基于深度强化学习的IRS辅助认知无线电系统波束成形算法

李国权*① 程 涛① 郭永存① 庞 宇② 林金朝②
①(重庆邮电大学通信与信息工程学院 重庆 400065)
②(光电信息感测与微系统重庆市重点实验室 重庆 400065)

摘 要：为进一步提升多用户无线通信系统的频谱利用率，该文提出了一种基于深度强化学习的智能反射面(IRS)辅助认知无线电网络次用户和速率最大化算法。首先在考虑次基站最大发射功率约束、次基站对主用户的干扰容限约束以及IRS相移矩阵单位模量约束的情况下，建立一个联合优化次基站波束成形和IRS相移矩阵的资源分配模型；然后提出了一种基于深度确定性策略梯度的主被动波束成形算法，联合进行变量优化以最大化次用户和速率。仿真结果表明，所提算法相对于传统优化算法在和速率性能接近的情况下具有更低的时间复杂度。

关键词：智能反射面；认知无线电；深度强化学习；波束成形
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2025)03-0657-09
DOI: 10.11999/JEIT240447

1 引言

未来第6代无线通信中，移动设备用户数量将呈爆炸式增长，频谱资源稀缺问题日益严重[1]。认知无线电(Cognitive Radio, CR)技术可以在不影响授权用户通信的情况下共享授权频段的空闲频谱资源，达到更高效利用稀缺频谱资源的目的[2]。智能反射面(Intelligent Reflecting Surface, IRS)是由许多无源反射单元组成的平面阵列，其中每个单元都能够在微控制器的操纵下，独立地控制对反射信号的相移，从而降低传输路径的损耗来提高传输质量，而且其低功耗特性特别适用于密集部署，因此近年来受到了广泛的关注，可在无额外频谱资源的情况下进一步增加无线通信系统的信道容量[3–5]。通过将IRS和CR应用于传统的无线通信系统中，则有望实现更高水平的频谱效率增强。

文献[6]研究了认知无线电网络(CR Networks, CRN)中IRS辅助的单输入单输出(Single-Input Single-Output, SISO)系统的最优主动和被动波束成形问题，其中次用户(Secondary User, SU)和主用户(Primary User, PU)存在交叉链路干扰，利用交替优化(Alternating Optimization, AO)和连续凸近似联合优化次基站(Secondary Base Station, SBS)发射功率和IRS相移矩阵来实现SU速率最大化。文献[7]研究了利用非正交多址(Non-Orthogonal Multiple Access, NOMA)和CR形成NOMA-IRS辅助CRN的频谱效率方法。文献[8]研究了IRS辅助下行多输入单输出(Multiple-Input Single-Output, MISO) CRN中，在PU的信干噪比(Signal to Interference and Noise Ratio, SINR)约束下，使SU速率最大化，提出一种采用半正定松弛，基于块坐标下降法的高效迭代算法联合优化了SBS发射功率和IRS反射相移。文献[9]研究了CRN中IRS辅助无线携能通信的波束成形问题，利用交替方向乘子法和AO算法，联合优化基站和IRS的有源和无源波束成形。总体来说，现有研究主要利用凸优化理论或启发式算法[6–10]等方法解决优化问题，求解问题的计算复杂度和计算成本较高。

深度强化学习(Deep Reinforcement Learning, DRL)是一种结合深度学习和强化学习的算法，是解决策略控制的理想工具，已在无线通信资源分配及IRS辅助无线通信系统中取得了明显效果[11–15]。文献[11]研究了在单用户MISO无线通信系统中，利用DRL优化IRS相移来最大化用户信噪比。文献[12]研究了在一个IRS辅助的MISO系统中，基于DRL优化基站波束成形和IRS相移来最大化用户和速率。文献[13]提出了一种基于深度确定性策略梯度(Deep Deterministic Policy Gradient, DDPG)的算法[16]来增强无线网络的防窃听和抗干扰性能。

为进一步提升多用户无线通信系统的频谱利用率，本文提出了一个IRS辅助的多用户MISO-CR系统，利用IRS辅助共享频谱资源的情况下实现主基站(Primary Base Station, PBS)与PU以及SBS与SU之间的通信，并提出了一种基于DDPG的SU和速率最大化相关的主被动波束成形算法，联合优化

收稿日期：2024-06-04；改回日期：2025-02-17；网络出版：2025-02-26
*通信作者： 李国权 ligq@cqupt.edu.cn
基金项目：国家自然科学基金 (U21A20447)，重庆市自然科学基金创新群体科学基金(cstc2020jcyj-cxttX0002)
Foundation Items: The National Natural Science Foundation of China (U21A20447), The Foundation for Innovative Research Groups of the Natural Science Foundation of Chongqing (cstc2020jcyj-cxttX0002)

SBS发射波束成形和IRS相移，在较低计算复杂度和满足PU干扰容限约束的情况下，最大限度地提高SU和速率，最后进行了仿真验证。论文主要贡献如下：

(1)建立了IRS辅助MISO-CRN多用户系统的无线传输模型，构建了SU和速率优化问题，在满足SBS最大发射功率约束，IRS相移单位模量约束和PU干扰容限约束的条件下，使SU和速率最大化。

(2)为解决上述变量耦合的非凸优化问题，提出了一种基于DRL的SBS发射波束成形和IRS相移优化算法。与AO算法交替优化发射波束成形和IRS相移不同，该算法能够同时优化SBS发射波束成形和IRS相移且具有更低的计算复杂度。

(3)仿真结果表明所提方案性能明显优于无IRS和随机IRS相移的方案。与传统凸优化算法相比，其时间复杂度并未随着IRS反射单元数量的增加而增加。

2 系统模型和问题提出

考虑如图1所示的IRS辅助多用户MISO-CR系统下行链路，其中PBS和SBS均配置有M副天线，PU和K个SU均为单天线，IRS反射单元/移相器数量为N，其相移可通过微控制器调整。SBS处部署有管理中心或者与管理中心具有高速有线链路，可以实时计算实现资源分配优化。

假设所有信道均为准静态平坦衰落，且SBS可以通过导频信号进行信道估计[17]，同时与PBS之间进行信息交互，获得系统的所有信道状态信息(Channel State Information, CSI)。在主用户链路中，PBS与PU、PBS与IRS、IRS与PU之间的信道矩阵分别为$\boldsymbol{h}_{\mathrm{pp}}\in\mathbb{C}^{1\times M}$，$\boldsymbol{G}_{\mathrm{pr}}^{\mathrm{H}}\in\mathbb{C}^{N\times M}$，$\boldsymbol{h}_{\mathrm{rp}}\in\mathbb{C}^{1\times N}$；在次用户链路中，SBS与SU、SBS与IRS、IRS与SU之间的信道矩阵分别为$\boldsymbol{h}_{\mathrm{d},k}\in\mathbb{C}^{1\times M}$，$\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}}\in\mathbb{C}^{N\times M}$，$\boldsymbol{h}_{\mathrm{r},k}\in\mathbb{C}^{1\times N}$；在干扰链路中，PBS与SU、SBS与PU之间的干扰信道矩阵分别为$\boldsymbol{h}_{\mathrm{ps},k}\in\mathbb{C}^{1\times M}$，$\boldsymbol{h}_{\mathrm{dp}}\in\mathbb{C}^{1\times M}$，其中，$k\in[1,\cdots,k,\cdots,K]$。令$\boldsymbol{\Theta}=\mathrm{diag}\left(\theta_1,\cdots,\theta_n,\cdots,\theta_N\right)$表示IRS相移矩阵，其中$\theta_n=\mathrm{e}^{\mathrm{j}\varphi_n}$表示IRS上第$n$个反射单元的相移。只考虑IRS首次反射的信号，则PU和SU k分别接收到的信号y_{p},y_k可表示为

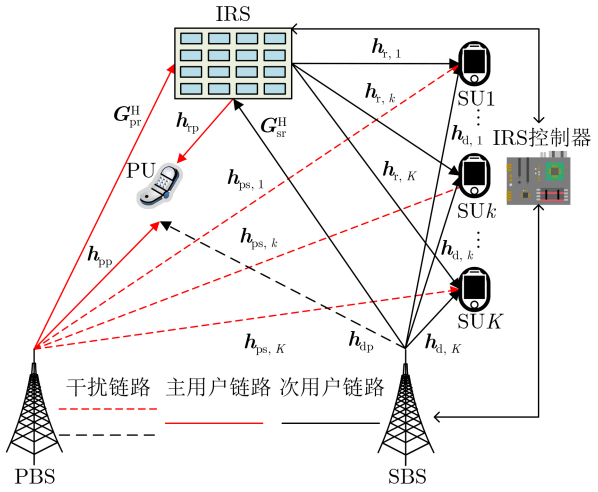

图 1 IRS辅助的认知无线电系统模型

$$y_{\mathrm{p}}=\underbrace{(\boldsymbol{h}_{\mathrm{pp}}+\boldsymbol{h}_{\mathrm{rp}}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{pr}}^{\mathrm{H}})\boldsymbol{w}_{\mathrm{p}}s_{\mathrm{p}}}_{\text{期望信号}}+\underbrace{(\boldsymbol{h}_{\mathrm{dp}}+\boldsymbol{h}_{\mathrm{rp}}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\sum_{k=1}^{K}\boldsymbol{w}_k s_k}_{\text{来自次基站的干扰}}+n_{\mathrm{p}} \tag{1}$$

$$y_k=\underbrace{(\boldsymbol{h}_{\mathrm{d},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\boldsymbol{w}_k s_k}_{\text{期望信号}}+\underbrace{(\boldsymbol{h}_{\mathrm{d},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\sum_{n\neq k}^{K}\boldsymbol{w}_n s_n}_{\text{多用户干扰}}+\underbrace{(\boldsymbol{h}_{\mathrm{ps},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{pr}}^{\mathrm{H}})\boldsymbol{w}_{\mathrm{p}}s_{\mathrm{p}}}_{\text{来自主基站的干扰}}+n_k \tag{2}$$

其中，$\boldsymbol{w}_{\mathrm{p}}\in\mathbb{C}^{M\times 1}$，$\boldsymbol{w}_k\in\mathbb{C}^{M\times 1}$分别表示PBS和SBS的波束成形向量，且$\boldsymbol{w}_{\mathrm{p}}=\sqrt{P_{\mathrm{PBS}}}\left(\boldsymbol{h}_{\mathrm{pp}}^{\mathrm{H}}/\|\boldsymbol{h}_{\mathrm{pp}}\|\right)$，$s_{\mathrm{p}}\in\mathbb{C}$和$s_k\in\mathbb{C}$均为零均值、单位方差的独立随机变量，分别表示PU和SU k的传输数据符号，$n_{\mathrm{p}}\sim\mathcal{CN}\left(0,\sigma_{\mathrm{p}}^2\right)$，$n_k\sim\mathcal{CN}\left(0,\sigma_k^2\right)$分别表示PU和SU k接收机处的加性高斯白噪声。因此，SU k的SINR可表示为

$$\gamma_k=\frac{\left|(\boldsymbol{h}_{\mathrm{d},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\boldsymbol{w}_k\right|^2}{\sum\limits_{n\neq k}^{K}\left|(\boldsymbol{h}_{\mathrm{d},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\boldsymbol{w}_n\right|^2+\left|(\boldsymbol{h}_{\mathrm{ps},k}+\boldsymbol{h}_{\mathrm{r},k}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{pr}}^{\mathrm{H}})\boldsymbol{w}_{\mathrm{p}}\right|^2+\sigma_k^2} \tag{3}$$

于是，第k个SU的可达速率为

$$R_k=\log_2\left(1+\gamma_k\right) \tag{4}$$

本文的目标是通过优化SBS发射波束成形矩阵$\boldsymbol{W}_{\mathrm{s}}=[\boldsymbol{w}_1,\cdots,\boldsymbol{w}_k,\cdots\boldsymbol{w}_K]\in\mathbb{C}^{M\times K}$和IRS相移矩阵$\boldsymbol{\Theta}$来最大化SU和速率，相应的优化问题表示为

$$
\begin{aligned}
(\mathrm{P1})\max_{\boldsymbol{W}_{\mathrm{s}},\boldsymbol{\Theta}}\ & \sum_{k=1}^{K}\log_2\left(1+\gamma_k\right)\\
\text{s.t. C1}\ & \sum_{k=1}^{K}\left|\left(\boldsymbol{h}_{\mathrm{dp}}+\boldsymbol{h}_{\mathrm{rp}}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}}\right)\boldsymbol{w}_k\right|^2\le P_0\\
\mathrm{C2}\ & \sum_{k=1}^{K}\left\|\boldsymbol{w}_k\right\|^2\le P_{\max}\\
\mathrm{C3}\ & |\theta_n|=1,\forall n\in[1,\cdots,n,\cdots,N]
\end{aligned}\tag{5}
$$

其中约束C1为SBS对PU的干扰容限约束，约束C2为SBS的总发射功率约束，约束C3为保证IRS相移矩阵是具有N个单位模分量的对角矩阵。由于单位模量约束以及相移矩阵$\boldsymbol{\Theta}$和波束成形向量$\boldsymbol{w}_k$的耦合，采用传统的凸优化方法计算复杂度和时间复杂度较高，求解比较困难。为解决该问题，本文提出了一种基于DRL的算法联合优化SBS波束成形和IRS相移矩阵，在较低复杂度的情况下最大限度地提高了SU和速率。

3　基于DRL的算法设计

DRL将深度学习的感知能力与强化学习的决策能力相结合，为复杂环境下的感知决策问题提供了解决思路。DDPG算法是确定性策略梯度算法和深度神经网络(Deep Neural Networks, DNN)的结合，是深度Q网络(Deep Q-Network, DQN)算法在连续动作空间中的扩展，可解决DQN算法无法直接应用于连续动作空间的问题。

3.1　DDPG算法框架及神经网络

DDPG可以在连续动作空间中更加有效地学习[16]。首先，根据演员-评论家(Actor-Critic, AC)方法构造策略函数μ和Q函数；然后，DDPG使用基于策略的演员神经网络和基于值的评论家神经网络分别去充当μ和Q函数的角色，如图2所示。$\mu(\boldsymbol{a}^{(t)}|\boldsymbol{s}^{(t)})$决定了在环境状态$\boldsymbol{s}^{(t)}$下智能体在第$t$个时间步长的动作$\boldsymbol{a}^{(t)}$，可以将其视为从状态到动作的映射。为了更好地探索，可通过添加噪声过程$\mathcal{N}$构造一个探索策略$\hat{\mu}(\boldsymbol{s})=\mu(\theta_\mu|\boldsymbol{s})+\mathcal{N}$，其中$\theta_\mu$表示策略$\mu$中神经网络的参数。$Q$函数定义为$Q_\mu(\theta|\boldsymbol{s}^{(t)},\boldsymbol{a}^{(t)})=\mathbb{E}_\mu(G^{(t)}|\boldsymbol{s}^{(t)},\boldsymbol{a}^{(t)})$，用于评估动作$a^{(t)}$对未来累计奖励的影响，其中$G^{(t)}=\sum_{\tau=0}^{\infty}\gamma^\tau r^{(t+\tau+1)}$表示累计奖励，$\gamma\in(0,1]$是折扣率。

为解决网络训练的不稳定即估计的Q值倾向于发散的问题，DDPG使用了DQN中的目标网络，因此DDPG算法框架存在4个神经网络：演员网络$\mu(\theta_\mu|\boldsymbol{s})$和目标演员网络$\mu'(\theta_{\mu'}|\boldsymbol{s})$，评论家网络$Q(\theta_Q|\boldsymbol{s},\boldsymbol{a})$和目标评论家网络$Q'(\theta_{Q'}|\boldsymbol{s},\boldsymbol{a})$，其中$\theta_{i/i'}(i=\mu/Q)$为需要训练的网络参数，且两两分别具有相同的网络结构，但参数不同。DDPG中使用了经验回放池$\mathcal{M}$，其包含了DQN中的经验数据组$(\boldsymbol{s}^{(t)},\boldsymbol{a}^{(t)},r^{(t)},\boldsymbol{s}^{(t+1)})$，经验回放池打破了经验之间的相关性，使神经网络能够更高效地训练。

本文使用的演员网络、评论家网络及目标网络的DNN结构如图3所示。所提出的评论家网络和演员网络的DNN结构均为完全连接的DNN，包含1个输入层、1个输出层以及2个维度分别为L_1和L_2的隐藏层。演员网络用于输出主被动波束成形策略；评论家网络用于评估演员网络并提供更新演员网络的梯度。使用批量归一化(Batch Normalization, BN)用于解决每一层输入数据分布的变化问题，并使新的分布更接近数据的真实分布。BN可加快训练速度，省略了随机失活、L1和L2正则化处理方法，提高了训练精度。在数据处理层，保持$\sum_{k=1}^{K}\|\boldsymbol{w}_k\|^2\le P_{\max}$和$|\theta_n|=1$，以实现式(5)中定义的功率约束C2和单位模量约束C3。为处理负值输入，本文的激活函数均采用tanh。

3.2　DDPG算法的深度神经网络训练

为了训练网络，首先从经验回放池$\mathcal{M}$中随机抽取小批量经验数据组$(\boldsymbol{s}^{(i)},\boldsymbol{a}^{(i)},r^{(i)},\boldsymbol{s}^{(i+1)})$

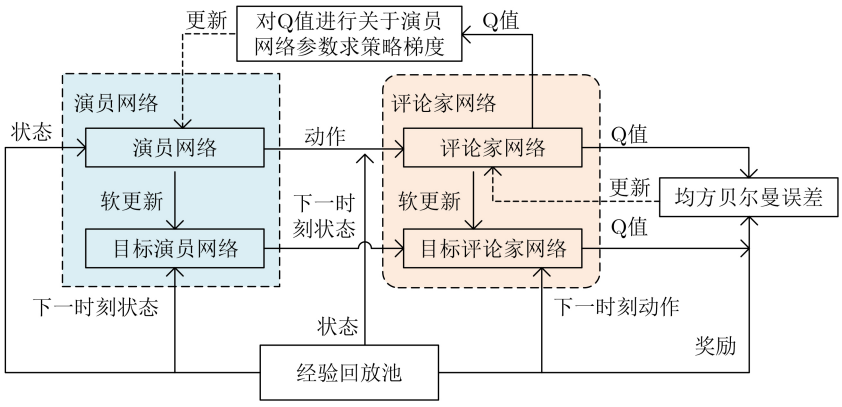

图 2 DDPG算法框架

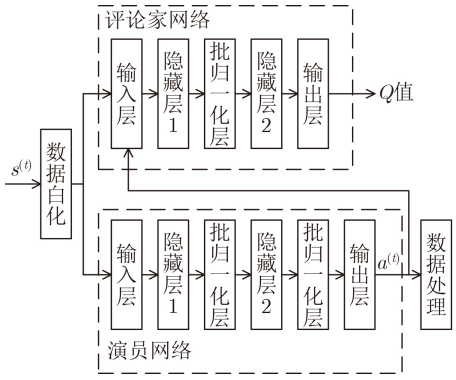

图 3 演员网络和评论家网络的DNN结构

$(i=1,\cdots,N_{\mathrm{B}})$，$N_{\mathrm{B}}$是小批量采样的大小。则第$i$个经验数据组$(\boldsymbol{s}^{(i)},\boldsymbol{a}^{(i)},r^{(i)},\boldsymbol{s}^{(i+1)})$所产生的Q值$y_i$为

$$y_i=r^{(i)}+\gamma Q'(\theta_{Q'}|\boldsymbol{s}^{(i+1)},\hat{\boldsymbol{a}}^{(i+1)}) \tag{6}$$

其中$\hat{\boldsymbol{a}}^{(i+1)}=\mu'(\theta_{\mu'}|\boldsymbol{s}^{(i+1)})$。评论家网络$Q(\theta_Q|s,\ a)$的损失函数可以定义为

$$L(\theta_Q)=\frac{1}{N_{\mathrm{B}}}\sum_{i=1}^{N_{\mathrm{B}}}(y_i-Q(\theta_Q|\boldsymbol{s}^{(i)},\boldsymbol{a}^{(i)}))^2 \tag{7}$$

性能目标函数$J(\mu_\theta)=\int_{\mathcal{S}}\rho^{\mu}(\boldsymbol{s})r(\boldsymbol{s},\ \boldsymbol{a})\mathrm{d}\boldsymbol{s}$用于衡量确定性策略梯度中的特定策略$\mu_\theta(\boldsymbol{s})$，其中$\rho^{\mu}(s)$表示策略$\mu$下的状态分布函数，$\boldsymbol{a}=\mu_\theta(\boldsymbol{s})$，$\mathcal{S}$表示环境的状态空间。当使用离线策略训练方法时，$\nabla_\theta J(\mu_\theta)=\int_S\rho^{\mu}(\boldsymbol{s})\nabla_\theta\mu_\theta(\boldsymbol{s})\nabla_a Q_\mu(\boldsymbol{s},\ \mu_\theta(\boldsymbol{s}))\mathrm{d}\boldsymbol{s}$即为策略梯度。基于蒙特卡罗方法，可以使用小批量数据获得$\nabla_\theta J(\mu_\theta)$的无偏估计。因此，演员网络的策略梯度可以重写为

$$\nabla_{\theta_\mu}J(\mu)\approx\frac{1}{N_{\mathrm{B}}}\sum_{i=1}^{N_{\mathrm{B}}}\nabla_a Q(\theta_Q|\boldsymbol{s}^{(i)},\boldsymbol{a}^{(i)})\nabla_{\theta_\mu}\mu(\theta_\mu|\boldsymbol{s}^{(i)}) \tag{8}$$

评论家网络和演员网络的更新如下

$$\theta_Q\leftarrow\theta_Q-\iota_Q\nabla_{\theta_Q}L\left(\theta_Q\right), \tag{9}$$

$$\theta_\mu\leftarrow\theta_\mu-\iota_\mu\nabla_{\theta_\mu}J\left(\mu\right), \tag{10}$$

其中ι_Q和ι_μ分别表示对评论家网络和演员网络进行更新的学习率。

目标网络通过软更新进行更新，可分别写为

$$\theta_{Q'}\leftarrow\tau_Q\theta_Q+(1-\tau_Q)\theta_{Q'}, \tag{11}$$

$$\theta_{\mu'}\leftarrow\tau_\mu\theta_\mu+(1-\tau_\mu)\theta_{\mu'}, \tag{12}$$

其中τ_Q和τ_μ分别表示对目标评论家网络和目标演员网络进行更新的学习率。

用于训练评论家网络和演员网络的优化器是Adam优化器，其自适应学习率分别为$\mu_{\mathrm{c}}^{(t)}=\lambda_{\mathrm{c}}\mu_{\mathrm{c}}^{(t-1)}$，$\mu_{\mathrm{a}}^{(t)}=\lambda_{\mathrm{a}}\mu_{\mathrm{a}}^{(t-1)}$，其中$\lambda_{\mathrm{c}}$和$\lambda_{\mathrm{a}}$为训练评论家网络和演员网络的衰减率。

3.3 基于DDPG的次用户和速率最大化算法流程

为了实现系统SU和速率最大化，基于本文所考虑的系统模型和目标函数对强化学习的基本要素进行设置如下。

(1)状态：状态$\boldsymbol{s}^{(t)}\in\mathcal{S}$表示在第$t$个时间步长从环境中获得的一组观测值，其中$\mathcal{S}$表示环境的状态空间。本文中的状态$\boldsymbol{s}^{(t)}$定义为

$$\boldsymbol{s}^{(t)}=\left[\boldsymbol{a}^{(t)},\boldsymbol{\gamma}_{\mathrm{SU}}^{(t)},\boldsymbol{H}\right] \tag{13}$$

其中，$\boldsymbol{a}^{(t)}$表示第t个时间步长智能体采取的动作，$\boldsymbol{\gamma}_{\mathrm{SU}}^{(t)}=\left[\gamma_{\mathrm{SU}1}^{(t)},\cdots,\gamma_{\mathrm{SU}k}^{(t)},\cdots,\gamma_{\mathrm{SU}K}^{(t)}\right]$表示动作$\boldsymbol{a}^{(t)}$作用于环境后时间步长$t$时SU的SINR，整个无线通信系统CSI为$\boldsymbol{H}=\left[\boldsymbol{h}_{\mathrm{pp}},\boldsymbol{G}_{\mathrm{pr}}^{\mathrm{H}},\boldsymbol{h}_{\mathrm{rp}},\boldsymbol{h}_{\mathrm{d},k},\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}},\boldsymbol{h}_{\mathrm{r},k},\boldsymbol{h}_{\mathrm{ps},k},\boldsymbol{h}_{\mathrm{dp}}\right]$。

(2)动作：动作$\boldsymbol{a}^{(t)}\in\mathcal{A}$表示根据观测到的状态$s^{(t)}$在时间步长$t$遵循策略$\mu$的一组行为选择，其中$\mathcal{A}$表示动作空间。本文中的动作$\boldsymbol{a}^{(t)}$定义为

$$\boldsymbol{a}^{(t)}=[\boldsymbol{W}_{\mathrm{s}}^{(t)},\boldsymbol{\Theta}^{(t)}] \tag{14}$$

其中$\boldsymbol{W}_{\mathrm{s}}^{(t)}$, $\boldsymbol{\Theta}^{(t)}$分别表示时间步长为$t$时的$\boldsymbol{W}_{\mathrm{s}}$和$\boldsymbol{\Theta}$。

(3)奖励：奖励$r^{(t)}$是环境在状态$\boldsymbol{s}^{(t)}$下执行动作$\boldsymbol{a}^{(t)}$后返回给智能体的第t个时间步长的奖励值。即时奖励评估了给定状态$\boldsymbol{s}^{(t)}$下动作$\boldsymbol{a}^{(t)}$的性能，智能体可以根据奖励r来调整自己的策略μ。考虑到本文的目标是最大化SU和速率，同时结合式(5)中的C1约束，奖励定义为

$$r^{(t)}=\left(\sum_{k=1}^{K}R_k\right)^{(t)}-p \tag{15}$$

其中R_k由式(4)计算得到，p是惩罚因子。为了动态设置p的值以适应不同的和速率情况，同时考虑到在式(4)中，$R_k=\log_2(1+\gamma_k)>0$恒成立，于是有

$$p=\begin{cases}10^{\lfloor\lg\sum_{k=1}^{K}R_{k,t}\rfloor},\sum_{k=1}^{K}\left|(\boldsymbol{h}_{\mathrm{dp}}+\boldsymbol{h}_{\mathrm{rp}}\boldsymbol{\Theta}\boldsymbol{G}_{\mathrm{sr}}^{\mathrm{H}})\boldsymbol{w}_k\right|^2>P_0\\0,\text{其它}\end{cases} \tag{16}$$

其中$\lfloor\cdot\rfloor$表示向下取整函数。

在每个时间步长t中，给定$\boldsymbol{a}^{(t)}=[\boldsymbol{W}_{\mathrm{s}}^{(t)},\boldsymbol{\Theta}^{(t)}]$，由动作$\boldsymbol{a}^{(t)}$计算所有SU在时间步长$t$的$\boldsymbol{\gamma}_{\mathrm{SU}}^{(t)}$，同时从环境中获得CSI $\boldsymbol{H}$以使智能体构建状态$s^{(t)}$。然后，根据式(15)计算相应的奖励$r^{(t)}$。接着，生成状态$\boldsymbol{s}^{(t+1)}=\left[\boldsymbol{a}^{(t)},\boldsymbol{\gamma}_{\mathrm{SU}}^{(t+1)},\boldsymbol{H}\right]$。接下来，以状态$\boldsymbol{s}^{(t+1)}$

为激励/输入，智能体基于θ_μ给出相应的动作$\boldsymbol{a}^{(t)}=[\boldsymbol{W}^{(t)},\boldsymbol{\Theta}^{(t)}]$。最后，将经验数据组$(\boldsymbol{s}^{(t)},\boldsymbol{a}^{(t)},r^{(t)},\boldsymbol{s}^{(t+1)})$保存到经验回放池以备DNN训练使用。如此循环往复直至达到停止条件。具体算法步骤如算法1所示。

3.4 复杂度分析

由于神经网络只能接受实数而不能接受复数作为输入，因此在状态$\boldsymbol{s}^{(t)}$以及动作$\boldsymbol{a}^{(t)}$的构建中，实部和虚部被分离为独立的输入输出，则主用户链路、次用户链路以及干扰链路的CSI维度分别为$D_1=2(NM+K_{\mathrm{PU}}N+K_{\mathrm{PU}}M)$，$D_2=2(NM+KN+KM)$以及$D_3=2(K_{\mathrm{PU}}M+KM)$，其中$K_{\mathrm{PU}}=1$为PU数量。因此，系统CSI的维度为$D_{\boldsymbol{H}}=D_1+D_2+D_3$。类似地，动作$\boldsymbol{a}^{(t)}=[\boldsymbol{W}_{\mathrm{s}}^{(t)},\boldsymbol{\Theta}^{(t)}]$的维度为$D_{\boldsymbol{a}}=2N+2MK$。$\gamma_{\mathrm{SU}}^{(t)}$表示时间步长$t$时SU的SINR，其维度为SU数量$D_{\mathrm{SINR}}=K$。因此状态的维度为$D_{\boldsymbol{s}}=D_{\boldsymbol{a}}+D_{\mathrm{SINR}}+D_{\boldsymbol{H}}$。

在DDPG算法中，评价家网络的输入层，第1个隐藏层，第2个隐藏层以及输出层的神经元数量分别是$D_{\boldsymbol{s}}$, L_1, L_2和1。而动作网络的输入层，第1个隐藏层，第2个隐藏层以及输出层的神经元数量分别是$D_{\boldsymbol{s}}$, L_1, L_2和$D_{\boldsymbol{a}}$。此外，算法需要执行$T_1\times T_2$步，因此，DDPG算法的复杂度为$\mathcal{O}\left(T_1T_2\left[(D_{\boldsymbol{s}}L_1+L_1L_2+L_2)+(D_{\boldsymbol{s}}L_1+L_1L_2+L_2D_{\boldsymbol{a}})\right]\right)$。

本文算法的空间复杂度是智能体和经验回放池$\mathcal{M}$大小的总和，即$\mathcal{O}(D_{\boldsymbol{s}}+D_{\boldsymbol{a}}+D_{\boldsymbol{s}+1}+1)$，其中，$D_{\boldsymbol{s}+1}$表示采用当前动作后进入的下一个状态的维度，1表示奖励的维度。

算法 1　基于DDPG的主被动波束成形算法训练

输入：IRS辅助的下行链路多用户MISO-CR系统的所有CSI

输出：最优动作$\boldsymbol{a}=\{\boldsymbol{W}_{\mathrm{s}},\boldsymbol{\Theta}\}$，Q值函数

初始化：大小为$\mathcal{D}$经验回放池$\mathcal{M}$，随机初始化演员和评论家网络参数θ_μ和θ_Q，赋值$\theta_{Q'}\leftarrow\theta_Q$，$\theta_{\mu'}\leftarrow\theta_\mu$

for episode $=1,2,3,\cdots,T_1$，进入循环

初始化发射波束成形矩阵$\boldsymbol{W}_{\mathrm{s}}^{(0)}$、相移矩阵$\boldsymbol{\Theta}^{(0)}$为单位矩阵作为$\boldsymbol{a}^{(0)}$

构建初始状态$\boldsymbol{s}^{(0)}$

for time steps$=1,2,3,\cdots,T_2$，进入循环

从演员网络中获取动作$a^{(t)}$

根据式(15)计算即时奖励$r^{(t)}$

根据式(3)计算所有SU的信干噪比$\gamma_{\mathrm{SU}}^{(t)}$

构建在动作$\boldsymbol{a}^{(t)}$下的状态$\boldsymbol{s}^{(t+1)}$

存储经验数据组$\left(\boldsymbol{s}^{(t)},a^{(t)},r^{(t)},\boldsymbol{s}^{(t+1)}\right)$到经验回放池中

从$\mathcal{M}$中随机抽取大小为N_{B}的小批量经验样本

根据式(6)得到目标Q值

根据式(7)得到在线评论家网络损失函数$L(\theta_Q)$

根据式(8)得到在线演员网络策略梯度$\nabla_{\theta_\mu}J(\mu)$

根据式(9)更新评论家网络参数ι_Q

根据式(10)更新演员网络参数ι_μ

根据式(11)更新目标评论家网络参数τ_Q

根据式(12)更新目标演员网络参数τ_μ

更新状态$\boldsymbol{s}^{(t)}\leftarrow\boldsymbol{s}^{(t+1)}$

end for

end for

4　仿真与分析

为评估算法的性能，本节设置了仿真参数并进行了仿真分析，如图4所示，其中PBS和SBS均配备有$M=8$副天线。PBS和SBS分别位于$(86,0,0)$和$(86,100,0)$，该系统含有一个位于$(1,48,0)$的PU，$K=2$个均匀随机分布在以$(1,52,0)$为圆心，半径为3，且$x>0$的区域内的SU。IRS位于$(0,50,2)$。

除非特别说明，PBS和SBS均配备有$M=8$副天线，IRS反射单元数量$N=8$，SU数量$K=2$，SBS发射功率$P_{\max}=30$ dBm，PBS发射功率$P_{\mathrm{PBS}}=l10$ dBm，SBS对PU的干扰容限约束阈值$P_0=-60$ dBm，所有SU接收天线处噪声均为$\sigma_s^2=-100$ dBm。

本文所有信道均被建模为$\boldsymbol{g}=\boldsymbol{h}\sqrt{C_0(d/d_0)^{-\alpha}}$，其中$d$为传输距离，$C_0=-30$ dB为在参考距离$d_0=1$ m时的路径损耗，α为路径损耗指数。设IRS到PU和所有SU之间的自由空间路径损耗指数均为2，即$\alpha_{\mathrm{I,PU}}=\alpha_{\mathrm{I,SU}}=2$，其余链路的路径损耗指数均设置为3。小尺度衰落服从莱斯分布，即$\boldsymbol{h}=\sqrt{\beta/(1+\beta)}\boldsymbol{h}^{\mathrm{LOS}}+\sqrt{1/(1+\beta)}\boldsymbol{h}^{\mathrm{NLOS}}$，其中，$\beta$为莱斯因子。设IRS到PU和第$k$个SU之间的莱斯因子为$\infty$，即$\beta_{\mathrm{I,PU}}=\beta_{\mathrm{I,SU}_k}=\infty$，其余链路的莱斯因子设置为0。$\boldsymbol{h}^{\mathrm{LOS}}$表示信道的视距部分[18]，$\boldsymbol{h}^{\mathrm{NLOS}}$为信道的非视距部分服从瑞利分布。基于DRL的DDPG算法的超参数设置如表1所示。

为了评估所提算法的有效性，本节首先将所提算法与基于AO的算法进行比较。此外，还以随机IRS反射相移方案和无IRS辅助方案作为基准进行

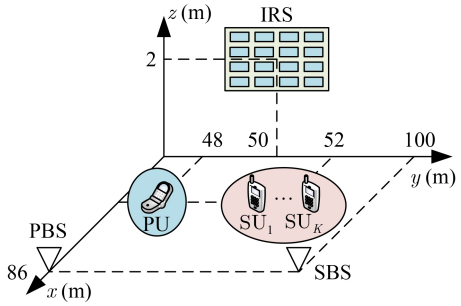

图 4 仿真场景图

了比较，这两种方案的SBS波束成形矩阵均由本文算法获得。所有展示的数据都是10个独立实现的结果的平均。

图5给出了不同SBS发射功率情况下，所提算法与另外3种基准对比算法的和速率性能比较。可以观察到，4种算法中，SU和速率均随着$P_{\max}$的增大而增大。同时，本文所提算法与无IRS以及IRS随机相移的方案相比，性能得到明显提升。此外，基于AO的方案整体上可达到的和速率高于本文所提算法，但是随着IRS反射单元的增加，基于AO的方案迭代出最优策略的时间相应也要增加。表2展示了在IRS反射单元数量N不断增加情况下，基于AO的方案和本文算法的运行时间对比。

为了更好地理解所提算法，本文研究了图6所示的IRS反射单元数量N对系统性能及算法收敛的影响，其中$M=4, K=4, P_{\max}=25$ dBm，使用式(17)计算平均奖励

表 1 DDPG算法参数

超参数	描述	参数值
γ	折扣率	0.99
ι_μ, ι_Q	演员、评论家网络的学习率	0.001
τ_μ, τ_Q	目标演员、目标评论家网络的学习率	0.001
λ_a, λ_c	训练演员、评论家网络的衰减率	0.00001
L_1, L_2	DNN隐藏层神经元数	1024
$\mathcal{D}$	经验回放池$\mathcal{M}$的大小	100000
T_1	回合数	10
T_2	每个回合的时间步长数	1000000
N_{B}	小批量采样的大小	16

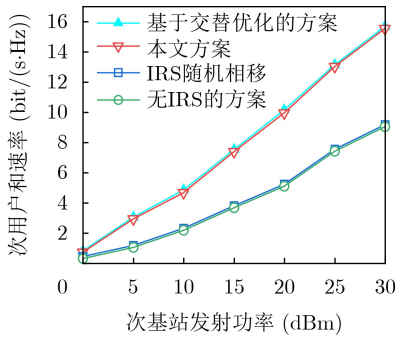

图 5 SBS发射功率与SU和速率的关系

表 2 不同算法运行时间对比

IRS反射单元数	基于交替优化(ms)	本文算法(ms)
N=4	968.76	16.24
N=10	1367.41	16.84
N=20	2248.25	16.36
N=30	3018.52	16.65

$$\text{average\_reward}(T_i)=\frac{\sum_{t=1}^{T_i}\text{reward}\,(t)}{T_i}, T_i=1,2,\cdots,T \tag{17}$$

其中T是最大时间步长数。从图6可以看出，奖励会随着时间步长t的增加而收敛。与发射功率相比，DRL对系统设置的变化具有更强的鲁棒性。具体来说，随着IRS反射单元数量N的增加，平均奖励也如预期的那样逐渐增加，但这并没有增加本文所提算法的收敛时间。

图7展示了在不同发射功率下，随着时间步长的增加，即时奖励$r^{(t)}$和平均奖励的变化情况。图7表明，基于DRL的算法从单位矩阵出发，能够在与环境的交互过程中学习，调整$\boldsymbol{W}_{\mathrm{s}}$和$\boldsymbol{\Theta}$来接近最优解。由图可知，算法所得到的即时奖赏$r^{(t)}$的方差会随着$P_{\max}$的增大而增大。出现这种现象的原因是$P_{\max}$越大，即时奖励的动态范围越大，导致波动越大，收敛性越差进而导致即时奖赏方差增大。

图8给出了不同SBS发射功率下，平均奖励与时间步长的关系，可以看出，发射功率对收敛速度和性能有着显著影响。当$P_{\max}\geq 10$ dBm时，性能差距远大于$P_{\max}=0$ dBm和$P_{\max}=10$ dBm之间的

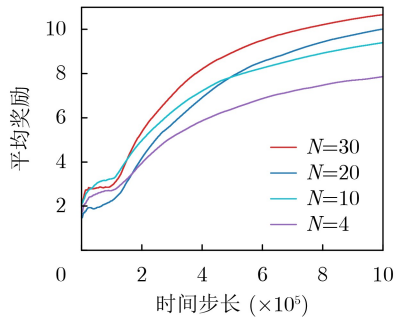

图 6 不同反射单元数量下算法的收敛性能

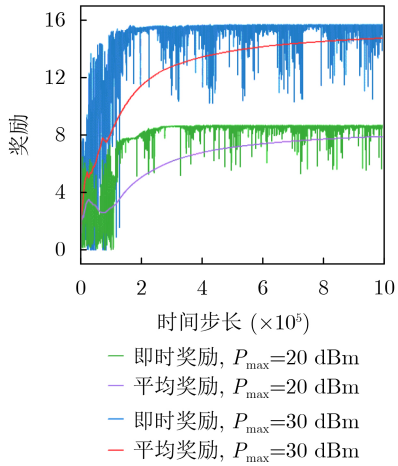

图 7 不同SBS发射功率下奖励与时间步长的关系

差距，即本文算法对高信噪比非常敏感，但其实现收敛的时间更长。

本文所提算法中，对评论家和演员网络使用恒定学习率和衰减率，同时通过仿真研究二者对本文算法的性能和收敛速度的影响。图9展示了不同学习率下，平均奖励与时间步长的关系。可以看出，不同的学习率对DRL算法的性能影响很大。其中，学习率为0.001时系统达到了最佳性能，而学习率为0.01的DRL性能最差，甚至对算法的收敛产生了最坏的影响，其原因是过大的学习率会增加振荡，使得性能急剧下降。综上所述，学习率的选择要适当，不能过大也不能过小。

图10比较了不同衰减率下，平均奖励与时间步长的关系，可以观察到不同的衰减率对算法的性能和收敛速度会产生一定的影响，但与学习率相比，衰减率对DRL性能和收敛速度的影响较小。

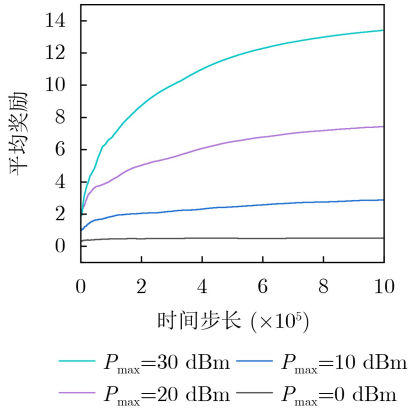

图 8 不同SBS发射功率下平均奖励与时间步长的关系

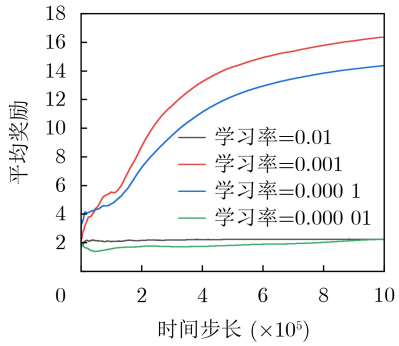

图 9 不同学习率下的平均奖励与时间步长的关系

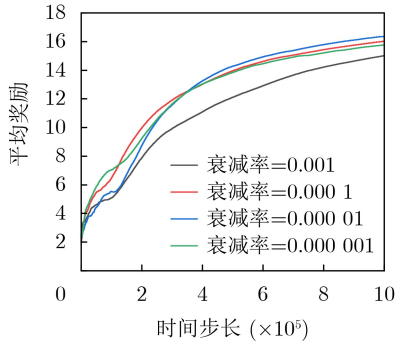

图 10 在不同衰减率下的平均奖励与时间步长的关系

5　结束语

针对当前无线通信系统频谱资源需求日益增长的问题，本文提出了一种基于DRL的IRS辅助CRN系统资源分配方案，通过优化SBS发射波束成形矩阵和IRS反射相移矩阵，实现SU和速率最大化。首先，针对IRS辅助的CRN进行了关于SU和速率最大化问题进行问题分析和问题建模；其次，研究了满足PU干扰容限约束、SBS最大发射功率约束以及IRS相移约束下的资源分配问题，提出了基于DDPG算法的联合波束成形算法；最后通过仿真实验，验证了所提算法在通信和速率方面的优异性能，与基于AO的凸优化算法相比，具有较低的计算复杂度及时间复杂度。未来可以在非完美信道、多输入多输出以及窃听场景下对所提资源分配算法的性能进行进一步的研究。

参 考 文 献

[1] LI Guoquan, HONG Zijie, PANG Yu, *et al*. Resource allocation for sum-rate maximization in NOMA-based generalized spatial modulation[J]. *Digital Communications and Networks*, 2022, 8(6): 1077–1084. doi: 10.1016/j.dcan.2022.02.005.

[2] LI Xingwang, ZHENG Yike, ALSHEHRI M D, *et al*. Cognitive AmBC-NOMA IoV-MTS networks with IQI: Reliability and security analysis[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(2): 2596–2607. doi: 10.1109/TITS.2021.3113995.

[3] 李国权, 党刚, 林金朝, 等. RIS辅助的MISO系统安全鲁棒波束赋形算法[J]. 电子与信息学报, 2023, 45(8): 2867–2875. doi: 10.11999/JEIT220894.

LI Guoquan, DANG Gang, LIN Jinzhao, *et al*. Secure and robust beamforming algorithm for RIS assisted MISO systems[J]. *Journal of Electronics & Information Technology*, 2023, 45(8): 2867–2875. doi: 10.11999/JEIT220894.

[4] CHEN Guang, CHEN Yueyun, MAI Zhiyuan, *et al*. Joint multiple resource allocation for offloading cost minimization in IRS-assisted MEC networks with NOMA[J]. *Digital Communications and Networks*, 2023, 9(3): 613–627. doi: 10.1016/j.dcan.2022.10.029.

[5] 熊军洲, 李国权, 王钥涛, 等. 基于有源智能反射面反射单元分组的反射调制系统[J]. 电子与信息学报, 2024, 46(7): 2765–2772. doi: 10.11999/JEIT231187.

XIONG Junzhou, LI Guoquan, WANG Yuetao, *et al*. A reflection modulation system based on reflecting element grouping of active intelligent reflecting surface[J]. *Journal of*

Electronics & Information Technology, 2024, 46(7): 2765–2772. doi: 10.11999/JEIT231187.

[6] GUAN Xinrong, WU Qingqing, and ZHANG Rui. Joint power control and passive beamforming in IRS-assisted spectrum sharing[J]. *IEEE Communications Letters*, 2020, 24(7): 1553–1557. doi: 10.1109/LCOMM.2020.2979709.

[7] LE A T, DO D T, CAO Haotong, *et al*. Spectrum efficiency design for intelligent reflecting surface-aided IoT systems[C]. GLOBECOM 2022 - 2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 2022: 25–30. doi: 10.1109/GLOBECOM48099.2022.10000937.

[8] YUAN Jie, LIANG Yingchang, JOUNG J, *et al*. Intelligent Reflecting Surface (IRS)-enhanced cognitive radio system[C]. ICC 2020 - 2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 2022: 1–6. doi: 10.1109/ICC40277.2020.9148890.

[9] WANG Zining, LIN Min, HUANG Shupei, *et al*. Robust beamforming for IRS-aided SWIPT in cognitive radio networks[J]. *Digital Communications and Networks*, 2023, 9(3): 645–654. doi: 10.1016/j.dcan.2022.10.030.

[10] LI Guoquan, ZHANG Hui, WANG Yuhui, *et al*. QoS guaranteed power minimization and beamforming for IRS-assisted NOMA systems[J]. *IEEE Wireless Communications Letters*, 2023, 12(3): 391–395. doi: 10.1109/LWC.2022.3189272.

[11] FENG Keming, WANG Qisheng, LI Xiao, *et al*. Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems[J]. *IEEE Wireless Communications Letters*, 2020, 9(5): 745–749. doi: 10.1109/LWC.2020.2969167.

[12] HUANG Chongwen, MO Ronghong, and 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. doi: 10.1109/JSAC.2020.3000835.

[13] YANG Helin, XIONG Zehui, ZHAO Jun, *et al*. Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications[J]. *IEEE Transactions on Wireless Communications*, 2021, 20(1): 375–388. doi: 10.1109/TWC.2020.3024860.

[14] ZHONG Canwei, CUI Miao, ZHANG Guangchi, *et al*. Deep reinforcement learning-based optimization for IRS-assisted cognitive radio systems[J]. *IEEE Transactions on Communications*, 2022, 70(6): 3849–3864. doi: 10.1109/TCOMM.2022.3171837.

[15] GUO Jianxin, WANG Zhe, LI Jun, *et al*. Deep reinforcement learning based resource allocation for intelligent reflecting surface assisted dynamic spectrum sharing[C]. 2022 14th International Conference on Wireless Communications and Signal Processing (WCSP), Nanjing, China, 2022: 1178–1183. doi: 10.1109/WCSP55476.2022.10039119.

[16] LILLICRAP T P, HUNT J J, PRITZEL A, *et al*. Continuous control with deep reinforcement learning[C]. 4th International Conference on Learning Representations, San Juan, Puerto Rico, 2016.

[17] WEI Yi, ZHAO Mingmin, ZHAO Minjian, *et al*. Channel estimation for IRS-aided multiuser communications with reduced error propagation[J]. *IEEE Transactions on Wireless Communications*, 2022, 21(4): 2725–2741. doi: 10.1109/TWC.2021.3115161.

[18] HAN Yu, TANG Wankai, JIN Shi, *et al*. Large intelligent surface-assisted wireless communication exploiting statistical CSI[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(8): 8238–8242. doi: 10.1109/TVT.2019.2923997.

李国权：男，教授，博士生导师，研究方向为无线资源管理、智能反射面优化等.

程 涛：男，硕士生，研究方向为无线资源管理、智能反射面.

郭永存：男，硕士生，研究方向为无线资源管理、智能反射面.

庞 宇：男，教授，博士生导师，研究方向为集成电路设计、无线通信和人工智能等.

林金朝：男，教授，博士生导师，研究方向为无线通信传输技术与优化等.

责任编辑：陈 倩

Deep Reinforcement Learning Based Beamforming Algorithm for IRS Assisted Cognitive Radio System

LI Guoquan[①] CHENG Tao[①] GUO Yongcun[①] PANG Yu[②] LIN Jinzhao[②]

[①](*School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications*, , *Chongqing* 400065, *China*)

[②](*Chongqing Key Laboratory of Optoelectronic Information Sensing and Microsystems*, *Chongqing* 400065, *China*)

Abstract:

Objective With the rapid development of wireless communication technologies, the demand for spectrum resources has significantly increased. Cognitive Radio (CR) has emerged as a promising solution to improve spectrum utilization by enabling Secondary Users (SUs) to access licensed spectrum bands without causing harmful interference to Primary Users (PUs). However, traditional CR networks face challenges in achieving high spectral efficiency due to limited control over the wireless environment. Intelligent Reflecting Surfaces (IRS) have recently been introduced as a revolutionary technology to enhance communication performance by dynamically reconfiguring the propagation environment. This paper aims to maximize the sum rate of SUs in an IRS-assisted CR network by jointly optimizing the active beamforming at the Secondary Base Station (SBS) and the passive beamforming at the IRS, subject to constraints on the maximum transmit power of the SBS, the interference tolerance of PUs, and the unit modulus of the IRS phase shifts.

Methods To address the non-convex and highly coupled optimization problem, a Deep Reinforcement Learning (DRL)-based algorithm is proposed. Specifically, the problem is formulated as a Markov Decision Process (MDP), where the state space includes the Channel State Information (CSI) of the entire system, the Signal-to-Interference-plus-Noise Ratio (SINR) in the SU network, and the action space consists of the SBS beamforming vectors and the IRS phase shift matrix. The reward function is designed to maximize the sum rate of SUs while penalizing violations of the constraints. The Deep Deterministic Policy Gradient (DDPG) algorithm is used to solve the MDP, owing to its ability to handle continuous action spaces. The DDPG framework consists of an actor network, which outputs the optimal actions, and a critic network, which evaluates these actions based on the reward function. The training process involves interacting with the environment to learn the optimal policy, and the algorithm is fine-tuned to ensure convergence and robustness under varying system conditions.

Results and Discussions Simulation results show that the proposed scheme achieves comparable sum rate performance with lower time complexity after optimization, compared to traditional optimization algorithms. The proposed algorithm significantly outperforms the no-IRS and IRS-random phase shift schemes (Fig. 5). The results demonstrate that the proposed algorithm achieves a sum rate close to that of alternating optimization-based approaches (Fig. 5), while substantially reducing computational complexity (Fig. 5, Table 2). Additionally, the impact of the number of IRS elements on the sum rate is examined (Fig. 6). As expected, the average reward increases with the number of reflecting elements, while the convergence time remains stable, indicating the robustness of the proposed algorithm. The DRL-based algorithm, starting from the identity matrix, can learn and adjust the beamforming vectors and phase shifts to approach the optimal solution through interaction with the environment (Fig. 7). It is also observed that the variance of the instantaneous reward increases with the transmit power. This is due to the larger dynamic range of the instantaneous reward at higher power levels, resulting in greater fluctuations and slower convergence. The relationship between average reward and time steps under different transmit power levels is presented, highlighting the sensitivity of the algorithm to high signal-to-noise ratios (Fig. 8). Moreover, it can be observed that a learning rate of 0.001 yields the best performance, while excessively high or low learning rates degrade performance (Fig. 9). The discount factor has a relatively smaller impact on performance compared to the learning rate (Fig. 10).

Conclusions This paper proposes a DRL-based algorithm for joint active and passive beamforming optimization in an IRS-assisted CR network. The algorithm utilizes the DDPG framework to maximize the sum rate of SUs while adhering to constraints on transmit power, interference, and IRS phase shifts. Simulation results demonstrate that the proposed algorithm achieves comparable sum rate performance to traditional optimization methods, with significantly lower computational complexity. The findings also highlight the impact of DRL parameter settings on performance. Future work will focus on extending the proposed algorithm to multi-cell scenarios and incorporating imperfect CSI to enhance its robustness in practical environments.

Key words: Intelligent Reflecting Surface (IRS); Cognitive Radio (CR); Deep Reinforcement Learning (DRL); Beamforming