作一个多智能体的场景，需要进行波束配置的基站在这个问题中被视为 agent。采用 MADDPG 算法，对问题进行求解。该算法采用了中心化训练及分布式执行的框架，即所有智能体分享同一个价值网络，该网络在训练过程中对每个智能体的决策网络作出指导，在执行时，每个策略网络独立地进行动作选择，所有智能体的动作集合为$\{a_1,a_2,\cdots,a_N\}$。每个智能体进行动作选择之后与环境进行交互，得到奖励$\{r_1,r_2,\cdots,r_N\}$。将所有智能体的动作、奖励及观测值储存到经验池中，$(\boldsymbol{X},\boldsymbol{X}',a_1,a_2,\cdots,a_N,r_1,r_2,\cdots,r_N)$，其中$\boldsymbol{X}=(o_1,o_2,\cdots,o_N)$代表所有智能体

的观测结果。之后，智能体从经验池中随机采样经验样本，作为策略网络的小批量训练样本数据，且价值网络将 Q 函数关于动作的梯度信息传到目标网络，用于更新预测策略网络参数μ_{θ_i}。

深度确定性策略梯度（deep deterministic policy gradient, DDPG） 使用深度神经网络对确定性策略函数$\mu(s,\boldsymbol{\theta})$进行逼近，其中$\boldsymbol{\theta}$为网络参数，$s$代表当前的状态，其输出为确定性动作值$a$。优化目标被定义为累计折扣奖赏：

$$J(\boldsymbol{\theta})=\mathrm{E}_{\boldsymbol{\theta}}\left[r_0+\gamma r_1+\gamma^2 r_2+\cdots\right] \tag{9}$$

为了找到最优的确定性行为$\boldsymbol{\theta}^*$，该问题等价于最大化目标函数$J(\boldsymbol{\theta})$中的策略，即：

$$\boldsymbol{\theta}^*=\arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \tag{10}$$

目标函数$J(\boldsymbol{\theta})$是关于参数$\boldsymbol{\theta}$的梯度，等价于动作函数$Q(s,a,\boldsymbol{w})$关于$\boldsymbol{\theta}$梯度的期望：

$$\nabla_\theta J_\beta(\boldsymbol{\theta})=\mathrm{E}_{s\sim\rho^\beta}\left[\nabla_\theta Q(s,a,\boldsymbol{w})\right] \tag{11}$$

式中，β表示行为策略，是一种探索性的策略，通过引入随机噪声影响动作的选择。ρ^β为状态的分布，即智能体根据行为策略产生的状态分布。

因此根据确定性策略$a=\mu(s,\boldsymbol{\theta})$，遵循链式法则对目标函数进行求导，得到策略网络的更新方式：

$$\nabla_\theta J_\beta(\boldsymbol{\theta})=\mathrm{E}_{s\sim\rho^\beta}\left[\nabla Q(s,a,\boldsymbol{w})\nabla_\theta\mu(s,\boldsymbol{\theta})\right] \tag{12}$$

利用小批量梯度上升法（mini-batch gradient ascent, MBGA），从经验池中随机采样获得 N 个小批量数据作为对期望值的采样估计：

$$\nabla_\theta \hat{J}_\beta(\boldsymbol{\theta})\approx\frac{1}{N}\sum_{i+1}^{N}\left[\nabla_a Q(s_i,a_i,\boldsymbol{w})\nabla_\theta\mu(s_i,\boldsymbol{\theta})\right] \tag{13}$$

价值网络：DDPG 使用深度网络对动作值函数$Q(s,a,\boldsymbol{w})$进行逼近，$\boldsymbol{w}$为网络参数。

与 DQN 一样，DDPG 利用 TD 误差的均方误差作为损失函数，它们的区别在于目标值y_i：

$$L(w)=\mathrm{E}\left[\begin{array}{l}(r+\gamma Q'(s',\mu'(s',\boldsymbol{\theta}'),w')-\\ Q(s,a,\boldsymbol{w}))^2\end{array}\right] \tag{14}$$

可以看出，目标值$y_i=r_i+\gamma Q'(s_i',\mu'(s_i',\boldsymbol{\theta}'),w')$涉及目标策略网络$\mu'$和目标价值网络$Q'$，这样可以使预测价值网络在学习时变得更加稳定，增强其收敛性。

价值网络的目标是最小化损失函数，故采用小批量随机梯度下降（mini-batch stochastic gradient descent, MSGD）法，从经验池 D 中随机采样获得 N 个小批量数据作为对期望值的采样估计：

$$\begin{aligned}&\nabla_w L(\boldsymbol{w})\approx 1/N\times\\&\sum_{i+1}^{N}(r_i+\gamma Q'(s'_i,\mu'(s'_i,\boldsymbol{\theta}'),w')-\\&\quad Q(s_i,a_i,\boldsymbol{w}))\nabla_w Q(s_i,a_i,\boldsymbol{w})\end{aligned} \tag{15}$$

式中，$\boldsymbol{\theta}'$和$\boldsymbol{w}'$分别代表目标策略网络μ'和目标值网络Q'的权重参数。DDPG 在每次迭代中利用经验回放机制，从样本池中随机选取 N 条转移样本。通过将价值网络中 Q 值函数对动作的梯度传递到策略网络，沿提升 Q 值的方向更新策略参数，以逐步逼近最优策略。预测价值网络参数采用小批量梯度下降优化，目标网络参数则依据文献 [19] 提出的软更新方法进行调整。

多智能体 DDPG（MADDPG）算法是对于每个智能体实现一个 DDPG 的算法。所有智能体共享一个中心化的 Critic 网络，该 Critic 网络在训练的过程中同时对每个智能体的 Actor 网络给出指导，而执行时每个智能体的 Actor 网络则是完全独立做出行动，即去中心化地执行。MADDPG 与 DDPG 的不同在于，MADDPG 中每个 agent 的策略网络能够获得其他 agent 的策略信息，假设每个 agent 的策略参数为$\boldsymbol{\theta}=\{\theta_1,\theta_2,\cdots\theta_N\}$，所有 agent 的策略集合为$\pi=\left\{\pi_{\theta_1},\pi_{\theta_2},\cdots,\pi_{\theta_N}\right\}$，对于确定性策略，梯度公式为：

$$\begin{aligned}&\nabla_{\theta_i}J(\theta_i)=\\&\mathrm{E}_{\boldsymbol{X}\sim D}\left[\nabla_{\theta_i}\mu_i(o_i)\nabla_{a_i}Q_i^{\mu}(\boldsymbol{X},a_1,\cdots,a_N)\right]\\&a_i=\mu_i(o_i)\end{aligned} \tag{16}$$

式中，$Q_i^{\mu}(\boldsymbol{X},a_1,a_2,\cdots,a_N)$为中心化的动作价值函数；$\boldsymbol{X}=(o_1,o_2,\cdots,o_N)$包含了所有 agent 的观测；$D$为存储数据的经验回放池，它存储的数据为$(\boldsymbol{X},\boldsymbol{X}',a_1,a_2,\cdots,a_N,r_1,r_2,\cdots,r_N)$。

同样，MADDPG 使用基于 TD 误差的均方误差作为损失函数，可表示为：

$$\begin{aligned}&L(\omega_i)=\mathrm{E}_{\boldsymbol{X},a,r,\boldsymbol{X}'}\left[(Q_i^{\mu}(\boldsymbol{X},a_1,a_2,\cdots,a_N)-y)^2\right]\\&y=r_i+\gamma Q_i^{\mu'}(\boldsymbol{X}',a'_1,a'_2,\cdots,a'_N)|_{a'_j=\mu'_j(o_j)}\end{aligned} \tag{17}$$

式中，$\mu'=(\mu'_{\theta_1},\mu'_{\theta_2},\cdots,\mu'_{\theta_N})$是更新价值函数中使用的目标策略的集合。

基于 MADDPG 的多小区 NOMA 波束赋形策略算法可归纳如算法 1 所示。

算法 1　基于 MADDPG 的多小区 NOMA 波束赋形

初始化预测策略网络$\mu_{\theta_i}(o)$和预测价值网络$Q_i^{\mu}(\boldsymbol{X},a_1,a_2,\cdots,a_N)$

初始化目标策略网络$\mu'_{\theta_i}(o)$和目标价值网络$Q_i^{\mu'}(\boldsymbol{X},a_1,a_2,\cdots,a_N)$

经验池D的容量为M，总迭代次数I，折扣系数γ、τ，随机小批量采样样本数量m

for communication round e= 1, 2, ⋯ do

初始化一个随机过程N，用于动作的探索

所有智能体的初始观测$\boldsymbol{X}$

repeat（情节中的每一个时间步t）

根据当前的预测策略网络和探索噪声来选择动作$a_i=\mu_{\theta_i}(o_i)+N_t$

执行动作$\boldsymbol{a}=(a_1,a_2,\cdots,a_N)$，获得奖赏$r$和新的观测$\boldsymbol{X}'$

将经验转换$(\boldsymbol{X},a,r,\boldsymbol{X}')$储存在经验回访池

从经验池D中随机采样小批量的m个经验转移样本，计算目标值

$$y=r_i+\gamma Q_i^{\mu'}(\boldsymbol{X}',a'_1,a'_2,\cdots,a'_N)|_{a'_j=\mu'_j(o_j)}$$

使用 MBGD，根据最小化损失函数来更新 critic 网络的参数$\boldsymbol{\omega}$

$$L(\boldsymbol{\omega})=\mathrm{E}_{\boldsymbol{X},a,r,\boldsymbol{X}'}\left[(Q_i^{\mu}(\boldsymbol{X},a_1,a_2,\cdots,a_N)-y)^2\right]$$

使用 MBGA 法，根据最大化目标函数更新策略网络参数θ_i

$$\nabla_{\theta_i}J(\theta_i)=\mathrm{E}_{\boldsymbol{X}\sim D}\begin{bmatrix}\nabla_{\theta_i}\mu_i(o_i)\\ \nabla_{a_i}Q_i^{\mu}(\boldsymbol{X},a_1,a_2,\cdots,a_N)\end{bmatrix}$$

对每个智能体i，更新目标 actor 网络和目标 critic 网络

end for

end for

4　仿真结果与分析

4.1　仿真设置

本研究使用仿真实验来测试提出的基于 MADDPG 的波束赋形训练算法，仿真基于 python 编程语言及 pytorch 深度学习框架。同时，为了说明提出的波束赋形算法的优势，将此算法与智能算法 DDPG 和非智能的启发式算法进行比较。仿真网络场景的参数设置如表 1 所示。

在仿真过程中首先使用智能化算法使智能体在环境中进行训练，在算法收敛之后运行，使用已训练好的模型进行验证。MADDPG 算法以及 DDPG 算法的具体参数如表 2 和表 3 所示。

表 1　系统仿真参数

仿真参数	仿真值
小区半径/m	200
天线数量/个	100
基站最大发射功率/dBm	46
总可用毫米波带宽/MHz	50
小区个数/个	6
RF链数量/个	3
用户噪声功率/dBm	−80

表 2　MADDPG 算法仿真参数

MADDPG仿真参数	仿真值
Actor网络学习率	0.016
Critic网络学习率	0.016
目标网络更新参数	0.01
衰减因子	0.9
全连接层神经元数	[128, 64]
经验回放池	100 000
批量样本数	256
每episode训练步	200

表 3　DDPG 算法仿真参数

DDPG仿真参数	仿真值
Actor网络学习率	0.02
Critic网络学习率	0.02
目标网络更新参数	0.01
衰减因子	0.9
全连接层神经元数	[128, 64]
经验回放池	100 000
批量样本数	256
每episode训练步	200

4.2　仿真结果分析

本研究通过仿真实验评估了 MADDPG 方案的性能。采用以下 3 种方案作比较。

1）基于贪婪策略的全局功率分配算法（Epsilon-greedy）：在此方案中，系统根据当前网络环境，基于贪婪策略，选择使得网络全局吞吐量最大的用户功率分配方案。

2）单智体 DDPG 算法（DDPG）：在此方案中，每个基站作为独立智能体，以自身吞吐量最大化为目标进行独立训练。与 MADDPG 算法的区别在于，每个基站独立训练本地模型，不与其他基站进行协作。

3）随机分配功率算法（Random）：基站随机分配功率给用户。

首先验证网络的收敛性，如图 3 及图 4 所示，MADDPG 算法的 TD loss 在 episode 内达到了较低值，表示此时参数更新已完成，价值网络已收敛。同样地，DDPG 算法也在 5 000 轮次内达到了较低的水平，趋于平缓，可得出本实验算法已经完成收敛。

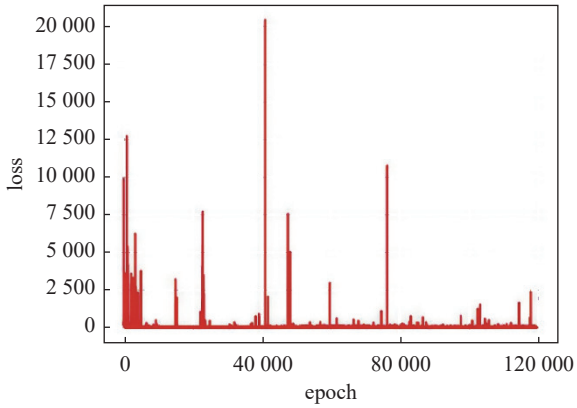

图 3 MADDPG 算法 TD-loss

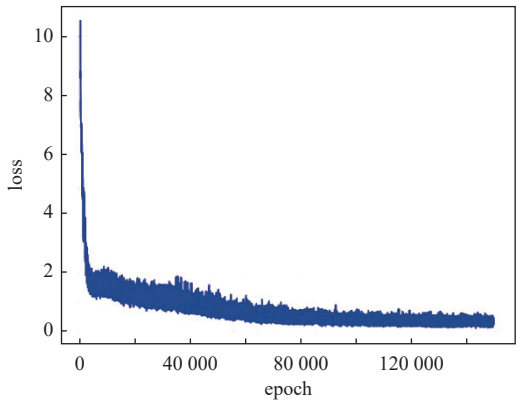

图 4 DDPG 算法 TD-loss

随后，仿真比较不同算法得到的 Reward（系统吞吐量），结果如图 5 所示。从图中可以看出，MADDPG 算法获得收益总是优于其他 3 种比较算法。这是因为与基于单智体的 DDPG 算法相比，MADDPG 考虑了多个智能体之间相互协作来改善系统性能。与全局的启发式算法 Epsilon-greedy 相比，Epsilon-greedy 算法只是根据当前的网络状态进行即时的优化决策，没有考虑系统的动态变化特征和规律对系统长时间性能的影响，而 MADDPG 算法能够基于采用强化学习的方法，优化系统的长效性能，从而提升系统平均吞吐量。DDPG 算法的吞吐量低于 Epsilon-greedy 算法，是因为基于单智体的 DDPG 算法没有充分考虑不同基站之间的相互影响关系，只从自身的局部最优出发作为优化决策，因此算法性能劣于基于贪婪策略，即从全局最优出发的 Epsilon-greedy 算法。

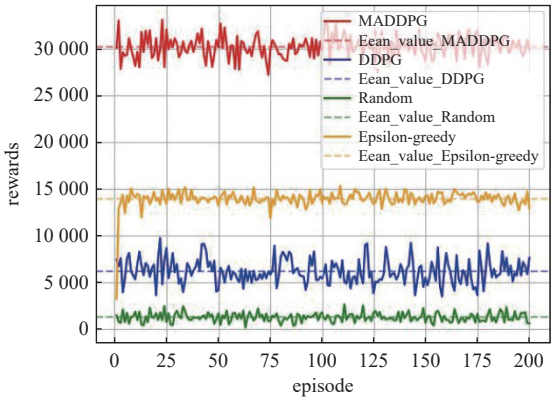

图 5 系统吞吐量对比图

图 6 给出了 4 种算法用户覆盖率的仿真比较结果。如图中所示，MADDPG 算法的用户覆盖率约为 55%，分别比 Epsilon-greedy 算法、Random 算法及 DDPG 算法高出 12%、13% 及 26%。与图 5 的结果类似，MADDPG 算法的用户覆盖率总是优于其他 3 种算法。

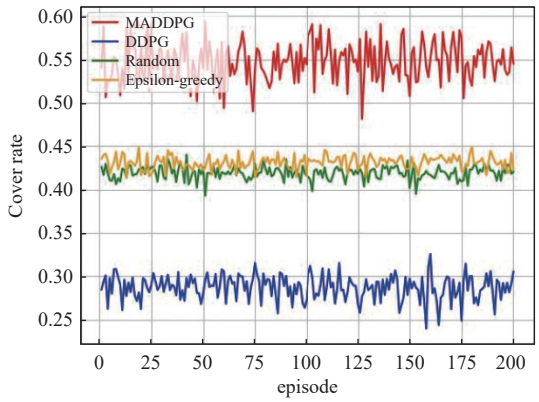

图 6 系统用户覆盖率对比图

分析用户数量对网络的影响，图 7 为用户数量分别为 60、80、100、120 时系统吞吐量均值的改变情况。而当用户数量逐渐增多，系统吞吐量则会逐渐上升。在不同网络规模下，MADDPG 始终取得最好的传输性能，并且在本研究场景下系统用户数量在 60～120 时，系统吞吐量相较于对比算法提升 2～3 倍。

图 8 为用户数量分别为 60、80、100、120 时系统用户覆盖率的变化情况。随着用户的增多，由于网络资源有限，算法无法同时满足所有用户的吞吐量需求，导致系统率越来越低，而 MADDPG 算

法在不同用户数量情况下，始终保持着较高的覆盖率。这是因为 MADDPG 不仅考虑了网络整体的长期吞吐量，还考虑了智能体之间的合作与竞争，能够更加高效地使用有限的网络资源来尽量满足更多用户的传输服务需求。相较于对比算法，在 60～120 的系统用户数量上能够在用户覆盖率上提升 20%～40%。

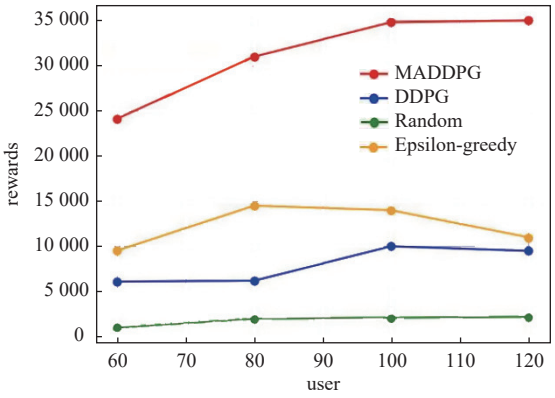

图 7　不同用户数量下系统奖励对比

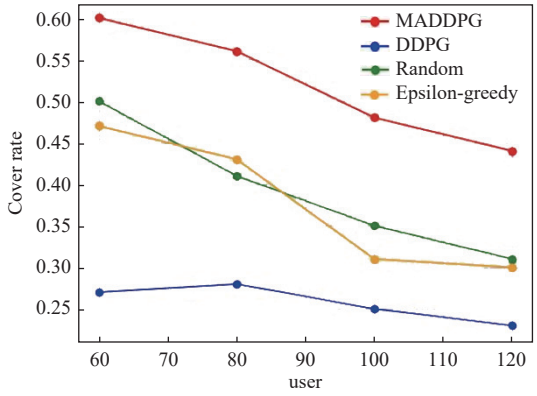

图 8　不同用户数量下系统用户覆盖率

综上所述，本文提出的基于 MADDPG 的波束赋形训练算法能够使系统达到更优的吞吐量以及用户覆盖率，优于基于单智体的 DDPG 算法、Epsilon-Greedy 算法和 Random 算法。

5　结束语

本文根据用户位置信息和信道情况，结合多智能体深度强化学习 MADDPG，设计了一种结合波束赋形在内的多小区 NOMA 资源分配策略算法。仿真结果表明，本文算法相较于智能算法 DDPG，以及非智能算法 Epsilon-Greedy 及 Random，能够合理利用获取的信息做出综合的资源分配决策，达到较好的系统吞吐量以及用户覆盖率。当系统用户数量在 60～120 时，系统吞吐量相较于对比算法提升 2～3 倍，用户覆盖率上提升 20%～40%。

参 考 文 献

[1] WANG X, MEI J, CUI S, et al. Realizing 6G: The operational goals, enabling technologies of future networks, and value-oriented intelligent multi-dimensional multiple access[J]. IEEE Network, 2023, 37(1): 10-17.

[2] ZHANG H, ZHANG Y, LIU X, et al. Resource allocation and mobility management for perceptive mobile networks in 6G[J]. IEEE Wireless Communications, 2024, 31(4): 223-229.

[3] WANG Z, ZHANG J, DU H, et al. A tutorial on extremely large-scale MIMO for 6G: Fundamentals, signal processing, and applications[J]. IEEE Communications Surveys & Tutorials, 2024, 26(3): 1560-1605.

[4] 刘承鹏, 张蔺, 陈智, 等. 下行协作 NOMA 系统中断概率分析与优化[J]. 电子科技大学学报, 2022, 51(5): 675-680.
LIU C P, ZHANG L, CHEN Z, et al. Outage probability analysis and optimization in downlink cooperative NOMA system[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(5): 675-680.

[5] SHIN W, VAEZI M, LEE B, et al. Non-orthogonal multiple access in multi-cell networks: Theory, performance, and practical challenges[J]. IEEE Communications Magazine, 2017, 55(10): 176-183.

[6] 胡浪涛, 毕松姣, 刘全金, 等. 基于深度强化学习的多小区 NOMA 能效优化功率分配算法[J]. 电子科技大学学报, 2022, 51(3): 384-391.
HU L T, BI S J, LIU Q J, et al. Multi-cell NOMA energy efficiency optimization power allocation algorithm based on deep reinforcement learning[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(3): 384-391.

[7] CHOI J. Non-orthogonal multiple access in downlink coordinated two-point systems[J]. IEEE Communications Letters, 2014, 18(2): 313-316.

[8] TIAN Y, NIX A R, BEACH M. On the performance of opportunistic NOMA in downlink CoMP networks[J]. IEEE Communications Letters, 2016, 20(5): 998-1001.

[9] HU Z, HAN C, DENG Y, et al. Multi-task deep reinforcement learning for Terahertz NOMA resource allocation with hybrid discrete and continuous actions[J]. IEEE Transactions on Vehicular Technology, 2024, 73(8): 11647-11663.

[10] BEYLERIAN A, OHTSUKI T. Coordinated non-orthogonal multiple access (CO-NOMA)[C]//2016 IEEE Global Communications Conference (GLOBECOM).

Washington, DC: IEEE, 2016: 1-5.

[11] LEE D, SEO H, CLERCKX B, et al. Coordinated multipoint transmission and reception in LTE-advanced: Deployment scenarios and operational challenges[J]. IEEE Communications Magazine, 2012, 50(2): 148-155.

[12] ISWARYA N, VENKATESWARI R. Deep reinforcement learning based resource allocation in NOMA[C]//2022 International Conference on Intelligent Innovations in Engineering and Technology (ICIIET). Coimbatore: IEEE, 2022: 286-293.

[13] LIU M, MENG Y, ZHANG Z. Deep reinforcement learning for resource allocation with mixed traffic in NOMA system[C]//2023 IEEE/CIC International Conference on Communications in China (ICCC). Dalian: IEEE, 2023: 1-2.

[14] CAO Y M, ZHANG G M, LI G B, et al. Joint resource allocation scheme based multi-agent DQN for massive MIMO-NOMA systems[C]//2022 14th International Conference on Communication Software and Networks (ICCSN). Chongqing: IEEE, 2022: 51-55.

[15] SAITO Y, BENJEBBOUR A, KISHIYAMA Y, et al. System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)[C]//2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC). London: IEEE, 2013: 611-615.

[16] DING Z, YANG Z, FAN P, et al. On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users[J]. IEEE Signal Processing Letters, 2014, 21(12): 1501-1505.

[17] BENJEBBOUR A, LI A, SAITO Y, et al. System-level performance of downlink NOMA for future LTE enhancements[C]//2013 IEEE Globecom Workshops (GC Wkshps). Atlanta: IEEE, 2013: 66-70.

[18] HANIF M F, DING Z, RATNARAJAH T, et al. A minorization-maximization method for optimizing sum rate in the downlink of non-orthogonal multiple access systems[J]. IEEE Transactions on Signal Processing, 2015, 64(1): 76-88.

[19] HAN J, LEE J G, KAMBER M. An overview of clustering methods in geographic data analysis[J]. Geographic Data Mining and Knowledge Discovery, 2009, 2: 149-170.

编 辑 刘飞阳