基于DDPG的智能反射面辅助无线携能通信系统性能优化

罗丽平，潘伟民
（广西民族大学电子信息学院，广西 南宁 530006）

摘 要：针对智能反射面（IRS, intelligent reflecting surface）辅助的多输入单输出（MISO, multiple input single-output）无线携能通信（SWIPT, simultaneous wireless information and power transfer）系统，考虑基站最大发射功率、IRS 反射相移矩阵的单位膜约束和能量接收器的最小能量约束，以最大化信息传输速率为目标，联合优化了基站处的波束成形向量和智能反射面的反射波束成形向量。为解决非凸优化问题，提出了一种基于深度强化学习的深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法。仿真结果表明，DDPG 算法的平均奖励与学习率有关，在选取合适的学习率的条件下，DDPG 算法能获得与传统优化算法相近的平均互信息，但运行时间明显低于传统的非凸优化算法，即使增加天线数和反射单元数，DDPG 算法依然可以在较短的时间内收敛。这说明 DDPG 算法能有效地提高计算效率，更适合实时性要求较高的通信业务。

关键词：多输入单输出；无线携能通信；智能反射面；波束成形；深度确定性策略梯度
中图分类号：TN929.5；TP18
文献标志码：A
doi：10.11959/j.issn.2096−3750.2024.00389

DDPG-based performance optimization algorithm for IRS-assisted simultaneous wireless information and power transfer systems

LUO Liping, PAN Weimin
College of Electronic Information, Guangxi Minzu University, Nanning 530006, China

Abstract: For the intelligent reflecting surface (IRS)-assisted multiple input single output (MISO) simultaneous wireless information and power transfer (SWIPT) system, the beam forming vector at the base station and the reflected beam forming vector of the IRS were jointly optimized, by considering the maximum transmit power of the base station, the unit modulus constraint of the IRS reflection phase shift matrix, and the minimum energy constraint of the energy receiver. The object was to maximize the spectrum efficiency. To solve the non-convex optimization problem, a deep deterministic policy gradient (DDPG) algorithm based on deep reinforcement learning was proposed. Simulation results show that the average reward of the DDPG algorithm is related to the learning rate. Under the condition of selecting the appropriate learning rate, the DDPG algorithm can obtain an average mutual information similar to that of the traditional optimization algorithm, but the running time is significantly lower than that of the traditional non-convex optimization algorithm. Even if the number of antennas and the number of reflective units are increased, the DDPG algorithm can still converge in a short period of time. This indicates that the DDPG algorithm can effectively improve the computational efficiency and is suitable for communication services with high real-time requirements.

Key words: multiple input single output, simultaneous wireless information and power transfer, intelligent reflecting surface, beam forming, deep deterministic policy gradient

收稿日期：2024−01−15；**修回日期：**2024−05−09
通信作者：罗丽平，luoliping@gxmzu.edu.cn
基金项目：广西科技重大专项（No. AA23073006）；广西民族大学研究生创新计划（No. gxun-chxs2022298）
Foundation Items: Guangxi Science and Technology Major Project (No. AA23073006), Guangxi Minzu University Graduate Innovation Program (No. gxun-chxs2022098)

0　引言

随着5G网络的大规模商用，学者们开始了6G技术的研究。新的用户需求、新的应用场景和新的网络趋势需要全新的通信模式，尤其是在物理层。目前常用的大规模多输入多输出（MIMO, multiple input multiple output）、毫米波通信和超密集网络等关键技术给5G通信网络提供了强有力的支撑，但是这些技术依然面临着两大挑战：一是在网络实施中的能耗问题，二是在恶劣传播环境下的信息可靠传输问题。现有技术缺乏对无线信道的主动控制，导致无法在恶劣环境下提供不间断的连接和令人满意的服务质量。智能反射面（IRS, intelligent reflecting surface）可以通过智能控制传输环境来提高系统频谱效率和能量效率，并具有低成本、容易部署的优势，能够突破上述两大瓶颈，成为6G的关键技术之一。

目前，将IRS与其他无线通信技术相结合已成为无线通信领域的研究热点。学者们纷纷考虑各种各样的应用场景，如将IRS应用在边缘计算、设备对设备（D2D, device to device）通信、无人机通信、认知无线电（CR, cognitive radio）网络、非正交多址接入（NOMA, non-orthogonal multiple access）、无线携能通信（SWIPT, simultaneous wireless information and power transfer）等[1-3]。SWIPT成为一个迅速兴起的研究领域，它可以通过射频信号同时传输信息和能量[4]。文献[5]从信息论的角度研究了具有点对点单天线高斯信道的SWIPT系统，并表征了无线携能通信中信息和能量之间的基本权衡。文献[6]研究了具有两个用户的多输入多输出SWIPT系统，其中一个用户收获能量，另一个用户解码信息。后续的研究扩展到多个发射机[7]和不完美信道状态信息（CSI, channel state information）[8]的情况。文献[9]研究了SWIPT-NOMA-CR网络中，考虑非理想CSI条件下，3种中继传输方案对次用户中断性能的影响。通过理论分析，得到了次用户中断概率的解析表达式，并使用蒙特卡洛仿真验证了理论结果。文献[10]研究分析了结合SWIPT和NOMA的认知中继网络中断性能，考虑了直接链路和两阶段传输模式，推导出了系统中断概率的解析和高信噪比下的渐近表达式，并通过蒙特卡洛仿真验证了理论分析。

基于IRS辅助的无线携能通信技术在物联网中有广泛的应用，通过合理部署IRS，可以大幅提升SWIPT网络的能量效率和频谱效率。

针对智能反射面辅助的通信系统多输入单输出（MISO, multiple input single output）SWIPT系统，文献[11]采用交替优化（AO, alternating optimization）和半定松弛（SDR, semidefinite relaxation）方法联合优化信息传输和能量传输的波束成形矩阵以及反射波束成形矩阵，最大化能量收集器接收的最小功率。文献[12]采用SDR和基于惩罚的流行优化算法最大化系统的可达速率。文献[13]采用SDR技术解决单天线移动联合设计波束成形向量和接收功率分配比问题，从而使基站的总发射功率最小。文献[14]针对太赫兹频段的SWIPT安全系统，提出了一种鲁棒波束成形策略，以最小化发射功率并满足中断率约束。通过伯恩斯坦不等式和半定规划技术，将非凸问题转化为凸问题，并采用交替优化算法求解。文献[15]研究了一种IRS辅助的异构网络，该网络结合了能量收集设备和非能量收集设备，在考虑能量消耗和传输时间调度的约束下，通过半定规划松弛和基于黎曼流形优化的方法，解决了总吞吐量最大化问题。文献[16]研究了毫米波系统中的保密速率优化，采用SWIPT架构为能量受限设备充电，提出了基于半定规划松弛的交替优化算法，仿真结果表明该算法有效地提升了保密速率。文献[17]研究了在IRS辅助下的MISO网络中，针对可能的窃听者和不完美信道状态信息，设计了鲁棒波束成形以最大化合法接收器的最小信息率，该方法采用了交替优化和连续凸近似方法，结合惩罚对偶分解处理IRS的单位模约束。文献[18]提出了一种IRS辅助的物联网能量传输策略，通过优化相位偏移和功率分配来提升网络吞吐量，该方法采用了交替优化算法和半定松弛处理非凸问题，闭式解简化了特殊情况下的计算。文献[19]提出了一种针对IRS辅助的无线携能通信系统的安全波束成形设计，以最大化能量采集器功率，该方法通过优化发射波束成形和IRS相移，采用松弛和半定规划方法解决非凸问题。文献[20]研究了MISO SWIPT广播信道，将优化问题解耦为具有固定辅助变量的子问题，通过凸优化技术找到最优辅助变量来计算全局最优传输比和传输协方差矩阵。数值结果表明，所提出的联合传输方案优于传统的传输方法。上述研究工作表

明，采用IRS辅助传输可以显著提升MISO SWIPT系统的性能。

尽管与IRS相关的优化问题可以通过非凸优化来解决[11-20]，但是由于优化变量通常是高维度的IRS反射波束成形向量，当系统参数（如信道系数）更新时，需要大量的运算，特别是IRS反射单元数较多时，将导致高额的计算复杂度。近年来，深度强化学习作为一种新兴技术，在解决海量数据分析、非线性非凸问题和高计算量问题上呈现良好的性能[21-25]。由于深度强化学习智能体可以通过与环境（如CSI和干扰）的交互学习做出最优决策，包括发射功率分配、波束成形矩阵设计等，因而被学者们用于解决通信系统中的资源优化问题[26]。文献[27]提出了一种基于深度确定性策略梯度（DDPG, deep deterministic policy gradient）算法来解决IRS辅助的MISO系统的接收信噪比（SNR, signal to noise ratio）最大化问题。结果表明，与基于优化理论的SDR方法相比，DDPG算法能够以更短的计算时间获取最优的SNR值。针对IRS辅助的MISO系统，文献[28]采用DDPG算法联合优化基站处和IRS处的波束成形向量，以最大化系统的遍历容量。针对IRS辅助的非正交多址网络，文献[29]采用DDPG算法，最大化系统的平均互信息。文献[30]利用深度强化学习（DRL, deep reinforcement learning）优化IRS辅助的多天线接入点下行传输，以最小化功率消耗。综上所述，基于DDPG的算法可以有效地解决无线通信系统的资源优化问题。与传统的优化理论算法相比，DDPG方法具有计算量少、运行时间短的优点。对于时延要求高的通信业务，运行时间和计算量会直接影响通信服务质量，降低资源优化算法的运算量对于时延受限系统非常重要。本文将DDPG算法应用在IRS辅助的MISO SWIPT系统中，通过求解最优的波束成形和反射波束成形向量，达到最大化信息传输速率的目标。本文的主要工作包括以下两点。

1) 基于马尔可夫决策过程，提出了DDPG资源优化算法，通过联合优化IRS反射相移矩阵和基站的波束成形向量，达到最大化系统平均互信息的目的。所提出的DDPG算法时间复杂度低，不需要无线环境的显式模型和具体的数学公式，能够离线学习关于环境的知识并适应环境，部署在线上时具有运算量少、时延低的优点。

2) 仿真结果表明，本文所提的DDPG优化算法能够通过观测瞬时回报来学习环境，并逐步改进其行为，从而获得最优的发射波束成形向量和反射相移矩阵并提高系统的信息传输速率。与传统的非凸优化算法相比，本文所提的DDPG算法在信息传输性能上与传统优化算法接近，且随着反射单元数的增加，传统算法时间复杂度呈线性增加，但DDPG算法的时间复杂度几乎不变。

1　系统模型与问题描述

1.1　系统模型

系统模型如图1所示，考虑一个IRS辅助的MISO SWIPT系统，其中，基站（BS, base station）向能量收集器（ER, energy harvesting）发送能量信号，向信号接收器（IR, information receiver）发送信息信号。假设BS有$N_t(N_t>1)$根发射天线，每个接收机安装单根天线，IRS通过控制M个反射单元来被动反射接收到的信号。$\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}}\in\mathbb{C}^{M\times 1}$，$\boldsymbol{h}_{\mathrm{r2}}^{\mathrm{H}}\in\mathbb{C}^{M\times 1}$，$\boldsymbol{h}_{1}^{\mathrm{H}}\in\mathbb{C}^{N_t\times 1}$，$\boldsymbol{h}_{2}^{\mathrm{H}}\in\mathbb{C}^{N_t\times 1}$，$\boldsymbol{Z}\in\mathbb{C}^{M\times N_t}$分别表示从IRS到IR、IRS到ER、BS到IR、BS到ER和BS到IRS的信道系数。假设BS和IRS都可以获得准确的CSI，BS基于所获得的CSI计算IRS相移，并通过专用的反馈信道将相移传回IRS[12]。IRS接收多径信号并通过反射单元反射这些信号，从而辅助BS信号传输。因此，IR的接收信号可以表示为式(1)，ER的接收信号可以表示为式(2)。

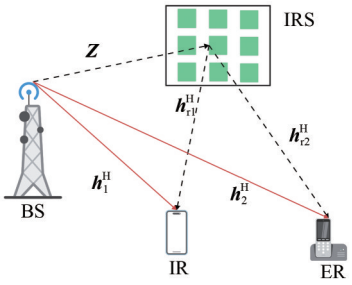

图1　系统模型

$$y_{\mathrm{IR}}=\left(\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{1}^{\mathrm{H}}\right)\boldsymbol{f}x+n_1 \tag{1}$$

$$y_{\mathrm{ER}}=\left(\boldsymbol{h}_{\mathrm{r2}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{2}^{\mathrm{H}}\right)\boldsymbol{f}x+n_2 \tag{2}$$

其中，$\boldsymbol{f}\in\mathbb{C}^{N_t\times 1}$表示发射波束成形向量，$x$为发射信号。$n_1\sim\subset\aleph(0,\sigma^2)$和$n_2\sim\subset\aleph(0,\sigma^2)$，分别表示IR和ER处的加性高斯白噪声，方差$\sigma^2=1$。$\boldsymbol{\phi}=\mathrm{diag}\left(\mathrm{e}^{\mathrm{j}\theta_1},\mathrm{e}^{\mathrm{j}\theta_2},\cdots,\mathrm{e}^{\mathrm{j}\theta_M}\right)$表示IRS的相移矩阵，$\theta_m\in[0,2\pi)$，$m$=1，2，…，$M$。

1.2　问题描述

假设输入信号服从高斯分布[12]，设基站最大发射功率为P，IRS反射角度受限，则信号接收器IR端的平均互信息可表示为式(3)

$$I_{\mathrm{G}}(x;y_{\mathrm{IR}})=\mathrm{lb}_2\left(1+\frac{\left|\left(\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{1}^{\mathrm{H}}\right)\boldsymbol{f}\right|^2}{\sigma^2}\right) \tag{3}$$

当能量接收器IE所要求的最小接收能量为$\bar{E}$时，通过联合优化基站处的发射波束成形向量$\boldsymbol{f}$和IRS的反射系数矩阵$\boldsymbol{\phi}$，以最大化IR端的平均互信息表示为式(4)

$$\bar{I}(x;y_{\mathrm{IR}})=\left|\left(\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{1}^{\mathrm{H}}\right)\boldsymbol{f}\right|^2 \tag{4}$$

优化问题描述为

$$\max_{\boldsymbol{f},\boldsymbol{\phi}}\left|\left(\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{1}^{\mathrm{H}}\right)\boldsymbol{f}\right|^2 \tag{5}$$

$$\text{s.t.}\quad \left|\left(\boldsymbol{h}_{\mathrm{r2}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z}+\boldsymbol{h}_{2}^{\mathrm{H}}\right)\boldsymbol{f}\right|^2\geqslant\bar{E} \tag{5a}$$

$$\|\boldsymbol{f}\|^2\leqslant P \tag{5b}$$

$$\boldsymbol{\phi}=\mathrm{diag}\left(\mathrm{e}^{\mathrm{j}\theta_1},\mathrm{e}^{\mathrm{j}\theta_2},\cdots,\mathrm{e}^{\mathrm{j}\theta_M}\right) \tag{5c}$$

其中，式(5a)为ER能量约束条件，式(5b)为发射功率约束条件，式(5c)为IRS相移矩阵$\boldsymbol{\phi}$的约束条件。需要特别说明的是，如果信号采用QPSK或16PSK等调制信号[12]，优化问题仍然是式(5)。

由于IRS的发射波束成形向量$\boldsymbol{f}$和相移矩阵$\boldsymbol{\phi}$是耦合的，并且相移矩阵受单位膜约束如式(5c)，能量约束条件为非凸约束，优化问题更加具有挑战性。虽然用传统的非凸优化理论可以求解，但是时间复杂度高，特别是随着IRS反射单元数的增加，计算时间成本更高。为了解决上述非凸优化算法的瓶颈，本文基于深度强化学习，提出了一种时间复杂度低的DDPG优化算法。

2　基于DDPG的联合优化算法

2.1　DDPG算法框架

深度强化学习中，智能体通过与环境的交互学习，选择最优的行动来最大化预期的奖励。智能体与环境之间的交互过程用状态s、动作a、奖励r和策略π表示。设s_t、a_t、r_t、π_t、s_{t+1}和a_{t+1}分别表示t时刻的状态、动作、奖励、策略及下一时刻步长的状态和动作。强化学习中，采用q值表示智能体在策略π和状态s下采取行动a时获得的未来奖励总和，记作$q_\pi(s,a)$，q值可以采用表格法获得。然而，对连续的作用空间，智能体难以通过传统表格法获得q值。深度强化学习采用深度学习中近似函数的概念，通过深度神经网络（DNN, deep neural network）获得$q_\pi(s,a)$函数，从而代替表格法，获得q值。

DDPG算法框架如图2所示，包括4个DNN，即价值网络、策略网络、目标价值网络和目标策略网络，参数分别为ω_q、ω_μ、ω_q'和ω_μ'。DDPG算法的思想是采用DNN逼近策略函数，从而实现连续动作的预测和优化，并在该动作—状态对下价值函数输出相应的q值。具体来讲，状态s_t作为输入，策略网络输出动作a_t。状态s_t和动作a_t作为输入，价值网络输出q值，记作$q_\pi(s_t,a_t;\omega_q)$。状态s_{t+1}作为输入，目标策略网络输出动作a'_{t+1}。状态s_{t+1}和动作a'_{t+1}为输入，目标价值网络输出q'值，记为$q'_\pi(s_{t+1},a'_{t+1};\omega'_q)$。

为了使DDPG算法与环境交互学习，按如下方法构造经验回放D。在t时刻，深度强化学习智能体在状态s_t下采取行动，获得奖励r_t并进入下一个状态s_{t+1}，将转换元组(s_t,a_t,r_t,s_{t+1})存储到经验回放D中。

深度强化学习智能体的学习过程包括策略评估阶段和策略改进阶段。这两个阶段相互作用，可以获得具有最高q值的最优动作。

策略评估阶段，从经验回放D中采样第k个转换元组(s_k,a_k,r_k,s_{k+1})，状态s_k和动作a_k作为价值网络的输入，q值$q_\pi(s_k,a_k;\omega_q)$作为价值网络的输出，估策略π。在状态s_{k+1}下，目标策略网络输出动作a'_{k+1}。在状态s_{k+1}和动作a'_{k+1}，网络输出q值$q'_\pi(s_{k+1},a'_{k+1};\omega'_q)$。

策略改进阶段，更新4个DNN以获得实现更高q值的更优策略。具体地，基于r_k、$q_\pi(s_k,a_k;\omega_q)$和$q_\pi(s_{k+1},a'_{k+1};\omega'_q)$，通过最小化均方贝尔曼误差（MSBE, mean squared Bellman error）来更新价值网络的参数ω_q。通过梯度下降法，依据$q_\pi(s_k,a_k;\omega_q)$函数中的q值来更新参数ω_μ。目标价值网络参数ω_q'和目标策略网络参数ω_μ'分别通过软更新方法更新。

价值网络和策略网络的结构如图3所示，两个网络都包括一个输入层、一个输出层和两个隐藏层，价值网络有两个归一化层$L-1$、$L-2$。策略网络的输入维度等于状态集的基数加上动作集的基数，输出维度等于1。两个隐藏层的维度分别为L_1和L_2。

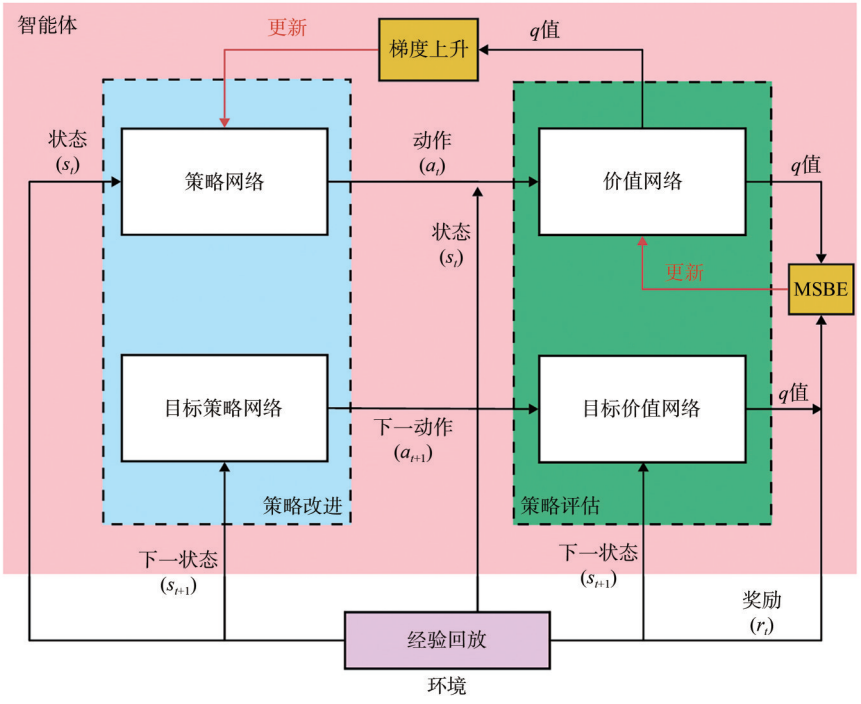

图2　DDPG算法框架

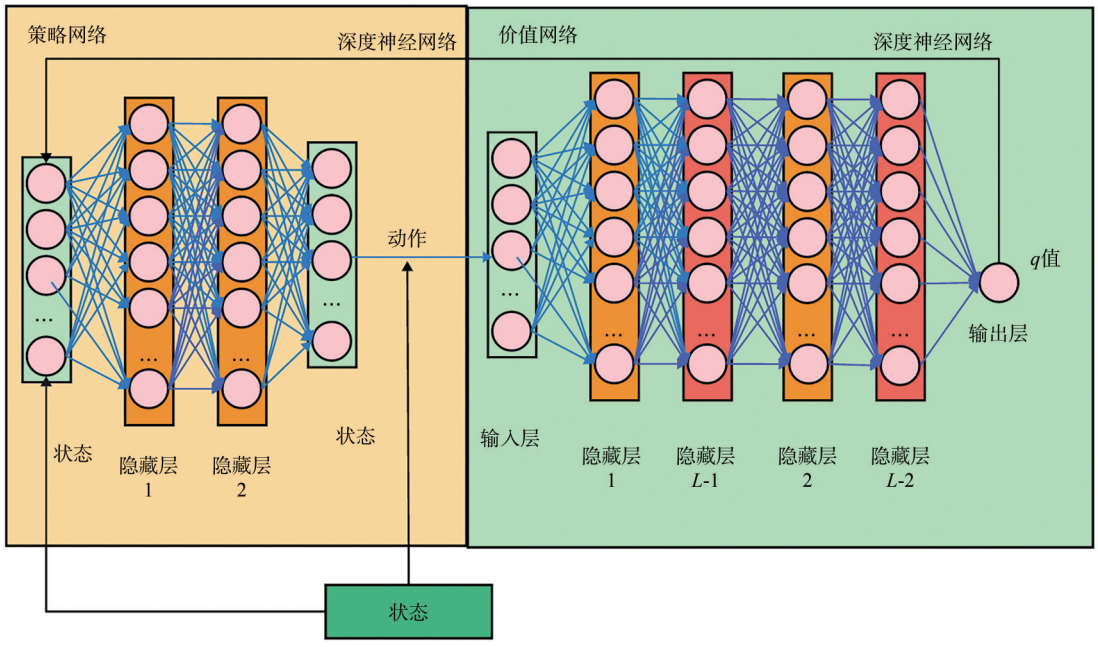

图3　价值网络和策略网络的结构

目标价值网络具有与价值网络相同的DNN结构。策略网络的输入维度等于状态集的基数，输出维度等于动作集的基数。目标策略网络具有与策略网络相同的DNN结构。

2.2　DDGP算法构建

下面将具体描述DDPG算法的状态（state）、动作（action）和奖励（reward）的构建。

状态（state）：t时刻的状态s_t由t时刻的发射功率、ER处的接收能量、$t-1$时刻的动作和所有通信链路的信道状态信息$\boldsymbol{H}_{\mathrm{G}}$组成，$\boldsymbol{H}_{\mathrm{G}}$如式(6)所示。

$$\boldsymbol{H}_{\mathrm{G}}=\left[\boldsymbol{h}_{\mathrm{r1}}^{\mathrm{H}},\boldsymbol{h}_{\mathrm{r2}}^{\mathrm{H}},\boldsymbol{h}_{1}^{\mathrm{H}},\boldsymbol{h}_{2}^{\mathrm{H}},\boldsymbol{Z}\right] \tag{6}$$

由于神经网络只能取实数而不能取复数作为输入，所以在构造状态s时，如果涉及复数，则实部和虚部将被分离为独立的输入端口。输入神经网络

的信道状态信息如式(7)

$$\boldsymbol{H}'_{\mathrm{G}} = [\,\mathrm{Re}(\boldsymbol{H}_{\mathrm{G}}), \mathrm{Im}(\boldsymbol{H}_{\mathrm{G}})\,] \tag{7}$$

状态s_t可以表示为式(8)

$$s_t = \left[a_{t-1}, P_t, E_t, \boldsymbol{H}'_{\mathrm{G}}\right] \tag{8}$$

其中，a_{t-1}为$t-1$时刻的动作，动作空间维度为$2N_t + M$。第二项和第三项为t时刻的发射功率和能量收集器接收到的能量，这两项都是实数动作空间，维度都为1，第四项为信道状态信息，因为$\boldsymbol{H}_{\mathrm{G}}$的维度为$2M + 2N_t + (M \times N_t)$，所以$\boldsymbol{H}'_{\mathrm{G}}$的维度为$2 \times \left(2M + 2N_t + \left(M \times N_t\right)\right)$，状态$s$的空间维度为$2N_t + M + \left(2M + 2N_t + \left(M \times N_t\right)\right) = 8N_t + 7M + 2$。

动作（action）：动作可以由$\boldsymbol{\phi}$的角度$\boldsymbol{\theta}$和波束成形向量$\boldsymbol{f}$表示。同样将实部和虚部分别输入神经网络，其中，波束成形向量$\boldsymbol{f}'$为

$$\boldsymbol{f}' = [\,\mathrm{Re}(\boldsymbol{f}), \mathrm{Im}(\boldsymbol{f})\,] \tag{9}$$

相移矩阵$\boldsymbol{\phi} = \left[\mathrm{e}^{\mathrm{j}\theta_1}, \mathrm{e}^{\mathrm{j}\theta_2}, \cdots, \mathrm{e}^{\mathrm{j}\theta_M}\right]$，所以角度$\boldsymbol{\theta}$可以表示为式(10)

$$\boldsymbol{\theta} = \left[\theta_1, \theta_2, \cdots, \theta_M\right] \tag{10}$$

其中，$\theta_M \in [0, 2\pi)$，所以t时刻的动作a_t为

$$a_t = \left[\boldsymbol{f}'_t, \boldsymbol{\theta}_t\right] \tag{11}$$

动作第一项的空间维度为$2N_t$，第二项的空间维度为M，所以动作的空间维度为$2N_t + M$。此外，为了便于智能体的探索，在策略网络输出的动作中加入了一个噪声κ，噪声κ的维度与动作空间的维度相同。所以，策略网络在t时刻的输出可以表示为

$$a_t = \pi\left(s_{t+1}; \omega_\mu\right) + \kappa \tag{12}$$

奖励（reward）：t时刻的奖励$r_t = \bar{I}\left(x; y_{\mathrm{IR}}\right)$，其中，$\bar{I}\left(x; y_{\mathrm{IR}}\right)$已在式(4)中给出，若ER处收集的能量$E = \left|\left(\boldsymbol{h}_{\mathrm{r2}}^{\mathrm{H}}\boldsymbol{\phi}\boldsymbol{Z} + \boldsymbol{h}_2^{\mathrm{H}}\right)\boldsymbol{f}\right|^2$不满足约束条件式(5a)，则奖励变更为$\bar{E} - E$，这是一个负奖励，鼓励智能体在不满足约束条件式(5a)的情况下使得动作输出接近最小收集能量$\bar{E}$，约束条件式(5b)采用归一化方法处理，约束条件式(5d)直接将$\boldsymbol{\theta}$限制在$[0, 2\pi)$。

2.3　神经网络参数的更新

1) 价值网络的更新：价值网络从经验回放D中采用一个小批量样本。将第t个元组$\left(s_t, a_t, r_t, s_{t+1}\right)$中的$s_t$和$a_t$输入价值网络，计算$q_t = \left(s_t, a_t; \omega_t\right)$得到$q_t$值。将$s_{t+1}$输入目标策略网络计算$a'_{t+1} = \pi\left(s_{t+1}; \omega'_\mu\right)$，将$a'_{t+1}$输入目标价值网络计算$q'_{t+1} = \left(s_{t+1}, a'_{t+1}; \omega'_q\right)$得到$q'_{t+1}$。此时可以计算损失函数为

$$\delta_t = q_t - \left(r_t + \gamma \times q'_{t+1}\right) \tag{13}$$

其中，$\gamma \in (0, 1]$表示价值的折扣率。使用梯度下降法更新价值网络的参数。

$$\omega_q = \omega_q - \alpha \times \delta_t \times \frac{\partial q_t = \left(s_t, a_t; \omega_q\right)}{\partial \omega_q} \tag{14}$$

其中，α为价值网络的学习率。

2) 策略网络的更新：策略网络$\pi\left(s; \omega_\mu\right)$的目标是更新$\omega_\mu$使价值网络输出的$q$值更大。所以需要计算$q$关于$\omega_\mu$的梯度如下

$$g = \frac{\partial q\left(s, \pi\left(s; \omega_\mu\right); \omega_q\right)}{\partial \omega_\mu} \tag{15}$$

式(15)可以展开为式(16)

$$g = \frac{\partial a}{\partial \omega_\mu} \times \frac{\partial q\left(s, a; \omega_q\right)}{\partial a} \tag{16}$$

计算梯度后做梯度上升来更新ω_μ，如式(17)

$$\omega_\mu = \omega_\mu + \beta \times g \tag{17}$$

其中，β为策略网络的学习率。

3) 目标价值网络和目标策略网络的更新：首先设计一个超参数$\tau \in (0, 1)$，目标价值网络$q'_t = \left(s_t, a'_t; \omega'_q\right)$和目标策略网络$\pi\left(s_t; \omega'_\mu\right)$的更新分别如式(18)和式(19)

$$\omega'_q = \tau \times \omega_q + (1-\tau) \times \omega'_q \tag{18}$$

$$\omega'_\mu = \tau \times \omega_\mu + (1-\tau) \times \omega'_\mu \tag{19}$$

基于DDPG的联合优化算法如算法1所示。

算法1　基于DDPG的联合优化算法

初始化：随机初始化价值网络和策略网络的参数ω_q、ω_μ。将价值网络和策略网络参数分别赋值给对应的目标网络即$\omega'_q = \omega_q$，$\omega'_\mu = \omega_\mu$。清空经验回放D，并将D的大小设置为D_{B}。初始化噪声分布$\varkappa$以进行动作探索。

输入：当前信道状态信息（CSI）矩阵$\boldsymbol{H}_{\mathrm{G}}$，噪声$\sigma$。

输出：通过策略网络获得最优动作a^*。

1) **for** episode j=1, 2, ⋯, J **do**

2) 重置状态s_0

3)　**for** time step t=1, 2, ⋯, T **do**

4)　通过式(12)获得a_t；

5)　通过式(5a)获得E_t；

6)　通过式(5b)获得P_t;

7)　获得给定动作a_t的下一状态s_{t+1}，并计算奖励r_t，然后将转移元组(s_t, a_t, r_t, s_{t+1})存储在经验回放D中;

8)　对经验回放D进行小批量采样，大小为N_B;

9)　通过式(14)更新价值网络的参数ω_q;

10)　通过式(17)更新策略网络的参数ω_μ;

11)　通过式(18)更新目标价值网络的参数ω'_q;

12)　通过式(19)更新目标策略网络的参数ω'_μ;

13)　将状态s_t赋值为s_{t+1};

14)　**end for**

15) **end for**

2.4　算法复杂度分析

价值网络中，输入层、第一全连接层、第二全连接层和输出层的维度分别为$10N_t + 8M + 2$、L_1、L_2和1。策略网络输入层、第一全连接层、第二全连接层和输出层的维度分别为$8N_t + 7M + 2$、L_1、L_2和$2N_t + M$。所以基于DDPG的算法复杂度为：

$$O\left[\begin{matrix}(10N_t + 8M + 2)L_1 + L_1L_2 + L_2 + \cdots \\ (8N_t + 7M + 2)L_1 + L_1L_2 + L_2(2N_t + M)\end{matrix}\right] \tag{20}$$

由式(20)可以看出，DDPG算法与发射天线数、反射单元数呈线性关系，计算复杂度较低。

3　仿真结果与分析

假设所有的信道状态信息都是完美已知的，IRS辅助的SWIPT MISO系统如图4所示[12]。

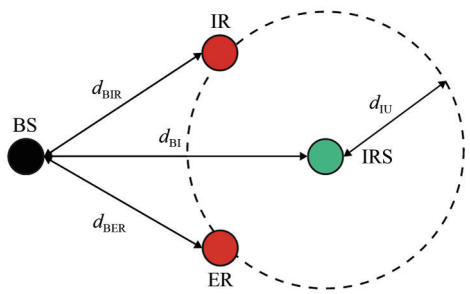

图4　IRS辅助的SWIPT MISO系统[12]

BS和IRS分别位于(0,0)和(50 m,0)，ER和IR两个用户在圆心为IRS、半径为d_{IU} = 30 m的圆上随机分布。假设所有信道都服从瑞利衰落分布，路径损耗模型为$PL = C_0(d/d_0)^{-\alpha}$，其中，$C_0$是参考距离$d_0$处的恒定路径损耗，$\alpha$是路径损耗指数，$d$表示通信链路的距离。仿真中，设定$C_0$ = 10 dB，d_0 = 10 m。路径损耗指数BS-IRS、IRS-IR、IRS-ER、BS-IR和BS-ER分别为α_{BI} = 3.0、α_{IIR} = 3.0、α_{IER} = 3.0、α_{BIR} = 3.5、α_{BER} = 3.5，信噪比SNR = $1/\sigma^2$，其中，σ^2为噪声功率。发射天线数N_t = 8，IRS的反射单元数量M = 10，BS与用户（IR和ER）之间的距离相等，即$d_{BIR} = d_{BER}$ = 40 m，最大发射功率P = 10 W，在ER处的最小收获能量$\bar{E}$ = 0.2 W。DDPG算法的超参数见表1。本文采用恒定学习率的策略网络和价值网络，学习率为0.000 1。

表1　　DDPG算法的超参数

超参数	参数值
价值网络学习率	α = 0.000 1
策略网络学习率	β = 0.000 1
软更新系数	τ = 0.005
奖励折扣因子	γ = 0.9
经验回放大小	D = 1 000 000
小批量大小	N_B = 128

1) 算法线上部署阶段：算法线上部署和其他3种不同方案都是在200个随机独立信道实现的平均值。M=10时4种方案下平均互信息与SNR的关系如图5所示，其中，AO-SDR为文献[12]提出的传统非凸优化方法。从图5中可以看出，本文所提出的DDPG算法性能优于随机IRS和无IRS辅助的系统性能，但略差于AO-SDR方案。虽然如此，DDPG算法可以通过离线训练的方式训练好权重参数，使得算法上线时可以用较少的时间学习到最优的策略，具有较低的时延，而AO-SDR计算复杂度较高，计算耗时长，对时延要求高的通信业务来说，本文所提的DDPG算法更适用。

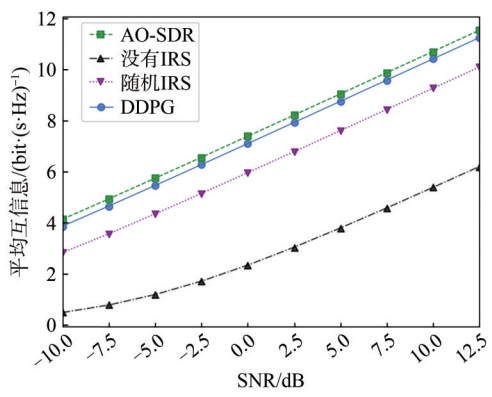

图5　M = 10时4种方案下平均互信息与SNR的关系

3种方案下平均互信息与反射单元数M的关系变化如图6所示，此实验取自文献[12]中的高斯输入的仿真结果。当M值增大时，用户的平均互信息也得到提升。此外，与图5结论一致，本文所提的DDPG算法的性能略低于AO-SDR算法，且随着M的增大，差距有所增加。这是因为随着M的增大，DDPG算法的动作空间也将增大，使得算法更加难以通过学习获得最优解。DDPG算法与AO-SDR算法运行时间对比如图7所示，从图7可以看出，随着M的增大，DDPG算法的运行时间远远低于AO-SDR算法，这说明，在性能和运行时间上存在折中，对于反射单元数量大、时延低的通信业务来说，在满足一定的传输速率的条件下，本文所提算法具有优越性。

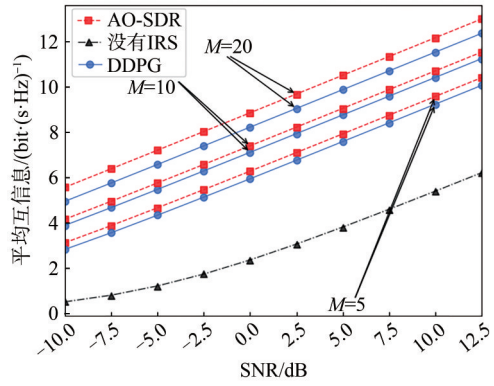

图6　3种方案下平均互信息与反射单元数M的关系变化

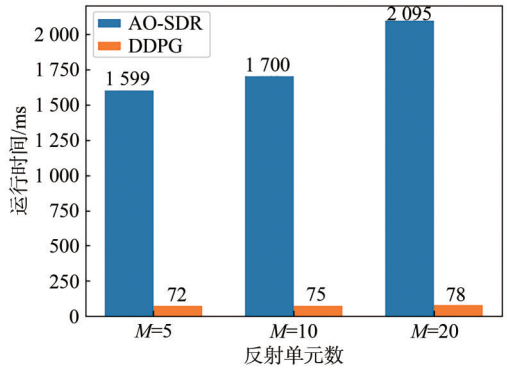

图7　DDPG算法与AO-SDR算法运行时间对比

2) 线下训练阶段：不同最大传输功率P下的奖励与步长关系如图8所示，考虑P = 60 W、P = 40 W、P = 20 W和P = 5 W这4种情况，可以看出，随着时间步长t的增加，不同P值下的奖励都会增加并收敛，这证明了算法的有效性。该算法在低信噪比下比高信噪比下收敛快。原因在于，信噪比越高，即时奖励的波动幅度越大，从而得到更差的收敛性。

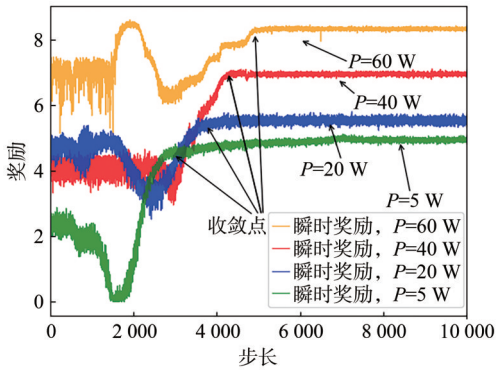

图8　不同最大传输功率P下的奖励与步长关系

DDPG算法与学习率的关系如图9所示，图9给出了学习率对DRL模型的性能和收敛速度的影响曲线。在模拟中，使用式(21)来计算平均奖励。

$$\text{AverageReward}(K_i)=\frac{\sum_{k=1}^{K_i}\text{reward}(k)}{K_i},\quad K_i=1,2,\cdots,K \tag{21}$$

其中，K是最大步长。

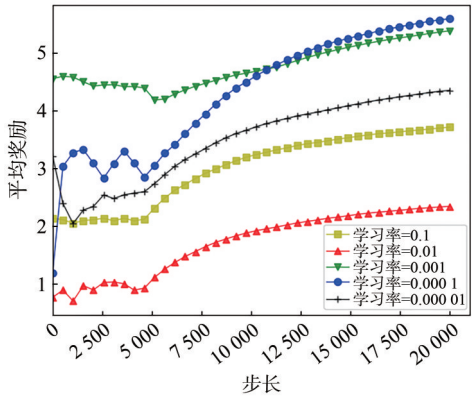

图9　DDPG算法与学习率的关系

从图9可以看出，不同的学习率对DDPG算法的性能有很大的影响。当学习率取0.1和0.01时，平均奖励较少，原因可能是较高的学习率导致算法持续震荡无法找到最优解。当学习率取0.001、0.000 1和0.000 01时，得到了较高的平均奖励，学习率为0.000 1的平均奖励最高，所以本文选取0.000 1作为本模型的学习率。

4　结束语

本文针对IRS辅助的MISO SWIPT系统，以最大化系统的平均互信息为目标，采用深度强化学习算法优化问题进行转化，提出基于DDPG的发射波

束成形和IRS相移矩阵联合优化算法，通过观测回报来学习环境，逐步改进其行为获得，从而达到与传统优化方法相同的性能。仿真结果表明，传统凸优化算法的计算时间是DDPG算法的22倍以上，说明DDPG算法时间复杂度远低于传统优化算法，但系统的平均互信息非常接近。此外，本文还研究了最大传输功率、最佳学习率对算法性能的影响。结果表明，传输功率越小，算法的收敛速度越快。过高或过低的学习率也会使DDPG算法的性能下降，因此，本文选择0.000 1作为DDPG模型的学习率。从复杂度分析结果看，DDPG算法的计算复杂度与天线数、反射单元数呈线性关系，时间复杂度较低，时延较小，特别适用于对时延要求高的通信系统。下一步继续研究将DDPG算法扩展到MIMO SWIPT系统，并进一步提高算法的收敛速度。

参考文献：

[1] 齐峰, 岳殿武, 孙玉. 面向6G的智能反射面无线通信综述[J]. 移动通信, 2022, 46(4): 65-73.
QI F, YUE D W, SUN Y. A survey of intelligent reflecting surface wireless communications toward 6G[J]. Mobile Communications, 2022,46 (4): 65-73.

[2] WU Q, ZHANG R. Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network[J]. IEEE Communications Magazine, 2020, 58(1): 106-112.

[3] 朱政宇, 王梓晅, 徐金雷, 等. 智能反射面辅助的未来无线通信：现状与展望[J]. 航空学报, 2022, 43 (2): 203-217.
ZHU Z Y, WANG Z X, XU J L, et al. Future wireless communication assisted by intelligent reflecting surface: state of art and prospects [J]. Acta Aeronautica et Astronautica Sinica, 2022, 43 (2) : 203-217.

[4] ALLAHZADEH S, DANESHIFAR E. Simultaneous wireless information and power transfer optimization via alternating convex-concave procedure with imperfect channel state information[J]. Signal Processing, 2021, 182: 107953.

[5] VARSHNEY L R. Transporting information and energy simultaneously[C]//Proceedings of the 2008 IEEE International Symposium on Information Theory. Piscataway: IEEE Press, 2008: 1612-1616.

[6] ZHANG R, HO C K. MIMO broadcasting for simultaneous wireless information and power transfer[J]. IEEE Transactions on Wireless Communications, 2013, 12(5): 1989-2001.

[7] XU J, LIU L, ZHANG R. Multiuser MISO beamforming for simultaneous wireless information and power transfer[J]. IEEE Transactions on Signal Processing, 2014, 62(18): 4798-4810.

[8] XIANG Z, TAO M. Robust beamforming for wireless information and power transmission[J]. IEEE Wireless Communications Letters, 2012, 1(4): 372-375.

[9] 马柱华, 罗丽平. 非理想顺序干扰消除和信道状态信息下SWIPT-NOMA-CR网络中断性能[J]. 物联网学报, 2023, 7 (1): 129-139.
MA Z H, LUO L P. Outage performance of SWIPT-NOMA-CR network with imperfect SIC and CSI[J]. Chinese Journal on Internet of Things, 2023, 7(1): 129-139.

[10] 王玉俊, 罗丽平. 基于无线携能和非正交多址接入的认知中继网络中断性能分析[J]. 中山大学学报(自然科学版)(中英文), 2023, 62 (1): 169-180.
WANG Y J, LUO L P. Outage performance analysis for cognitive relay networks based on SWIPT and NOMA[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2023, 62 (1): 169-180.

[11] TANG Y, MA G, XIE H, et al. Joint transmit and reflective beamforming design for IRS-assisted multiuser MISO SWIPT systems[C]// Proceedings of the ICC 2020-2020 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2020: 1-6.

[12] LIU Z, ZHU X, CHEN B, et al. Joint transmission design for IRS-assisted MISO SWIPT systems[J]. Signal Process, 2022, 200: 108649.

[13] SHI Q, LIU L, XU W, et al. Joint transmit beamforming and receive power splitting for MISO SWIPT systems[J]. IEEE Transactions on Wireless Communications, 2014, 13(6): 3269-3280.

[14] ZHU Z, XU J, SUN G, et al. Robust beamforming design for IRS-aided secure SWIPT terahertz systems with non-linear EH model[J]. IEEE Wireless Communications Letters, 2022, 11(4): 746-750.

[15] ZHU Z, LI Z, CHU Z, et al. Intelligent reflecting surface-assisted wireless powered heterogeneous networks[J]. IEEE Transactions on Wireless Communications, 2023, 22(12):9881-9892.

[16] ZHU Z, MA M, SUN G, et al. Secrecy rate optimization in nonlinear energy harvesting model-based mmWave IoT systems with SWIPT[J]. IEEE Systems Journal, 2022, 16(4): 5939-5949.

[17] NIU H, CHU Z, ZHOU F, et al. Robust design for intelligent reflecting surface-assisted secrecy SWIPT network[J]. IEEE Transactions on Wireless Communications, 2022, 21(6): 4133-4149.

[18] ZHU Z, LI Z, CHU Z, et al. Resource allocation for intelligent reflecting surface assisted wireless powered IoT systems with power splitting[J]. IEEE Transactions on Wireless Communications, 2022, 21(5): 2987-2998.

[19] 朱政宇, 徐金雷, 孙钢灿, 等. 基于IRS辅助的SWIPT物联网系统安全波束成形设计[J]. 通信学报, 2021, 42 (4): 185-193.
ZHU Z Y, XU J L, SUN G C, et al. Secure beamforming design for IRS-assisted SWIPT Internet of things system[J]. Journal on Communications, 2021, 42 (4):185-193.

[20] LEE H, LEE K J, KIM H, et al. Joint transceiver optimization for MISO SWIPT systems with time switching[J]. IEEE Transactions on Wireless Communications, 2018, 17(5): 3298-3312.

[21] ATAPATTU S, FAN R, DHARMAWANSA P, et al. Reconfigurable intelligent surface assisted two-way communications: perfor-

mance analysis and optimization[J]. IEEE Transactions on Communications, 2020, 68(10): 6552-6567.

[22] THRUN S, LITTMAN M L. Reinforcement learning: an introduction[J]. AI Magazine, 2000, 21(1): 103-103.

[23] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge: MIT press, 2016.

[24] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[25] HUANG C, ALEXANDROPOULOS G C, ZAPPONE A, et al. Deep learning for UL/DL channel calibration in generic massive MIMO systems[C]//Proceedings of the ICC 2019-2019 IEEE International Conference on Communications (ICC). Piscataway: IEEE Press, 2019: 1-6.

[26] LUONG N C, HOANG D T, GONG S, et al. Applications of deep reinforcement learning in communications and networking: a survey[J]. IEEE Communications Surveys & Tutorials, 2019, 21(4): 3133-3174.

[27] FENG K, WANG Q, LI X, et al. Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems[J]. IEEE Wireless Communications Letters, 2020, 9(5): 745-749.

[28] HUANG C, MO R, 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.

[29] SHEHAB M, CIFTLER B S, KHATTAB T, et al. Deep reinforcement learning powered IRS-assisted downlink NOMA[J]. IEEE Open Journal of the Communications Society, 2022, 3: 729-739.

[30] LIN J, ZOU Y, DONG X, et al. Deep reinforcement learning for robust beamforming in IRS-assisted wireless communications[C]// Proceedings of the GLOBECOM 2020-2020 IEEE Global Communications Conference. Piscataway: IEEE Press, 2020: 1-6.

[作者简介]

罗丽平（1980– ），女，博士，广西民族大学电子信息学院教授、博士生导师，主要研究方向为新一代无线通信技术。

潘伟民（1999– ），男，广西民族大学电子信息学院硕士生，主要研究方向为智能反射面、无线携能通信和深度强化学习技术。