基于深度强化学习的多小区 NOMA 能效优化功率分配算法

胡浪涛*，毕松姣，刘全金，吴建岚，杨 瑞

(安庆师范大学电子工程与智能制造学院 安徽 安庆 246133)

【摘要】在下行多小区非正交多址接入系统中，功率分配是决定系统性能的关键因素之一。由于多小区系统间的功率优化问题的非凸性，获得最优功率分配在求解上非常困难。为此提出了一种基于深度强化学习最大化能效的功率分配算法，将深度 Q 网络作为动作–状态值函数，将系统能效直接设置为奖励函数，优化信道功率分配，使系统能量效率最大化。仿真结果表明，该算法比加权最小均方误差、分式规划、最大功率和随机功率算法等能够获得更高的系统能量效率，在算法计算复杂度、收敛速度和稳定性方面也有较好表现。

关 键 词 深度 Q 网络； 能量效率； 非正交多址接入； 功率分配； 强化学习

中图分类号 TN911.22 **文献标志码** A doi:10.12178/1001-0548.2021193

Multi-Cell NOMA Energy Efficiency Optimization Power Allocation Algorithm Based on Deep Reinforcement Learning

HU Langtao*, BI Songjiao, LIU Quanjin, WU Jianlan, and YANG Rui

(School of Electronic Engineering and Intelligent Manufacturing, Anqing Normal University Anqing Anhui 246133)

Abstract In a downlink multi-cell non-orthogonal multiple access system, power allocation is one of the key factors to determine system performance. Due to the non-convexity of the power optimization problem among multi-cell systems, it is very difficult to obtain the optimal power allocation. The power allocation algorithm based on deep reinforcement learning is proposed to maximize energy efficiency in this paper, which is simple and efficient. The algorithm takes the deep Q network as the action-state value function, system energy efficiency is directly set as a reward function, which optimizes channel power allocation and maximizes system energy efficiency. The simulation results show that the algorithm of proposed scheme is more effective than the weighted minimum mean square error, fractional programming, maximum power and random power algorithms in achieving higher system energy efficiency. The scheme also has better performances in algorithm calculation complexity, convergence speed and stability.

Key words deep Q network; energy efficiency; non-orthogonal multiple access; power allocation; reinforce learning

近年来，随着移动用户数量的爆炸式增长，多小区间的功率分配问题引起了广泛关注。此外，小区内和小区间的干扰管理对于提高蜂窝网络系统的能量效率也很重要。为了解决移动用户密度大的问题，非正交多址接入技术成为当前通信系统的研究热点之一[1-5]。

非正交多址接入 (non-orthogonal multiple access，NOMA) 技术的基本思想是在发送端采用非正交方式发送信号，在接收端采用串行干扰删除技术，从而实现信号的正确解调。已有很多文献研究了 NOMA 系统的功率分配问题。文献 [1] 提出一种单输入单输出情况下的优化问题，随后将单输入单输出解决方案扩展为多输入多输出场景，在满足每个用户的最小速率要求的服务质量和总功率约束条件下使总容量最大化。文献 [2] 将深度强化学习 (deep reinforce learning, DRL) 应用于无授权 NOMA 系统的决策中，旨在减轻冲突并提高未知网络环境中的系统吞吐量。文献 [3] 研究了包含任意用户的

收稿日期：2021－07－21；修回日期：2021－12－01
基金项目：国家自然科学基金 (61603003，62171002)；安徽省教育厅自然科学基金 (KJ2019A0554)
作者简介：胡浪涛 (1982－)，男，博士，副教授，主要从事无线通信中的信号处理和机器学习方面的研究.
*通信作者：胡浪涛，E-mail：hulangtao@aqnu.edu.cn

单个 NOMA 簇，目标是在满足每个用户所需的最小数据速率下最大化能量效率。文献 [4] 研究了集群中多用户多输入多输出 NOMA 系统中最大化能量效率的功率分配方案。

很多功率优化函数是非凸的，且优化问题是非确定性多项式 (non-deterministic polynomial, NP) 难题，机器学习技术被引入用于解决功率优化问题。机器学习包括监督学习、非监督学习和强化学习等。监督学习需要训练样本带有类别标签，通过训练深度神经网络逼近已给出的标签，文献 [6-7] 给出了关于监督学习的实现方案。无监督学习的训练样本没有标签，文献 [8-9] 相继提出了多种无监督学习研究方案。强化学习讨论一个智能体如何在未知环境里面最大化能获得的奖励。因为监督学习需要提前给出类别标签，非监督学习在学习过程中无反馈，强化学习在近年来成为无线通信中功率分配的热门技术[10-14]。文献 [10] 将 Actor-critic 算法应用于 NOMA 系统中不同认知无线电之间的功率分配，其目的是满足认知无线电最小数据速率要求的同时，最大化系统能量效率。文献 [11] 研究使用深度 Q 网络 (deep Q networks, DQN) 算法，旨在最大化整个网络的能量效率。文献 [12] 考虑动态无线网络中发射功率和信道的联合决策优化问题，通过构造 DQN 解决状态空间过大的复杂决策问题，提高系统能量效率。文献 [13] 提出基于 Actor-Critic 算法研究混合能源异构网络中用户调度和资源分配的最优策略，目的是最大化系统的能量效率。

本文针对单输入单输出的下行多小区 NOMA 系统，研究了一种 DRL 的功率分配算法 (energy efficient power allocation-DQN，EEPA-DQN)，将 DQN 作为动作–状态值函数，目的是优化信道功率分配，使系统能量效率最大化。将基站到用户的单个信道视为一个智能体，使用经验回放池将数据进行集中训练，分步执行时使用该智能体学习到的策略。仿真结果表明，EEPA-DQN 算法与加权最小均方误差 (weight minimum mean square error, WMMSE)[15]、分式规划 (fractional programming, FP)[16]、最大功率 (maximal power, MP)[17] 和随机功率 (random power, RP)[18] 等算法相比，得到的能量效率更高，收敛速度更快。

1　下行多小区 NOMA 系统模型

考虑多小区下行非正交多址接入系统，每个小区的中心设置有一个基站 (BS)。基站通过信道向用户发送信号，小区和用户的索引集分别为 $c=\{1,2,\cdots,C\}$和$k=\{1,2,\cdots,K\}$。小区 c 的第 k 个用户表示为$U_{c,k}$，每个小区包含两个用户。在时隙 t，小区 c 基站和用户 k 之间的信道增益描述为：

$$g_{c,k}^{t}=\beta_{c,k}\left|h_{c,k}^{t}\right|^{2} \tag{1}$$

式中，$h_{c,k}^{t}$是小尺度衰落；$\beta_{c,k}$是大尺度衰落。蜂窝网络系统模型如图 1 所示，系统包含 C 个小区。假设用户$U_{c,1}$为近端用户，用户$U_{c,2}$为远端用户，基站到$U_{c,1}$和$U_{c,2}$的信道增益分别为$g_{c,1}$和$g_{c,2}$，则$\left|g_{c,1}\right|^{2}>\left|g_{c,2}\right|^{2}$。小区 c 基站的发射功率为p_c，为$U_{c,1}$和$U_{c,2}$分配的功率分别为$p_{c,1}$和$p_{c,2}$，$p_c=p_{c,1}+p_{c,2}$，$U_{c,1}$为近端用户，且$p_{c,1}<p_{c,2}$。

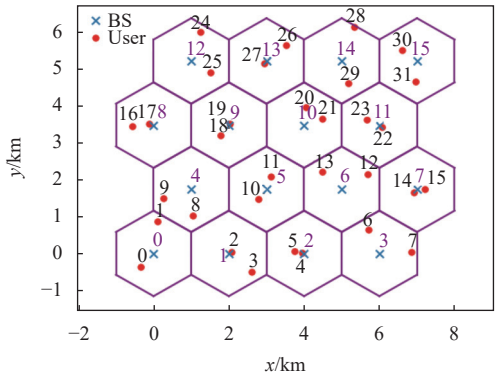

图 1　蜂窝网络模型

基站向不同用户发送消息，每个基站发送给用户的叠加信号表示为：

$$x_c=\sum_{k=1}^{K}\sqrt{p_{c,k}}s_{c,k} \tag{2}$$

式中，$p_{c,k}$和$s_{c,k}$分别为基站 c 中用户 k 的发送功率和信号。用户$U_{c,k}$的接收信号$y_{c,k}$表示为：

$$y_{c,k}=g_{c,k}x_c+n_{c,k} \tag{3}$$

式中，$n_{c,k}$是$U_{c,k}$接收到的高斯白噪声，均值为 0，方差为σ^2。

在下行多小区 NOMA 系统中，接收机使用串行干扰消除 (successive interference cancellation, SIC) 技术消除用户间干扰。已知$U_{c,1}$是近端用户，$U_{c,2}$是远端用户，故$U_{c,1}$首先解调出来自$U_{c,2}$干扰信号，并用 SIC 技术消除$U_{c,2}$带来的干扰；再解调出自身的信号。$U_{c,2}$可以直接解调出所接收的信号。在接收端，由 SIC 技术可得到$U_{c,1}$和$U_{c,2}$的信干扰比分别为：

$$\sin r_{c,1}^t = \frac{g_{c,1}^t p_{c,1}^t}{\sum_{c'\in D_c} g_{c',k}^t \sum_k p_{c',k}^t + \sigma^2} \tag{4}$$

$$\sin r_{c,2}^t = \frac{g_{c,2}^t p_{c,2}^t}{g_{c,2}^t p_{c,1}^t + \sum_{c'\in M_c} g_{c',k}^t \sum_k p_{c',k}^t + \sigma^2} \tag{5}$$

式中，M_c是干扰小区集；p 是基站的发射功率；$g_{c,2}^t p_{c,1}^t$为小区 c 内$U_{c,1}$对$U_{c,2}$的干扰；$\sum_{c'\in D_c} g_{c',k}^t \sum_k p_{c',k}^t$为小区间用户的干扰；$\sigma^2$表示附加噪声功率。

该链路的下行链路和速率根据远、近端用户可分别表示为$R_{c,1}$、$R_{c,2}$：

$$R_{c,1}^t = B\log_2(1+\sin r_{c,1}^t) \tag{6}$$

$$R_{c,2}^t = B\log_2(1+\sin r_{c,2}^t) \tag{7}$$

式中，B 为信道带宽。单个小区的和速率为两个用户和速率之和，定义为：

$$R_c^t = R_{c,1}^t + R_{c,2}^t \tag{8}$$

单个小区的能量效率η_c^t定义为：

$$\eta_c^t = \frac{R_c^t}{p_c^t + N_0} \tag{9}$$

式中，p_c^t为小区 c 在时隙 t 的发射功率；N_0为电路固定损耗。

NOMA 系统的能量效率定义为：

$$\eta^t = \sum_{c=1}^{C} \eta_c^t \tag{10}$$

优化目标是在最大发射功率约束下最大化系统的能量效率，描述如下：

$$\max_p \eta^t \tag{11}$$

$$\text{s.t.}\quad 0 \leqslant p_c^t \leqslant P_{\max}$$

式中，$P_{\max}$为基站发射功率的最大值。上述优化问题是一个非凸的问题，很难用最优化的方法解决这个问题。根据强化学习的理论[19]，对于一个给定的马尔可夫随机过程，尤其当系统是动态时，深度强化学习可以找到最优的决策动作，解决这个功率分配的优化问题。深度神经网络可以近似为一个函数，可以利用神经网络建立起状态s^t到价值$Q_\pi(s^t,a^t)$的一个映射，状态值会从时隙t跳转到下一时隙 $(t+1)$，通过训练神经网络找到最大的$Q_\pi(s^t,a^t)$值对应的动作值a^t(基站的发射功率)。

2 EEPA-DQN 算法设计

2.1 深度 Q 网络简介

强化学习算法讨论一个智能体如何在一个复杂不确定的环境里获得最大化的奖励。本文采用深度强化学习 DQN 算法，基于离散时间马尔可夫决策过程 (Markov decision process, MDP)，在有限的动作和状态空间中最大化获得的奖励。在时隙 t，从环境中获取状态$s^t \in S$，智能体选择动作$a^t \in A$，并与环境交互，获得奖励$r^t \in R$并转换到下一个状态s^{t+1}，其中，A 是动作集合，S 是状态集合，P 是当前状态转移到下一个状态的状态转移概率，R 是奖励集合。强化学习框图如图 2 所示。

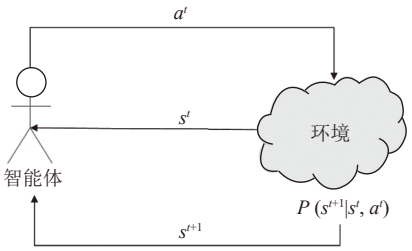

图 2 强化学习模型

由于状态可以是连续的，DQN 将 Q-learning 与神经网络相结合，用于解决无限状态空间的问题，即用神经网络代替 q-table，并在 q-table 的基础上提出两个创新点[14]。

1) 经验回放。由于 Q-learning 算法得到的样本前后是相关的，为了打破数据之间的关联性，在网络训练过程中使用经验回放机制。从以往的状态转移中随机采样(s^t,a^t,r^t,s^{t+1})进行训练。经验回放可以减少智能体所需的学习经验，解决样本关联性和效率利用的问题。

2) 固定 Q 目标。固定 Q 目标主要解决算法不稳定的问题，当前值网络复制得到目标值网络，起初两个网络具有相同的参数。随着训练次数的增加，当前值网络参数 θ 实时更新，每隔 N 步，当前值网络的参数复制给目标值网络，进行目标值网络的更新，通过当前值网络和目标值网络计算当前 Q 值和目标Q^{target}值，通过最小化当前 Q 值和目标Q^{target}值之间的均方误差优化更新网络参数。引入目标值网络后，在一段时间内目标Q^{target}值是保持不变的，从而降低了当前 Q 值和目标Q^{target}值之间的相关性，提升了算法的稳定性。

采用$Q_\pi(s^t,a^t)$函数作为状态−动作值函数。在时隙 t，给定状态s^t，选择动作a^t，即时奖励为r^t，

Q 值函数为：

$$Q_\pi(s^t,a^t)=\mathrm{E}_\pi[R^t|s^t=s,a^t=a]$$
$$R^t=\sum_{\tau=0}^{\infty}\gamma^\tau r^{t+\tau+1} \tag{12}$$

式中，$\gamma\in[0,1)$是平衡当前奖励和未来奖励的折扣因子；E[·]是期望算子。Q 函数是一个度量指标，用来评估在政策 π 下所选动作a^t对学习过程获得的预期累积折扣奖励的影响。

Q 函数满足贝尔曼方程:

$$Q_\pi(s^t,a^t)=\mathrm{E}_\pi[r^{t+1}|s^t=s,a^t=a]+\gamma\sum_{s'\in S}P_{ss'}^a\left(\sum_{a'\in A}\pi(s',a')Q^\pi(s',a')\right) \tag{13}$$

式中，$P_{ss'}^a=P[s^{t+1}=s'|s^t=s,a^t=a]$表示智能体在状态 s 时，采取动作 a 到下一个状态s'的转移概率。

与最优策略相关的最优 Q 函数为：

$$Q^*(s^t,a^t)=r^{t+1}\ (s^t=s,a^t=a,\pi=\pi^*)+\gamma\sum_{s'\in S}P_{ss'}^a\max_{a'\in A}Q^*(s',a') \tag{14}$$

式中，π^*为 Q-learning 算法搜索最优策略。

由式 (14) 递归求解得到最优$Q^*(s^t,a^t)$值。对 Q 函数的更新为：

$$Q^*(s^t,a^t)\leftarrow(1-\alpha)Q^*(s^t,a^t)+\alpha(r^{t+1}+\gamma\max_{a'}Q_\pi(s^{t+1},a')) \tag{15}$$

式中，α是用以更新 Q 函数的学习率。在维度较大的状态空间和动作空间实时更新$Q(s^t,a^t)$难以实现，选用 DQL 与深度神经网络 (deep neural network，DNN) 结合，即 DQN。

DQN 中 Q 值函数由参数 θ 决定，θ 包括神经网络中的权重和偏差。与式 (15) 中直接更新 Q 函数不同，DQN 的参数 θ 更新如下：

$$\theta^{t+1}=\theta^t+\lambda(y^t-Q(s^t,a^t;\theta^t))\nabla Q(s^t,a^t;\theta^t) \tag{16}$$

式中，λ 代表学习率，用以更新网络参数，使得当前$Q(s^t,a^t;\theta^t)$值逐渐朝向目标值。智能体从经验重放池中随机采样批量数据(s^t,a^t,r^t,s^{t+1})来训练 DQN 网络，DQN 的训练流程如图 3 所示。

本文研究在最大发射功率约束下，最大化系统的能量效率。令 γ=0，式 (12) 可描述为$\max Q=\max_{a\in A}\mathrm{E}_\pi[r^t|s^t=s,a^t=a]$。针对功率分配问题，$a=p^t$，$s=g^t$，奖励函数设置为$r^t=\eta^t$，最大 Q 值可表示为：

$$\max Q=\max_{0\leqslant p_c^t\leqslant P_{\max}}\mathrm{E}_\pi[\eta^t|g^t,p^t] \tag{17}$$

选择最优 Q 值，得到最佳动作a^*:

$$a^*=\arg\max_a Q(s,a;\theta_q) \tag{18}$$

训练过程中采用动态的 ε-greedy 策略控制探测概率[20]，定义为：

$$\varepsilon_k=\varepsilon_1-\frac{k-1}{N_e-1}(\varepsilon_1-\varepsilon_k)\ \ k=1,2,\cdots,N_e \tag{19}$$

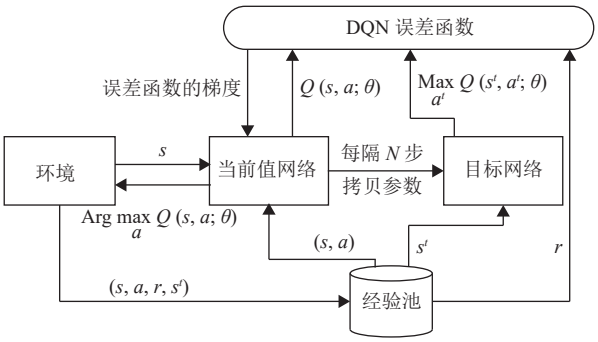

图 3　DQN 训练流程

2.2　基于 DQN 的下行多小区 NOMA 系统设计

本文使用免模型两步训练框架，由于数据驱动算法对数据量要求较高，为了减少在线训练的压力，使用 DRL 算法对 DQN 进行离线训练；再将训练过的 DQN 在真实场景中进行动态微调。基站到用户的下行链路信道可视为一个智能体，环境是下行多小区 NOMA 系统，智能体和环境进行交互，智能体选择一个动作a^t，得到一个奖励r^t，进入下一个状态s^{t+1}。下行多小区 NOMA系统研究的是一个多智能体问题，训练数据及参数较单智能体更为复杂。故引入经验回放技术，经验回放池中包括状态s^t、动作a^t、奖励r^t和下一个状态s^{t+1}等数据，利用经验回放池数据对 DQN 网络进行集中训练，分步执行时使用该智能体学习到的策略。

本文将 DQN 的思想引入 NOMA 系统的功率分配中，即 EEPA-DQN 算法，旨在最大化系统的能量效率。EEPA-DQN 的 3 个重要组成元素为状态、动作和奖励，具体如下。

状态：状态的选取很重要，为了降低输入维度，在时隙 t 开始时，智能体根据来自接收机处干扰源的当前接收功率对干扰源按从大到小进行排序。保留前 Z 个对用户 k 下一个动作有较强干扰的信息源，Z 以外的基站到用户的下行链路及干扰信号的信道增益均视为零。最佳发射功率p^t和当前的信道增益g^t相关，但这种设计使得 DQN 的性能变差。因此，本文基于文献 [21]，通过 3 个主要特征

描述智能体的状态，即对数归一化干扰集$I_{c,k}^{t}$、前一时隙的奖励η^{t-1}和前一时隙的发射功率p^{t-1}，表示为状态$s_{c,k}^{t}=\{I_{c,k}^{t},\eta_{c,k}^{t-1},p_{c,k}^{t-1}\}$。

动作：基站的发射功率作为智能体的动作。NOMA 系统中，基站到用户的发射功率是一个连续的变量，受最大功率的限制，而 DQN 只能处理离散的动作空间，所以需要将其量化，0 和$P_{\max}$之间发射功率被量化为$|A|$个数量级，量化后的动作为：

$$A\in\{0,P_{\min},\cdots,P_{\max}\} \tag{20}$$

式中，$P_{\min}$表示非零发射功率最小值；$P_{\max}$表示发射功率最大值，最大值与最小值之间的功率为等间隔增大的$|A-3|$个动作。在时隙 t，每个智能体只允许选择一个动作$a^{t}\in A$更新它的功率策略。

奖励：奖励函数的设计决定强化学习算法的收敛速度和程度，智能体目的是最大化的系统的累计收益，若想要让智能体较快的达到目标，提供奖励函数应使智能体在最大化收益的同时可实现系统能量效率最大化。故本文中将系统能量效率用作奖励函数。

算法：EEPA-DQN 算法

输入：状态，状态包含对数归一化干扰集$I_{c,k}^{t}$、前一时隙的奖励η^{t-1}和前一时隙的发射功率p^{t-1}

输出：动作，动作为功率

初始化：回合数为 M，每个回合时隙数为 T，学习率为 λ，探索率为ε，初始化经验池 D，批样本数量为 N，随机初始化 EEPA-DQN 中参数 θ

for m=1: M

初始化状态s^{1}

for t=1: T

以ε的概率随机选取动作a^{t}，否则选择最优的动作$a^{t}=\arg\max_{a}Q(s,a;\theta_{q})$；

执行动作a^{t}，得到奖励r^{t}和下一个状态s^{t+1}；

将$(s^{t},a^{t},r^{t},s^{t+1})$放入 D 中；

以$(y^{t}-Q(s^{t},a^{t};\theta^{t}))^{2}$为损失函数训练 EEPA-DQN；

$s^{t}\leftarrow s^{t+1}$；

更新 θ；

end for

end for

3 下行多小区 NOMA 系统仿真

3.1 下行多小区 NOMA 系统参数设置

本文研究下行多小区 NOMA 系统，模拟一个小区数 C=16 的蜂窝网络，在每一个小区内配备一个中心基站，每个基站可同时为 K=2 个用户服务。假设某一小区的两层之内的小区设置为干扰用户，即干扰层数 I=2；用户被随机分配在$d\in[r_{\min},r_{\max}]$内，$r_{\min}=0.01$ km和$r_{\max}=1$ km分别为小区内基站到用户最短距离和最长距离。信道模拟小尺度衰落，小尺度衰落服从独立的瑞利分布，使用 Jakes 模型，路径损耗以$\beta=120.9+37.6\lg d+10\lg z$进行模拟，$d$ 是基站与用户之间的距离，距离越大路径损耗值越大，z 为对数正态随机变量，标准差为 8 dB[20]。

为确保智能体能快速做出决策，网络结构不宜过于复杂，EEPA-DQN 算法为一个输入层、两个隐藏层和一个输出层的结构较简单的神经网络。隐藏层采用 ReLU 激活函数，输出层的激活函数是线性的。将前 12 个小区视为干扰源，功率电平数$|A|=10$。为了减少在线计算的压力，采用离线训练。在前 100 次迭代训练中，只能随机选择动作，在探索阶段使用自适应贪婪策略[22]。训练得到的 EEPA-DQN 具有较强的泛化能力，每次迭代包含 1 000 个时隙，每 10 个时隙从经验回放记忆中随机抽取一批样本训练 EEPA-DQN，使用 Adam[23] 算法作为优化器，NOMA 无线通信系统参数设置见表 1。

表 1 NOMA 无线通信系统参数设置

参数	值
折扣因子γ	0
经验池大小D	50 000
批样本数量N	256
学习率λ	0.000 1
初始探索率ε_1	0.25
最终探索率$\varepsilon_{\mathrm{Ne}}$	0.000 1
隐藏层	128(1), 64(2)
回合数M	20 000
信道带宽B/Hz	1
每个回合时隙数T	1 000
最小发射功率$P_{\min}$/dBm	5
最大发射功率$P_{\max}$/dBm	38
最大多普勒频率f_p/Hz	10
电路固定损耗功率N_0/dBm	30
加性高斯白噪声功率σ^2/dBm	−114

3.2 功率分配算法比较

在对 EEPA-DQN 算法进行实验仿真的同时，

将本文提出的 EEPA-DQN 算法与 FP、WMMSE、MP 和 RP 算法进行实验比较。FP、WMMSE 这两个算法是非常经典的考虑多小区间干扰的功率分配算法，均为迭代的算法，都需要全局实时的跨小区信道状态信息 (channel state information, CSI)，对于基站来说它的开销庞大[24]。深度神经网络具有一定的学习本领，在进行网络的特征提取时具有一定的智能和泛化性能。另一个优点是 DQN 的算法复杂度较低。表 2 列出了不同算法的单次 CPU 运行时间。从表 2中可以看出基于强化学习的功率分配算法复杂度较低。EEPA-DQN 算法分别比 FP、WMMSE、MP 和 RP 算法快 13.0 倍、14.1 倍、15.1 倍和 13.0 倍左右，硬件配置为：Intel(R) Xeon(R) CPU E3-1230 v5；软件为：python 3.7，TensorFlow 1.15.0。仿真的下行多小区 NOMA 系统小区数目为 16。

表 2　单次执行时间

算法	时间/s
EEPA-DQN	9.9×10^{-4}
FP	1.296×10^{-2}
WMMSE	1.396×10^{-2}
MP	1.496×10^{-2}
RP	1.292×10^{-2}

不过，EEPA-DQN 算法计算复杂度与神经网络的层数呈线性关系，且随着维数的增加，计算变得复杂。图 4 展示了 EEPA-DQN 算法得到的平均能效比 FP、WMMSE、MP 和 RP 分配算法有显著提高。因此，EEPA-DQN 算法可有效地最大化系统的能量效率。

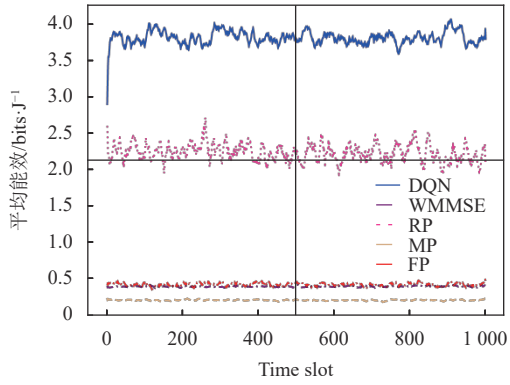

图 4　5 种功率分配算法平均能量效率

NOMA 是非正交多址技术，OMA 代表传统的正交多址。当多个用户的信号在相同的信道资源上传输时，NOMA 可以实现更高的频谱效率[25]、更大的系统容量和低传输延迟[26]。从图 5 中可以看出，随着迭代次数的增加，两种多址方案的系统平均能量效率都增加了。NOMA 的功率分配与接收端处 SIC 过程相关，将较高的功率分配给路径损耗较大的用户，提高了用户的速率，使 NOMA 系统比 OMA 系统可实现更大的系统平均能量效率，且算法更为稳定。

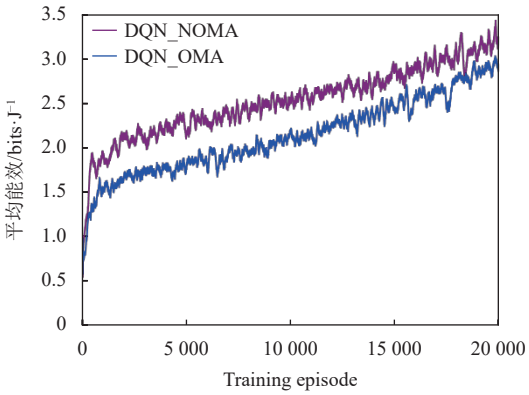

图 5　NOMA 与 OMA 平均能量效率

3.3　折扣因子选择

折扣因子是可选择的一个经验值，对于大多数应用而言，增加 γ 更有利于 DQN。本文算法取 $\gamma=0,0.2,0.4,0.6,0.8$，仿真结果表明，$\gamma=0$时能效值明显高于其他值，但考虑到信号传输过程中存在路径损耗，智能体与未来回报之间的相关性相对较少，γ 应选取较小的值。图 6 仿真了不同 γ 值时，EEPA-DQN 训练过程中的下行多小区 NOMA 系统平均能量效率，随着训练次数的增加，平均能量效率逐渐增加，且在$\gamma=0$时达到最高能效。图 7 仿真了不同 γ 值在不同小区数时，EEPA-DQN 训练过程中的平均能效。仿真实验考虑了小区数 C=9, 16, 36, 64 的情况，通过图 7 可知，这 4 种情况下小区数为 9 时所能达到的能效最高，目标小区周围的干扰小区数目越多，外围到目标小区距离越大，干扰会越来越小，所以最外围的干扰小区的干扰功率就非常小。最后仿真了不同小区数目的 NOMA 系统的能效。由式 (4)、式 (5) 可知，随着小区数的增加，如小区数为 36、64 时，小区间的干扰随之增强，所达到的能效随着小区数量的增加而下降，γ=0 时仍能保持较高的能效，从而验证了本文算法在 γ=0 时有一定的泛化性能。

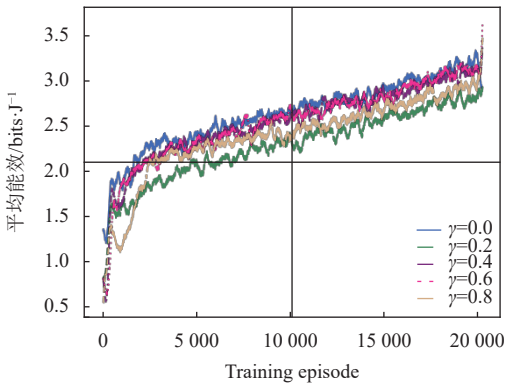

图 6 不同 γ 值时系统平均能量效率

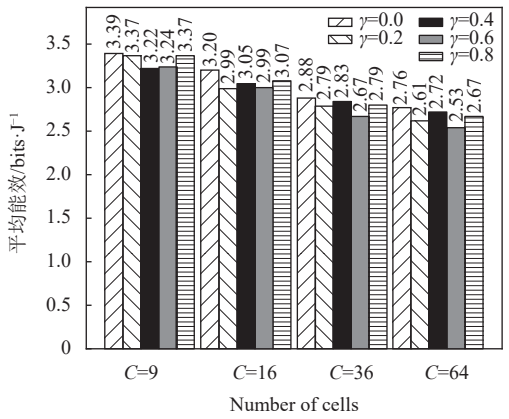

图 7 不同 γ 值不同小区数时平均能量效率

3.4 学习率

通过实验评估不同学习率对 EEPA-DQN 算法的影响。图 8 展示不同学习率下的平均能量效率与训练回合的关系，学习率 Ir=0.01, 0.001, 0.000 1 这 3 种情况，平均能效均有上升趋势。当学习率设置为 0.000 1 时，算法相对于其他两个取值更为稳定，且平均能效可达到最高；当学习率为 0.01 时，可观察到算法稳定性较差。通过以上分析，EEPA-DQN 算法的学习率设置为 0.000 1。

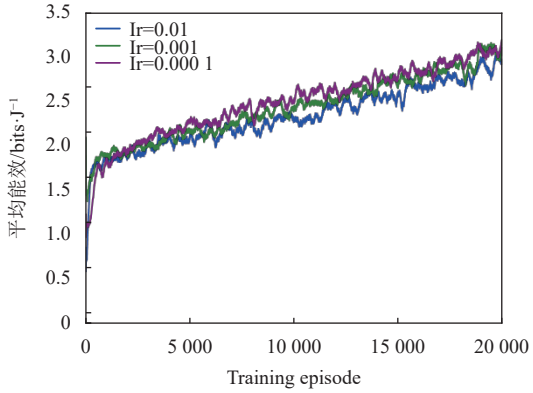

图 8 不同学习率值时平均能量效率

4 结 束 语

本文研究了一种基于强化学习的下行多小区 NOMA 系统的功率分配问题，旨在最大化系统的能效。由于功率优化问题具有非凸性，本文选用免模型驱动的 DQL 算法，将 DQL 与神经网络相结合以解决状态连续的问题。仿真结果表明，本文算法将含有两个隐藏层的 EEPA-DQN 逼近动作−值函数，同时，本文算法扩展到大规模场景也有较好的性能，但算法的稳定性还有待提高。

参 考 文 献

[1] WANG C L, CHEN J Y, CHEN Y J. Power allocation for a downlink non-orthogonal multiple access system[J]. IEEE Wireless Communications Letters, 2016, 5(5): 532-535.

[2] ZHANG J Z, TAO X F, WU H C, et al. Deep reinforcement learning for throughput improvement of the uplink grant-free NOMA system[J]. IEEE Internet of Things Journal, 2020, 7(7): 6369-6379.

[3] ZHANG Y, WANG H M, ZHENG T X, et al. Energy-efficient transmission design in non-orthogonal multiple access[J]. IEEE Transactions on Vehicular Technology, 2017, 66(3): 2852-2857.

[4] ZENG M, YADAV A, DOBRE O A, et al. Energy-efficient power allocation for MIMO-NOMA with multiple users in a cluster[J]. IEEE Access, 2018, 6: 5170-5181.

[5] 田心记, 黄玉霞, 李晓静. NOMA 系统中最大化能量效率的功率分配[J]. 电子科技大学学报, 2021, 50(1): 1-7.
TIAN X J, HUANG Y X, LI X J. Power allocation with maximizing energy efficiency for NOMA system[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(1): 1-7.

[6] YE H, LI G Y, JUANG B H, et al. Power of deep learning for channel estimation and signal detection in OFDM systems[J]. IEEE Wireless Communications Letters, 2018, 7(1): 114-117.

[7] JEON Y S, HONG S N, LEE N. Supervised-learning-aided communication framework for MIMO systems with low-resolution ADCs[J]. IEEE Transactions on Vehicular Technology, 2018, 67(8): 7299-7313.

[8] XUE S Y, MA Y, YI N, et al. Unsupervised deep learning for MU-SIMO joint transmitter and noncoherent receiver design[J]. IEEE Wireless Communications Letters, 2019, 8(1): 177-180.

[9] SHE C Y, SUN C J, GU Z Y, et al. A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning[J]. Proceedings of the IEEE, 2021, 109(3): 204-246.

[10] LIANG W, NG S X, SHI J, et al. Energy efficient transmission in underlay CR-NOMA networks enabled by reinforcement learning[J]. China Communications, 2020, 17(12): 66-79.

[11] SHI D P, TIAN F, WU S C. Energy efficiency optimization in heterogeneous networks based on deep

reinforcement learning[C]//2020 IEEE International Conference on Communications Workshops (ICC Workshops). Dublin: IEEE, 2020: 1-6.

[12] FAN H R, ZHU L, YAO C H, et al. Deep reinforcement learning for energy efficiency optimization in wireless networks[C]//2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). Chengdu: IEEE, 2019: 465-471.

[13] WEI Y F, YU F R, SONG M, et al. User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach[J]. IEEE Transactions on Wireless Communications, 2018, 17(1): 680-692.

[14] NASIR Y S, GUO D N. 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.

[15] SHI Q J, RAZAVIYAYN M, LUO Z Q, et al. An iteratively weighted mmse approach to distributed sum-utility maximization for a MIMO interfering broadcast channel[J]. IEEE Transactions on Signal Processing, 2011, 59(9): 4331-4340.

[16] SHEN K M, YU W. Fractional programming for communication systems—part I: Power control and beamfor-ming[J]. IEEE Transactions on Signal Processing, 2018, 66(10): 2616-2630.

[17] CHIANG M, HANDE P, LAN T, et al. Power control in wireless cellular networks[J]. Foundations and Trends in Networking, 2008, 2(4): 381-533.

[18] TAO L W, YANG W W, YAN S H, et al. Covert communication in downlink NOMA systems with random transmit power[J]. IEEE Wireless Communications Letters, 2020, 9(11): 2000-2004.

[19] SUTTON R S. 强化学习[M]. 北京: 电子工业出版社, 2018.
SUTTON R S. Reinforcement learning[M]. Beijing: Publishing House of Electronics Industry, 2018.

[20] 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.

[21] MENG F, CHEN P, WU L N. Power allocation in multi-user cellular networks with deep Q learning approach[C]//2019 IEEE International Conference on Communications (ICC). Shanghai: IEEE, 2019: 1-6.

[22] SUTTON S, BARTO A G. Reinforcement learning: An introduction[M]. Cambridge: MIT Press, 1998.

[23] KINGMA D P, BA J L. Adam: A method for stochastic optimization[EB/OL]. [2021-03-23]. https://arxiv.org/abs/1412.69-80.

[24] HUANG C W, MO R H, YUEN C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(8): 1839-1850.

[25] SAITO Y, KISHIYAMA Y, BENJEBOUR A, et al. Non-orthogonal multiple access (NOMA) for cellular future radio access[C]//2013 IEEE 77th Vehicular Technology Conference (VTC Spring). [S.l.]: IEEE, 2013: 1-5.

[26] WAN D H, WEN M W, JI F, et al. Non-orthogonal multiple access for cooperative communications: challenges, opportunities, and trends[J]. IEEE Wireless Communications, 2018, 25(2): 109-117.

编　辑　税　红