深度强化学习下连续和离散相位 RIS 毫米波通信

胡浪涛*，杨 瑞，刘全金，吴建岚，嵇 文，吴 磊

（安庆师范大学 电子工程与智能制造学院，安庆 246133）

摘要 在分布式智能反射面 (RIS) 辅助多用户毫米波 (mmWave) 系统中，利用深度强化学习 (DRL) 理论学习并调整基站发射波束赋形矩阵和 RIS 相位偏转矩阵，联合优化发射波束赋形和相位偏转，实现加权和速率最大化。即在离散动作空间中，设计了功率码本与相位码本，提出了用深度 Q 网络 (DQN) 算法进行优化发射波束赋形与 RIS 相位偏转矩阵；在连续动作空间中，采用双延迟策略梯度 (TD3) 算法进行优化发射波束赋形与 RIS 相位偏转矩阵。仿真分析比较了在不同码本比特数下离散动作空间和连续动作空间下系统的加权和速率。与传统的凸优化算法以及迫零波束赋形随机相位偏转算法进行了对比，强化学习算法的和速率性能有明显提升，连续的 TD3 算法的和速率超过凸优化算法 23.89%，在码本比特数目为 4 时，离散的 DQN 算法性能也优于传统的凸优化算法。

关键词 深度 Q 网络 (DQN)；深度强化学习；双延迟策略梯度；毫米波；智能反射面

中图分类号 TN928 **文献标志码** A DOI 10.12178/1001-0548.2022285

Continuous vs Discrete: Phase Performance Comparison of RIS-Assisted Millimeter Wave Communication Based on Deep Reinforcement Learning

HU Langtao*, YANG Rui, LIU Quanjin, WU Jianlan, JI Wen, and WU Lei

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

Abstract In this paper, in the distributed Reconfigurable Intelligence Surface (RIS) assisted multi-user millimeter wave (mmWave) system, the deep reinforcement learning (DRL) theory is used to learn and adjust transmit beamforming matrix at the base station and phase shift matrix at the RIS, and jointly optimize the transmit beamforming matrix and phase shift matrix to maximize the weighted sum-rate. Specifically, in the discrete action space, we first design the power codebook and the phase codebook, and propose the Deep Q Network(DQN) algorithm to optimize the beamforming matrix and phase shift matrix; then, in the continuous action space, the Twin Delayed Deep Deterministic (TD3) policy gradient algorithm is used to optimize the beamforming matrix and phase shift matrix. The weighted sum-rates of the system in discrete action space and continuous action space with different number of codebook bits are compare through simulation. In addition, compared with the traditional convex optimization algorithm and the zero-forcing precoding with a random PBF algorithm, the sum-rate performance of DRL algorithm is significantly improved, and the sum-rate of the continuous TD3 algorithm exceeds the convex optimization algorithm by 23.89%, and the performance of the discrete DQN algorithm exceeds the traditional convex optimization algorithm when the number of codebook bits is 4.

Key words deep Q network(DQN); deep reinforcement learning; delayed deep deterministic policy gradient; millimeter wave; reconfigurable intelligence surface

与传统的 6 GHz 以下的通信相比，具有千兆赫带宽可用性的毫米波（Millimeter Wave, mmWave）通信具有更高的容量和传输速率[1-2]。但毫米波信号的传输距离较短，且易受到障碍物的影响。因此，引入智能反射面（Reconfigurable Intelligence Surface, RIS）来增强毫米波信号的传输和接收。与有源放大转发中继不同，RIS 基本由无源反射元件组成，没有 RF 射频单元，具有低成本、低功耗、可编

收稿日期：2022−08−22；修回日期：2023−05−29

基金项目：国家自然科学基金（62171002）；安徽省教育厅自然科学基金（KJ2020A0497）

作者简介：胡浪涛，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 E-mail: hulangtao@aqnu.edu.cn

程、易部署等特点[3]。此外，RIS 的每个智能超表面单元可以调整其振幅和相位参数，以增强基站（Base Station, BS）的输入信号实时反射给用户，从而经济有效地提高网络性能[3-6]。

最近，RIS 辅助通信的场景已得到了广泛的关注[7-8]。文献 [7] 研究了 RIS 辅助的无人机通信系统的物理层安全。文献 [8] 将 RIS 部署到多用户 MIMO 通信中，并提出了一个基于并行因子分解的信道估计框架，以展开所产生的级联信道模型。RIS 处的无源波束赋形可由 BS 通过 RIS 控制。因此，为了使 RIS 的增益最大化，基站和 RIS 的波束赋形通常是联合设计的[9-10]。文献 [9-10] 的波束赋形设计均为连续相位；在文献 [11-13] 中，波束赋形设计问题被推广到离散相位，其中文献 [11-12] 研究了 RIS 处的离散反射波束赋形，文献 [13] 研究了基站处的离散发射波束赋形。大多数研究假设 BS 和 RIS 之间存在丰富的散射[15-16]，但涉及毫米波传输时，应考虑低阶 BS-RIS 信道[9]。文献 [17] 从 mmWave 的角度研究了 RIS 的潜在应用，其中弱 BS—用户链路可通过 RIS 的反射增益进行补偿。

上述研究主要是通过传统凸优化算法来解决 RIS 的波束赋形问题，而传统凸优化算法求解问题时大多采用交替迭代的方式，求解的结果强烈依赖于初始值，且计算复杂性会因通信的复杂度增加而急速增加，对大规模系统效率较低。受深度强化学习（Deep Reinforcement Learning, DRL）可解决无线通信中具有非凸特性的复杂问题、允许通信实体学习、能够提供自主决策以及对高维数据处理等优点的启发，一些研究者尝试利用 DRL 来解决无线通信中的一些问题[13-14, 18-20]。文献 [13] 研究了同构蜂窝网络中干扰信道的信道容量，利用 DRL 提出了一种分布式动态下行波束赋形协调方法，并根据码本设计了离散化的基站发射波束赋形矩阵。文献 [14] 研究了基于 DRL 的多小区非正交多址接入（Non-orthogonal Multiple Access, NOMA）能效优化功率分配问题。文献 [18] 研究了基于 DRL 的异构蜂窝网络中用户关联与资源分配。文献 [19-20] 分别研究了基于 DRL 的 RIS 辅助多用户多输入单输出系统和 RIS 辅助隐蔽通信系统，并且均利用 DRL 联合设计基站发射波束赋形与 RIS 相位偏转矩阵，以提高系统性能。然而，文献 [13] 虽引入了 RIS，但仅研究了基站处的离散发射波束赋形；文献 [18] 并没有引入 RIS 这一先进技术；文献 [19-20] 研究的联合设计均为连续波束赋形。此前的 RIS 辅助 mmWave 通信系统中，基于码本的离散波束赋形向量和离散相位的联合设计还未被研究。现阶段，大多数研究还是围绕连续的算法，但使用离散的算法也有其优点，离散算法的复杂度低，且连续相位和离散相位的性能对比也有很重要的意义。

基于上述研究背景，本文研究了在无直视链路的场景下分布式 RIS 辅助多用户 mmWave 通信系统，目标是实现最大化加权和速率。本文基于 DRL 提出两种联合优化方法，一种是基于深度 Q 网络（Deep Q Network, DQN）算法的离散化发射波束赋形和相位偏转矩阵联合优化方法，另一种是基于双延迟策略梯度（Delayed Deep Deterministic Policy Gradient, TD3）算法的连续发射波束赋形和相位偏转矩阵联合优化方法。本文主要研究工作如下：

1）基于 DRL 的 RIS 辅助多用户 mmWave 通信系统中，采用离散的动作空间，设计了功率码本和相位码本，通过 DQN 算法设计了发射波束赋形和相位偏转矩阵联合优化算法，实现最大化加权和速率；

2）基于 DRL 的 RIS 辅助多用户 mmWave 通信系统中，采用连续的动作空间，通过 TD3 算法设计了发射波束赋形和相位偏转矩阵联合优化算法，实现最大化加权和速率；

3）对比分析离散动作空间和连续动作空间的 DRL 算法的系统和速率、两种算法的复杂度，以及与传统凸优化算法、迫零随机波束赋形算法进行了仿真对比分析。

1　系统模型与问题公式化

1.1　系统模型

本文研究了图 1 所示的 RIS 辅助的多用户 mmWave 通信系统模型。在 RIS—用户的直视链路被障碍物阻塞的情况下，该系统利用 RIS 的反射功能协助 BS 与用户之间进行通信。BS 配备由N个天线组成的均匀线性阵列（Uniform Linear Array, ULA），系统设置G个 RIS 单元服务于K个单天线移动用户。每个 RIS 单元配备有M个无源反射元件组成的统一平面阵列（Uniform Planar Array, UPA）。令$M = M_{\mathrm{az}} \times M_{\mathrm{el}}$，其中$M_{\mathrm{az}}$为 RIS 单元水平方向的元件数，$M_{\mathrm{el}}$为 RIS 单元垂直方向的元件数。BS 的所有天线同时传输K个数据流，信号首先到达 RIS 反射面，再被反射面反射到用户。本文假设从 BS 到第g个 RIS 单元的毫米波信道矩阵为$\boldsymbol{W}_g \in \mathbb{C}^{M \times N}$，

第g个 RIS 单元到第k个用户之间的信道矢量为$\boldsymbol{h}_{g,k}\in\mathbb{C}^{M\times 1}$。

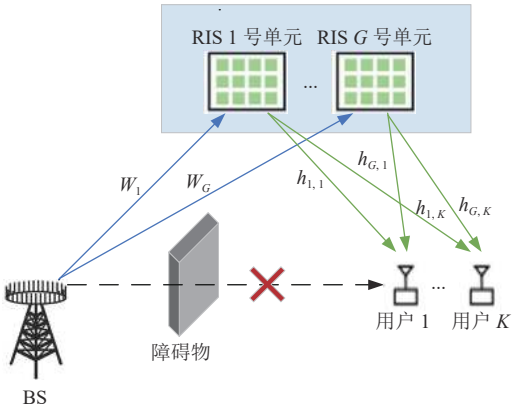

图 1 RIS 辅助的多用户毫米波通信系统模型

第k个用户处的接收信号为：

$$y_k=\sum_{g=1}^{G}\boldsymbol{h}_{g,k}^{\mathrm{H}}\boldsymbol{\Phi}_g^{\mathrm{H}}\boldsymbol{p}_k s_k+\sum_{g=1}^{G}\boldsymbol{h}_{g,k}^{\mathrm{H}}\boldsymbol{\Phi}_g^{\mathrm{H}}\boldsymbol{W}_g\sum_{j=1,j\neq k}^{K}\boldsymbol{p}_j s_j+u_k \tag{1}$$

式中，第g个 RIS 单元处的相位偏转矩阵为$\boldsymbol{\Phi}_g=\sqrt{\eta}\mathbf{diag}\left(\left[\theta_{g,1},\cdots,\theta_{g,M}\right]^{\mathrm{H}}\right)\in\mathbb{C}^{M\times M}$；$\eta$是反射系数，$\theta_{g,m}=\mathrm{e}^{\mathrm{j}\varphi_{g,m}}$，$\varphi_{g,m}\in[0,2\pi)$为 RIS 元件的相位偏移；$s_k$为 BS 对第$k$个用户的发射信号；$u_k$是方差为$\sigma_u^2$的循环对称复高斯噪声；$\boldsymbol{P}=[\boldsymbol{p}_1,\cdots,\boldsymbol{p}_k]\in\mathbb{C}^{N\times K}$为 BS 的波束赋形矩阵。

为了保持 BS 的发射功率，添加如下约束条件：

$$\mathbb{E}\left\{\mathbf{tr}\left(\boldsymbol{P}\boldsymbol{P}^{\mathrm{H}}\right)\right\}\leqslant P_t \tag{2}$$

式中，P_t是 BS 允许的总发射功率。

根据广泛应用的三维 Saleh-Valenzuela 信道模型[21]，从 BS 到第g个 RIS 单元的毫米波信道矩阵$\boldsymbol{W}_g$可以表述为：

$$\boldsymbol{W}_g=\sum_{\ell=0}^{N_p}v^{(\ell)}\boldsymbol{\alpha}_B(\phi_B^{(\ell)})\boldsymbol{\alpha}_I^{\mathrm{H}}(\phi_I^{(\ell)},\theta_I^{(\ell)}) \tag{3}$$

式中，N_p表示非视距 (non line of sight, NLoS) 路径的数量；$\ell=0$表示视距 (line of sight, LoS) 路径，$v^{(\ell)}$是第ℓ路径的复增益。$\theta^{(\ell)}$和$\phi^{(\ell)}$分别表示二维 RIS 的仰角和方位角；$\boldsymbol{\alpha}_B(\phi)=\frac{1}{\sqrt{N}}[\mathrm{e}^{-\mathrm{j}\frac{2\pi d}{\lambda}\phi_i}]_{i\in I(N)}$，$\boldsymbol{\alpha}_I(\phi,\theta)=\boldsymbol{\alpha}_I^{\mathrm{az}}(\phi)\otimes\boldsymbol{\alpha}_I^{\mathrm{el}}(\theta)$为 ULA 和 UPA 的阵列导向矢量，$\alpha_I^{\mathrm{az}}(\phi)=\frac{1}{\sqrt{N}}[\mathrm{e}^{-\mathrm{j}\frac{2\pi d}{\lambda}\phi_i}]_{i\in I(N)}$，$\alpha_I^{\mathrm{el}}(\theta)=\frac{1}{\sqrt{N}}[\mathrm{e}^{-\mathrm{j}\frac{2\pi d}{\lambda}\theta_i}]_{i\in I(N)}$，$\lambda$为载波波长，$d$为天线间距，$I(N)=\left\{n-\frac{N-1}{2},n=0,1,\cdots,N-1\right\}$。本文假设 ULA 和 UPA 的阵元间距都为$\lambda/2$。

通常情况下，由于路径损耗严重，可以忽略两个以上反射的发射功率，只考虑 LoS[15]。因此，第g个 RIS 单位与第k个用户之间的信道为：

$$\boldsymbol{h}_{g,k}=\sqrt{M}v_k\vartheta_r\vartheta_t\boldsymbol{\alpha}_t(\phi_k) \tag{4}$$

式中，v_k为信道增益；ϑ_r与ϑ_t分别为接收、发射天线单元增益；$\boldsymbol{\alpha}_t$为 RIS 的阵列导向矢量。本文假设涉及的所有信道状态信息都是完全已知的。

1.2 优化问题建模

本文旨在通过联合优化波束赋形矩阵$\boldsymbol{P}$与相位偏转矩阵$\boldsymbol{\Phi}_g$来最大化所提系统的下行总和速率。则第k个用户处的干扰信噪比为：

$$\rho_k=\frac{\left|\sum_{g=1}^{G}\boldsymbol{h}_{g,k}^{\mathrm{H}}\boldsymbol{\Phi}_g^{\mathrm{H}}\boldsymbol{W}_g\boldsymbol{p}_k\right|^2}{\sum_{j=1,j\neq k}^{K}\left|\sum_{g=1}^{G}\boldsymbol{h}_{g,k}^{\mathrm{H}}\boldsymbol{\Phi}_g^{\mathrm{H}}\boldsymbol{W}_g\boldsymbol{p}_j\right|^2+\sigma_u^2} \tag{5}$$

因此，本系统的最大和速率问题可以表示为：

$$\max_{\boldsymbol{\Phi}_g,\boldsymbol{P}}\sum_{k=1}^{K}\omega_k\log_2(1+\rho_k) \tag{6}$$

$$\text{s.t.}\quad \mathbf{tr}\{\boldsymbol{P}\boldsymbol{P}^{\mathrm{H}}\}\leqslant P_t \tag{6.1}$$

$$\theta_{g,m}\in\mathcal{F}_c,\forall_g,\forall_m \tag{6.2}$$

式中，ω_k为第k个用户数据的权重；P_t为 BS 所允许的最大发射功率。式 (6.1) 为通信施加的功率约束；式 (6.2) 为对 RIS 施加的相位约束，连续相位集合为$\mathcal{F}_c=\left\{\theta_{g,m}=\mathrm{e}^{\mathrm{j}\varphi_{g,m}}\middle|\varphi_{g,m}\in[0,2\pi)\right\}$，离散相位集合为$\mathcal{F}_c=\left\{\theta_{g,m}=\mathrm{e}^{\mathrm{j}\varphi_{g,m}}\middle|\varphi_{g,m}\in\left\{\frac{2\pi i}{2^B}\right\}_{i=0}^{2^B-1}\right\}$，$B$是以 bit 数表示的相位分辨率。

2 深度强化学习 DRL

2.1 DRL 概述

DRL 是指智能体通过“试错”的方式与环境进行交互学习，并自动调整策略以找到最佳动作，达到最优性能的学习过程[22]。以下是一些用于完整描述 DRL 学习过程的基本元素：状态、动作、奖励、策略和价值函数。

1) 状态：令$\mathcal{S}$表示所有可能存在状态的集合，则$S_t\in\mathcal{S}$表示 t 时刻观测到的状态；

2) 动作：令$\mathcal{A}$表示所有可能动作的集合，则$A_t\in\mathcal{A}$表示 t 时刻得到的动作；

3) 奖励：奖励R_t反映了智能体从当前状态S_t转

移到下一个状态S_{t+1}后行为表现的好坏。智能体的目标是最大化回报$G_t=\sum_{\tau=0}^{\infty}\gamma^{\tau}R_{t+\tau+1}$，$\gamma\in(0,1]$为奖励的折扣率；

4) 策略：策略$\pi(a|s)=\mathbb{P}(A_t=a|S_t=s)$表示在输入状态为$s$的情况下采取动作$a$的概率；

5) 价值函数：价值函数$Q(s,a)=\mathbb{E}_{\pi}[G_t|S_t=s,A_t=a]$是对未来奖励的预测，用于评估动作选择对于学习过程获得的预期未来累积折扣奖励的影响。Q 函数满足贝尔曼期望方程[19]：

$$\begin{aligned}&Q_{\pi}(S_t,A_t)=\mathbb{E}_{\pi}[R_{t+1}|S_t=s,A_t=a]+\\&\gamma\mathbb{E}_{\pi}[G_{t+1}|S_t=s,A_t=a]=\mathbb{E}_{\pi}[R_{t+1}|S_t=s,A_t=a]+\\&\gamma\sum_{S'\in\mathcal{S}}P(S'|s,a)\sum_{A'\in\mathcal{A}}\pi(A'|S')Q_{\pi}(S',A')\end{aligned}\tag{7}$$

因此，满足最优策略π^*的最优 Q 函数表达式为：

$$\begin{aligned}&Q^*(S_t,A_t)=R_{t+1}(S_t=s,A_t=a,\pi=\pi^*)+\\&\gamma\sum_{S'\in\mathcal{S}}P(S'|s,a)\max_{A'\in\mathcal{A}}Q^*(S',A')\end{aligned}\tag{8}$$

Q 函数的更新[23]为：

$$\begin{aligned}&Q^*(S_t,A_t)\leftarrow Q^*(S_t,A_t)+\mu[R_{t+1}+\\&\gamma\max_{A'\in\mathcal{A}}Q^*(S_{t+1},A')-Q^*(S_t,A_t)]\end{aligned}\tag{9}$$

式中，$\mu\in(0,1]$为训练网络的学习率，用于更新 Q 函数。$R_{t+1}+\gamma\max_{A'\in\mathcal{A}}Q^*(S_{t+1},A')$为时序差分 (temporal-difference, TD) 目标，$Q^*(S_t,A_t)$要向 TD 目标靠近。在 off-policy 的 DRL 算法中，将训练过程中产生的每一步四元组数据(S_t,A_t,R_t,S_{t+1})存放到经验回放缓冲池$\mathcal{M}$中，训练网络时每次从$\mathcal{M}$中随机采样一批数据进行训练。

2.2　深度 Q 网络 (DQN)

DQN 算法的最终更新目标是让$Q^*(S_t,A_t)$逼近 TD 目标。对于一组数据(S_t,A_t,R_t,S_{t+1})，Q 网络的损失函数为：

$$\mathcal{L}(\theta)=\arg\min_{\theta}\frac{1}{2}\left\|Q_{\theta}(S_t,A_t)-\left(R_{t+1}+\gamma\max_{A'\in A}Q_{\theta'}(S_{t+1},A')\right)\right\|^2\tag{10}$$

式中，θ与θ'分别为评价网络和目标网络的参数。评价网络用来计算原来的损失函数中的$Q_{\theta}(S_t,A_t)$项。目标网络用来计算原来的损失函数中的$Q_{\theta'}(S_{t+1},A')$项。为了让更新目标更稳定，目标网络会使用评价网络的一套比较旧的参数。评价网络在训练中实时更新，而目标网络的参数按照预定的频率与评价网络同步一次，即$\theta'\leftarrow\theta$。

2.3　双延迟确定策略梯度 (TD3)

TD3 算法中每个 Actor 网络都伴有两套 Critic 网络用于估算 Q 值，并取相对较小的值作为更新目标，其目的是缓解自举和最大化造成的高估。Critic 网络的 TD 目标为：

$$\begin{aligned}Q_{\text{t arg et}}=R_{t+1}+\gamma\min(&Q_{\theta'_1}(S_{t+1},\pi_{\varphi'}(S_{t+1})),\\&Q_{\theta'_2}(S_{t+1},\pi_{\varphi'}(S_{t+1})))\end{aligned}\tag{11}$$

式中，$Q_{\theta'_1}$与$Q_{\theta'_2}$分别为两个目标评价网络，θ'_1与θ'_2分别为对应目标网络的参数；$\pi_{\varphi'}$为目标策略网络；φ'为网络参数。为了使得预估更加准确，网络更具有健壮性，本文在计算Q_{target}时，为目标策略网络输出的动作添加一个 Ornstein-Uhlenbeck(OU) 噪声，则式 (11) 可变换为：

$$\begin{aligned}Q_{\text{target}}=R_{t+1}+\gamma\min(&Q_{\theta'_1}(S_{t+1},\pi_{\varphi'}(S_{t+1})+\varepsilon),\\&Q_{\theta'_2}(S_{t+1},\pi_{\varphi'}(S_{t+1}))+\varepsilon)\end{aligned}\tag{12}$$

式中，ε为 OU 噪声。因此，Critic 网络的更新[24]为：

$$\theta_i\leftarrow\theta_i-\alpha\delta_i\Delta_{\theta}Q(S_t,A_t;\theta_i)\quad i=1,2\tag{13}$$

$$\delta_i=Q_i-Q_{\text{target}}\quad i=1,2\tag{14}$$

$$Q_i=Q(S_t,A_t;\theta_i)\quad i=1,2\tag{15}$$

式中，α为更新训练 Critic 网络的学习率；$\delta_{i,t}$为 TD 误差；$Q_{i,t}$为 Critic 网络做的预测。此外，为了解决值函数与策略耦合问题，TD3 算法还采用延迟更新，确保 TD 误差足够小。Actor 网络的更新[24]为：

$$\varphi\leftarrow\varphi+\beta\nabla_{\varphi}\pi(S_t;\varphi)\nabla_A Q(S_t;\pi(S_t;\varphi);\theta_1)\tag{16}$$

式中，β为更新训练 Actor 网络的学习率。在降低频率更新的同时，使用软更新[24]：

$$\theta'_i\leftarrow\tau\theta_i+(1-\tau)\theta'_i\quad i=1,2\tag{17}$$

$$\varphi'\leftarrow\tau\varphi+(1-\tau)\varphi'\tag{18}$$

式中，τ为更新目标评价网络和目标策略网络的学习率。

3　基于 DRL 的发射波束赋形与相位偏转联合设计

本节主要介绍两种联合设计算法，一种是在离散动作空间中，利用图 2 所示的 DQN 神经网络结构，设计了功率码本与相位码本，提出了用 DQN 算法进行联合优化发射波束赋形与 RIS 相位偏转矩阵；另一种是在连续动作空间中，利用图 3 所示的 TD3 神经网络结构，采用 TD3 算法进行联合优化发射波束赋形与 RIS 相位偏转矩阵。

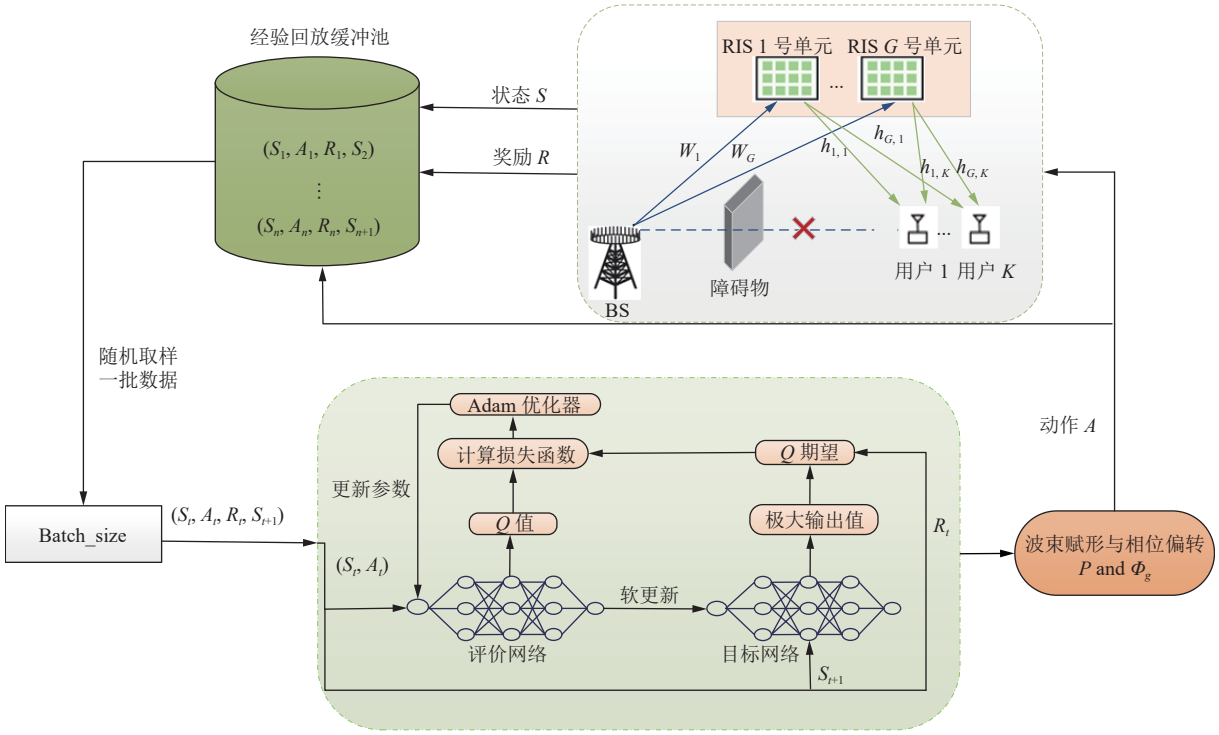

图 2 基于 DQN 的离散化联合设计网络结构图

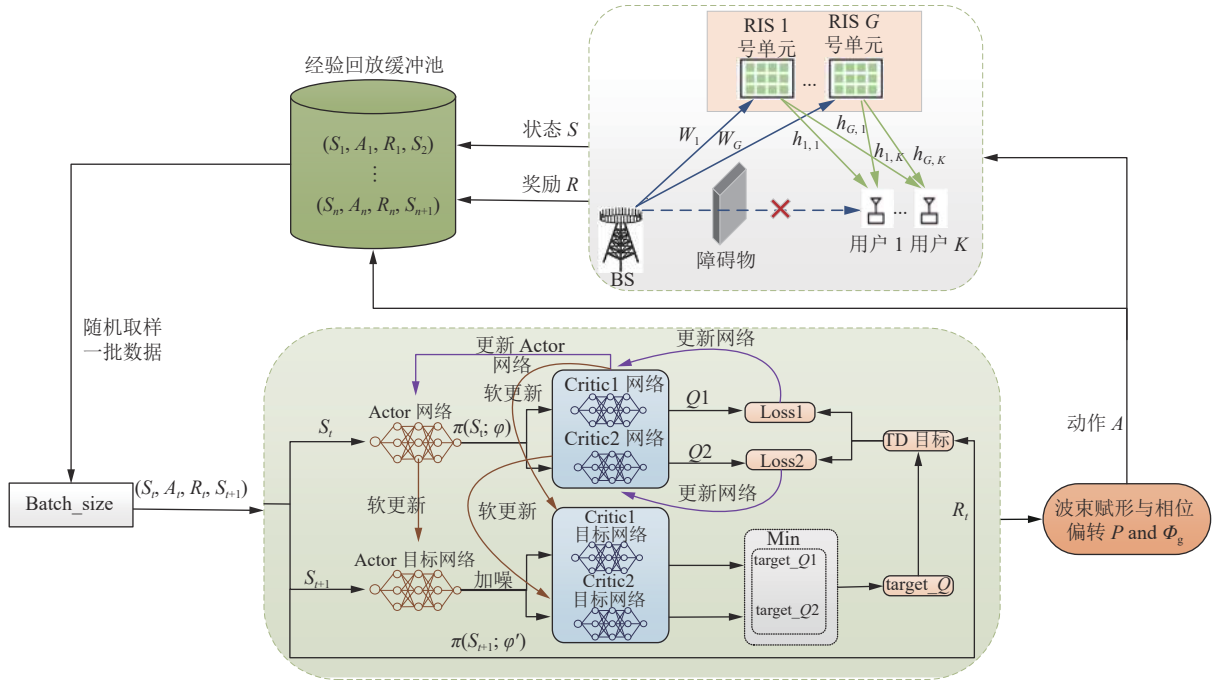

图 3 基于 TD3 的连续联合设计网络结构图

3.1 基于 DRL 的离散化发射波束赋形与相位偏转联合设计

本小节利用 DQN 算法处理离散动作空间问题，最大化加权和速率。DQN 算法中网络均为四层前馈神经网络。在评价网络的输出层放置一个归一化过程，以满足功率约束条件。DQN 的关键组件定义如下。

1) 状态：当前 t 时刻的状态S_t包含所有用户的信道信息$\boldsymbol{W}_g^{(t)}, \boldsymbol{h}_{g,k}^{(t)}, \forall g,k$，以及 t−1 时刻的动作$\boldsymbol{\Phi}_g^{(t-1)}, \boldsymbol{p}_k^{(t-1)}, \forall g,k$。因此，$S_t$可定义为：$S_t = [\boldsymbol{W}_g^{(t)}, \boldsymbol{h}_{g,k}^{(t)}, \boldsymbol{\Phi}_g^{(t-1)}, \boldsymbol{p}_k^{(t-1)}]^{\mathrm{T}}, \forall g,k$。

2) 动作：当前 t 时刻的动作A_t包括发射波束赋

形$\boldsymbol{p}_k^{(t)}$和 RIS 相位偏转$\boldsymbol{\Phi}_g^{(t)}$，则$A_t$可定义为：$A_t=[\boldsymbol{\Phi}_g^{(t)},\boldsymbol{p}_k^{(t)}]^{\mathrm{T}},\forall g,k$。而 DQN 处理的动作为离散的，因此将动作进行离散化处理。对于$\boldsymbol{p}_k^{(t)}$，首先将其拆分为两部分：

$$\boldsymbol{p}_k^{(t)}=\sqrt{q_g^{(t)}}\tilde{\boldsymbol{p}}_k^{(t)} \tag{19}$$

式中，$\sqrt{q_g^{(t)}}=\left\|\boldsymbol{p}_k^{(t)}\right\|^2$代表 BS 在 t 时刻的发射功率，且满足$0\leqslant q_g^{(t)}\leqslant P_t$；$\tilde{\boldsymbol{p}}_k^{(t)}$代表发射波束的方向，$\tilde{\boldsymbol{p}}_k^{(t)}\in[0,2\pi)$。将 BS 的可用发射功率电平在 0～$P_t$ 间均匀取q_{pow}个值，并将所选功率电平集合定义为功率集$\mathcal{P}=\left\{0,\frac{1}{q_{\mathrm{pow}}-1}P_t,\frac{2}{q_{\mathrm{pow}}-1}P_t,\cdots,P_t\right\}$。此外，定义一个由$q_{\mathrm{code}}$个码向量$\boldsymbol{c}_q\in\mathbb{C}^{N\times 1}$组成的并覆盖$\tilde{\boldsymbol{p}}_k^{(t)}$在$[0,2\pi)$上任意方向的码本$C$，其中$q\in\{0,1,\cdots,q_{\mathrm{pow}}-1\}$。令码本矩阵为$C=[\boldsymbol{c}_0,\boldsymbol{c}_1,\cdots,\boldsymbol{c}_{q_{\mathrm{code}}-1}]\in\mathbb{C}^{N\times q_{\mathrm{code}}}$，$C$中的每一列代表波束的一个方向。本文采用了文献 [25] 中的码本矩阵。用$C[n,q]$代表第q个码中的第n个天线元素的相位偏转，$C[n,q]=\frac{1}{\sqrt{N}}\exp\left\{\mathrm{j}\frac{2\pi}{S}\left|\frac{n \bmod \left(q+\frac{q_{\mathrm{code}}}{2},q_{\mathrm{code}}\right)}{q_{\mathrm{code}}/S}\right|\right\}$，其中，$S$代表每个天线元件的可用相位值的数量。因此，BS 可以分别从功率集$\mathcal{P}$与码本C中选取$q_g^{(t)}$与$\boldsymbol{c}^{(t)}$来确定其波束赋形矩阵。

3) 奖励：在 t 时刻，通过给定的瞬时信道信息$\boldsymbol{W}_g^{(t)},\boldsymbol{h}_{g,k}^{(t)},\forall g,k$以及从评价网络得到的动作$\boldsymbol{\Phi}_g^{(t)},\boldsymbol{p}_k^{(t)}$可确定奖励，奖励设置为系统的和速率。

DQN 优化算法见算法 1。

算法 1：离散化发射波束赋形与 RIS 相位偏转联合设计算法

输入：$\boldsymbol{W}_g^{(t)},\boldsymbol{h}_{g,k}^{(t)},\forall g,k$

输出：最优动作$A=\{\boldsymbol{\Phi}_g,\boldsymbol{P}\}$，$Q$ 值函数

初始化：回合数 E，每回合时隙 T，目标网络参数θ'，评价网络参数θ，经验回放缓冲池 M，发射波束赋形矩阵 $\boldsymbol{P}$，相位偏转矩阵$\boldsymbol{\Phi}_g$

for $i=0,1,\cdots,E-1$ do

　　收集初始状态$S_0=\{\boldsymbol{W}_g,\boldsymbol{h}_{g,k}\},\forall g,k$

　　for $j=0,1,\cdots,T-1$ do

　　根据ε–贪婪策略选择动作A_t；

　　与环境交互得到A_t对应的奖励R_t，并得到下一时刻的状态S_{t+1}；

　　将S_t、A_t、R_t和S_{t+1}放入回放缓冲池 M 中；

　　从 M 中随机抽取一批数据$\{S_t,A_t,R_t,S_{t+1}\}$；

　　根据式 (10) 计算损失函数；

　　利用损失函数梯度更新评价网络参数θ；

　　每隔T_{step}步，更新一次目标网络参数θ'；

　　更新状态，将神经网络的输入设置为S_{t+1}。

　　end

end

DQN 算法的网络参数设置见表 1。

表 1　表格 1 DQN 算法超参数描述

参数	描述	值
γ	对未来奖励的折扣率	0.6
μ	网络的学习率	0.000 05
batch_size	批处理数据的大小	32
T_{step}	目标网络延迟同步更新的步数	100
$\mathcal{M}$	经验回放缓冲池的大小	50 000
E	回合数	1 000
T	每回合的步数	10 000

3.2　基于 DRL 的连续发射波束赋形与相位偏转联合设计

本小节利用 TD3 算法处理连续动作空间问题，最大化加权和速率。TD3 算法中 Actor 网络与 Critic 网络都是四层前馈神经网络。为满足功率约束条件，在隐藏层和 Actor 网络的输出层放置归一化过程。TD3 的关键组件定义如下：

1) 状态：TD3 算法中的状态组成与 DQN 算法相似，不同之处在于 t–1 时刻的动作为连续值，$S_t=[\boldsymbol{W}_g^{(t)},\boldsymbol{h}_{g,k}^{(t)},\boldsymbol{\Phi}_g^{(t-1)},\boldsymbol{p}_k^{(t-1)}]^{\mathrm{T}},\forall g,k$。

2) 动作：TD3 算法也将发射波束赋形与相位偏转作为策略网络的输出。智能体通过强化学习选择动作$A_t=[\boldsymbol{\Phi}_g^{(t)},\boldsymbol{p}_k^{(t)}]^{\mathrm{T}},\forall g,k$。

3) 奖励：与 DQN 算法奖励计算方式相同，以和速率作为奖励。

TD3 优化算法见算法 2。

算法 2：连续发射波束赋形与 RIS 相位偏转联合设计算法

输入：$\boldsymbol{W}_g^{(t)},\boldsymbol{h}_{g,k}^{(t)},\forall g,k$

输出：最优动作$A=\{\boldsymbol{\Phi}_g,\boldsymbol{P}\}$，$Q$ 值函数

初始化：回合数E，每回合时隙T，策略网络参数φ，目标策略网络参数φ'，两个评价网络参数θ_1与θ_2，两个目标评价网络参数θ_1'与θ_2'，经验回放缓冲池$\mathcal{M}$，发射波束赋形矩阵$\boldsymbol{P}$，相位偏转矩阵$\boldsymbol{\Phi}_g$

for $i=0,1,\cdots,E-1$ do

　　$n=0$；

　　收集初始状态$S_0=\{\boldsymbol{W}_g,\boldsymbol{h}_{g,k}\},\forall g,k$；

　　for $j=0,1,\cdots,T-1$ do

以S_t作为输入，策略网络输出相应动作A_t；

与环境交互得到A_t对应的奖励R_t，并得到下一时刻的状态S_{t+1}；

将S_t、A_t、R_t和S_{t+1}放入回放缓冲池$\mathcal{M}$中；

从$\mathcal{M}$中随机抽取一批数据$\{S_t, A_t, R_t, S_{t+1}\}$；

$n = n + 1$；

得到两个评价网络输出Q_{θ_1}、Q_{θ_2}；

根据式 (12) 计算得到Q_{target}；

根据式 (13)、式 (14) 与式 (15) 更新评价网络；

if n % T_{step}== 0 then

根据式 (16) 更新策略网络；

根据式 (17)、(18) 更新φ'、θ_1'与θ_2'；

end if

更新状态，将神经网络的输入设置为S_{t+1}；

end

end

TD3 算法的网络参数设置见表 2。

表 2　TD3 算法超参数描述

参数	描述	值
γ	对未来奖励的折扣率	0.99
α	更新训练评价网络的学习率	0.001
β	更新训练策略网络的学习率	0.000 1
τ	更新目标策略网络与目标价值网络的学习率	0.001
λ	训练评价网络和训练策略网络的衰减率	0.000 01
batch_size	批处理数据的大小	16
$\mathcal{M}$	经验回放缓冲池的大小	100 000
T_{step}	策略网络延迟更新的步数	4
E	回合数	1 000
T	每回合的步数	10 000

4　仿真结果与分析

信道增益设置为$v_k \sim \mathcal{CN}(0, 10^{-0.1PL(r)})$[17, 26]，其中$PL(r) = \vartheta_a + 10\vartheta_b \lg(r) + \xi$，$\xi \sim \mathcal{N}(0, \sigma_\xi^2)$[3]。通过设置参数$\sigma_u^2 = -85$ dBm，$\vartheta_t = 9.82$ dBi，$\vartheta_r = 0$ dBi，$\vartheta_a = 61.4$，$\vartheta_b = 2$以及$\sigma_\xi = 5.8$ dB来实现信道。建立一个二维坐标平面，位于原点位置的 BS 具有$N = 20$根天线，用户均匀随机分布在以(40 m, 0 m)为中心，10 m为半径的圆中。设置$G = 2$个 RIS 单元，坐标位置分别为(40 m, 30 m)和(30 m, 40 m)。除非另有说明，否则本文总发射功率设置为$P_t = 30$ dBm[17]。

本文将文献 [11] 所提算法与迫零 (zero forcing, ZF)+随机 PBF 算法设为基准算法。图 4 展示了不同算法中和速率与 RIS 元件个数的关系，将本文所提算法与两种基准算法进行对比分析。从图中可以看到随着 RIS 元件个数的增加，每个算法的和速率均在增长，且本文提出的方法均优于迫零波束赋形随机相位偏转算法。在基于 DQN 离散化联合优化的算法中，和速率随着bit的增大而逐渐增长。当bit = 1时，和速率略高于迫零波束赋形随机相位偏转算法，且在$M = 120$处，和速率高出迫零波束赋形随机相位偏转算法 45.3%；当bit = 2时，和速率增益相较于bit = 1的情况更为显著；当bit = 4时，仿真结果已经可以达到甚至超过文献 [11] 的效果，且在$M = 60$处，和速率性能高出文献 [11] 所提方法 4.5%。仿真时bit取 1、2、4 这 3 种情况，因为bit越往后增大，和速率差距越来越小，根据功率损耗[25-26] 计算公式$1/\left(2^B/\pi \sin\left(2^B/\pi\right)\right)^2$可以计算 2-bit与 4-bit场景下功率损耗分别为 0.9 dB 与 0.03 dB，当bit无穷大时，功率损耗为 0 dB；当bit增加时，系统的反馈链路的开销也会增大。综合上述因素，bit可以选择为 4。基于 DQN 的方法处理离散动作，其和速率性能差于连续动作的方法。从仿真图 4 可以发现本文提出的基于 TD3 连续联合优化的方法得到的效果最好，性能最优，且在$M = 120$处，和速率高出文献 [11] 仿真结果 30.7%。

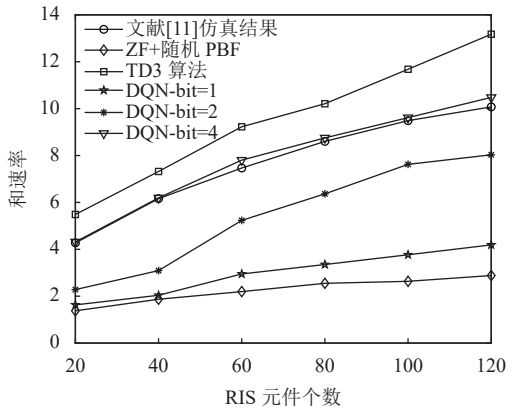

图 4　当$N_p = 2, K = 2$时，和速率与 RIS 大小的关系

图 5 展示了在 DQN 算法下，改变码本比特数目对系统性能的影响。从图中可以看到，随着比特数目的增大，系统性能也在逐渐提高，并且当bit = 4时算法收敛最快。图 6 展示了在相同用户个数和 RIS 元件个数的条件下 DQN 与 TD3 算法分别对系统性能的影响。从图 6 中可以看出基于 DQN 算法的系统在收敛速度上高于基于 TD3 算法的系统，但从整体系统性能来看，基于 TD3 算法的系统性能高于基于 DQN 算法的系统性能。除去 TD3 算法处理连续动作空间，而 DQN 算法处理离散动作空间这一点，TD3 算法使用了 4 个 Critic 网

络，并选取最小 Q 值作为目标值，缓解自举和最大化造成的高估，使得系统更加稳定。

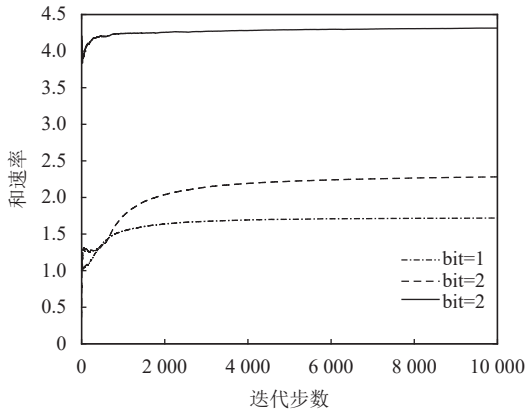

图 5　基于 DQN 的联合设计算法收敛度

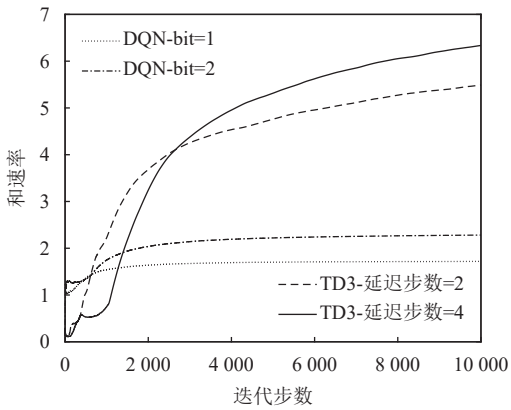

图 6　当$N_p=2,K=2,M=20$时，基于 DQN 算法与基于 TD3 算法的系统性能比较

图 7 评估了本文所提 TD3 算法在不同用户个数的条件下发射功率与和速率的关系，随着发射功率从 20 dBm 增长到 45 dBm，所有多用户的情况都呈现出上升趋势。随着用户数量的增加，在相同发射功率的条件下得到的和速率也会增加。

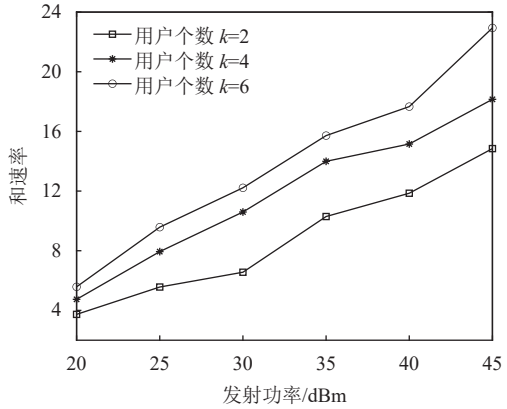

图 7　TD3 算法下当$N_p=2,M=20$时，和速率与发射功率大小的关系

图 8 显示了在 TD3 算法下，改变用户个数和 RIS 元件个数对系统性能的影响。从图中可以看到，改变用户个数和 RIS 元件个数，和速率会随着用户个数以及 RIS 元件个数的增加而增加。

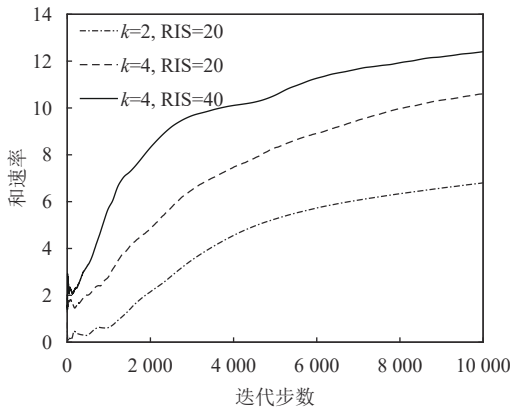

图 8　基于 TD3 的联合设计算法收敛度

图 9、图 10、图 11 都基于 TD3 算法分别研究了奖励折扣率γ、不同衰减率λ和策略网络延迟同步更新步数T_{step}对算法性能的影响。设置参数$\gamma=\{0.96,0.98,0.99,0.995\}$，$\lambda=\{0.001,0.000\,1,0.000\,01,0.000\,001\}$和$T_{\text{step}}=\{2,3,5\}$。图 9 与图 10 表明网络参数中$\gamma$、$\lambda$对系统奖励和速率的影响很小，并且在$\gamma=0.99$和$\lambda=0.000\,01$的条件下和速率最大。图 11表明策略网络延迟同步更新的步数对系统性能影响较小。

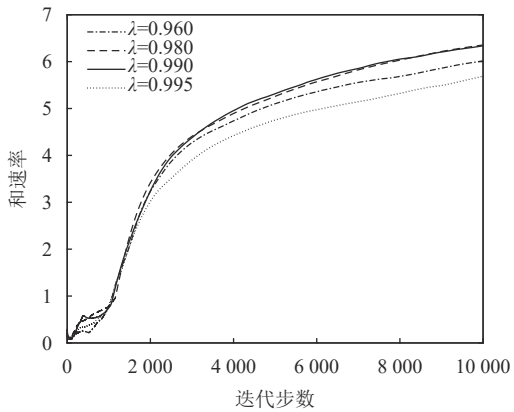

图 9　TD3 算法下不同折扣率下的和速率与迭代步数的关系

本文算法中的网络均为 4 层前馈神经网络。DQN 算法中的评价网络与 TD3 算法中的 Actor 网络的每层神经元个数分别为$|S|$、$2|S|$、$2|A|$、$|A|$，DQN 算法中目标网络与 TD3 算法中 Critic 网络的每层神经元个数分别为$|S|$、$2|S|+|A|$、100，所以 DQN 算法的计算复杂度为$O(ET((|S|\times 2|S|+2|S|\times 2|A|+2|A||A|)+(|S|(2|S|+|A|)+(2|S|+|A|)\times 100)))$，TD3 算法的计算复杂度为$O(2ET((|S|\times 2|S|+2|S|\times 2|A|+2|A||A|)+2(|S|(2|S|+|A|))+(2|S|+|A|)\times 100)))$，其中，$|S|$

为状态数，$|A|$为动作数，E为回合数，T为每回合的步数。文献[11]中凸优化算法复杂度为$O((M+1)^{4.5})$，M为智能反射面元件个数。与凸优化算法相比，本文所用强化学习算法在训练时复杂度较高，但当训练阶段结束后进行预测时，算法的复杂度会变得非常小。

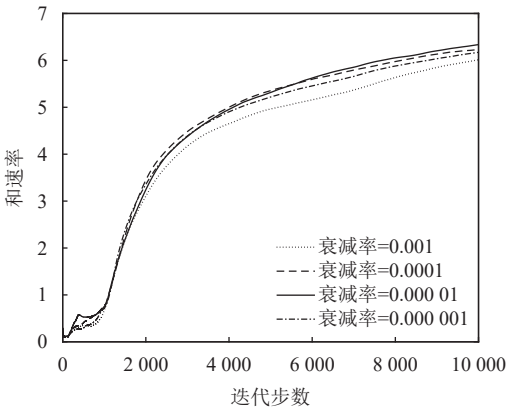

图 10 TD3 算法下衰减率对和速率的影响

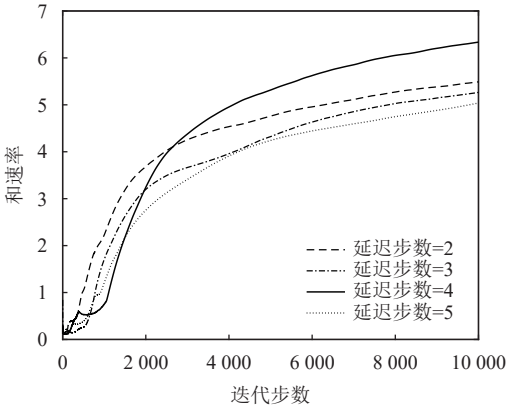

图 11 TD3 算法下延迟步数对和速率的影响

5 结束语

为了支持毫米波多用户通信传输，本文引入了分布式部署 RIS 单元来辅助毫米波通信，并基于 DRL 技术的最新进展，提出了离散化和连续两种情况下的发射波束赋形和相位偏转的联合设计，实现最大化 RIS 辅助毫米波通信系统的加权和速率。本文提出的基于 DRL 的算法具有较强的鲁棒性，因此很容易适应各种通信系统设置。

参考文献

[1] GUAN K, PENG B, HE D P, et al. Channel sounding and ray tracing for intrawagon scenario at mmwave and sub-mmwave bands[J]. IEEE Transactions on Antennas and Propagation, 2021, 69(2): 1007-1019.

[2] 肖振宇, 刘珂, 朱立鹏. 无人机机间毫米波阵列通信技术[J]. 通信学报, 2022, 43(10): 196-209.
XIAO Z Y, LIU K, ZHU L P. Millimeter-Wave array enabled UAV-to-UAV communication technology[J]. Journal on Communications, 2022, 43(10): 196-209.

[3] HUANG C W, ZAPPONE A, ALEXANDROPOULOS G C, et al. Reconfigurable intelligent surfaces for energy efficiency in wireless communication[J]. IEEE Transactions on Wireless Communications, 2019, 18(8): 4157-4170.

[4] SHAO X D, YOU C S, MA W Y, et al. Target sensing with intelligent reflecting surface: Architecture and performance[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(7): 2070-2084.

[5] ZHANG Z J, DAI L L, CHEN X B, et al. Active RIS vs Passive RIS: Which will prevail in 6G?[J]. IEEE Transactions on Communications, 2023, 71(3): 1707-1725.

[6] HUANG C W, YANG Z H, ALEXANDROPOULOS G C, et al. Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design[J]. IEEE Journal on Selected Areas in Communications. 2021, 39(6): 1663-1677.

[7] 胡浪涛, 毕松姣, 刘全金, 等. 基于强化学习的智能超表面辅助无人机通信系统物理层安全算法[J]. 电子与信息学报, 2022, 44(7): 2407-2415.
HU L T, BI S J, LIU Q J, et al. Physical layer security algorithm of reconfigurable intelligent surface-assisted unmanned aerial vehicle communication system based on reinforcement learning[J]. Journal of Electronics & Information Technology, 2022, 44(7): 2407-2415.

[8] WEI L, HUANG C W, ALEXANDROPOULOS G C, et al. Channel estimation for RIS-empowered multi-user MISO wireless communications[J]. IEEE Transactions on Communications, 2021, 69(6): 4144-4157.

[9] 郭海燕, 杨震, 邹玉龙, 等. 基于主被动波束成形联合优化的双 RIS 辅助抗干扰通信方法[J]. 通信学报, 2022, 43(7): 21-30.
GUO H Y, YANG Z, ZOU Y L, et al. Double-RIS assisted anti-jamming communication method based on joint active and passive beamforming optimization[J]. Journal on Communications, 2022, 43(7): 21-30.

[10] WU Q, ZHANG R. Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5394-5409.

[11] XU P, CHEN G J, YANG Z, et al. Reconfigurable intelligent surfaces-assisted communications with discrete phase shifts: how many quantization levels are required to achieve full diversity?[J]. IEEE Wireless Communications Letters, 2021, 10(2): 358-362.

[12] WU Q Q, ZHANG R. Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts[J]. IEEE Transactions on Communications, 2020, 68(3): 1838-1851.

[13] GE J G, LIANG Y C, JOUNG J, et al. Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination[J]. IEEE Transactions on Communications, 2020, 68(10): 6070-6085.

[14] 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.

[15] GUO H Y, LIANG Y C, CHEN J, et al. Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks[J]. IEEE Transactions on Wireless Communications, 2020, 19(5): 3064-3076.

[16] HAN Y, TANG W K, JIN S, et al. Large intelligent surface-assisted wireless communication exploiting statistical CSI[J]. IEEE Transactions on Vehicular Technology, 2019, 68(8): 8238-8242.

[17] WANG P L, FANG J, YUAN X J, et al. Intelligent reflecting surface-assisted millimeter wave communications: Joint active and passive precoding design[J]. IEEE Transactions on Vehicular Technology, 2020, 69(12): 14960-14973.

[18] ZHAO N, LIANG Y C, NIYATO D, et al. Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks[J]. IEEE Transactions on Wireless Communications, 2019, 18(11): 5141-5152.

[19] 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.

[20] YANG H L, XIONG Z H, ZHAO J, et al. Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications[J]. IEEE Transactions on Wireless Communications, 2021, 20(1): 375-388.

[21] MEI H B, YANG K, LIU Q, et al. 3D-Trajectory and phase-shift design for ris-assisted UAV systems using deep reinforcement learning[J]. IEEE Transactions on Vehicular Technology, 2022, 71(3): 3020-3029.

[22] CHU Z, HAO W M, XIAO P, et al. Intelligent reflecting surface aided multi-antenna secure transmission[J]. IEEE Wireless Communications Letters, 2020, 9(1): 108-112.

[23] SUTTON R S, BARTO A G. Reinforcement learning: An introduction[M]. Cambridge: MIT press, 2018.

[24] FUJIMOTO S, HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]// International Conference on Machine Learning. New York: PMLR, 2018: 1587-1596.

[25] ZHOU W X, CUI Z F, LI B, et al. Beamforming codebook design and performance evaluation for 60GHz wireless communication[C]//2011 11th International Symposium on Communications & Information Technologies (ISCIT). Piscataway: IEEE, 2011: 30-35.

[26] AKDENIZE M R, LIU Y P, SAMIMI M K, et al. Millimeter wave channel modeling and cellular capacity evaluation[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(6): 1164-1179.

编　辑　叶　芳