利用A2C-ac的城轨车车通信资源分配算法

王瑞峰[①] 张 明*[①] 黄子恒[①] 何 涛[②]
[①](兰州交通大学自动化与电气工程学院 兰州 730070)
[②](兰州交通大学自动控制研究所 兰州 730070)

摘 要：在城市轨道交通列车控制系统中，车车(T2T)通信作为新一代列车通信模式，利用列车间直接通信来降低通信时延，提高列车运行效率。在T2T通信与车地(T2G)通信并存场景下，针对复用T2G链路产生的干扰问题，在保证用户通信质量的前提下，该文提出一种基于多智能体深度强化学习(MADRL)的改进优势演员-评论家(A2C-ac)资源分配算法。首先以系统吞吐量为优化目标，以T2T通信发送端为智能体，策略网络采用分层输出结构指导智能体选择需复用的频谱资源和功率水平，然后智能体做出相应动作并与T2T通信环境交互，得到该时隙下T2G用户和T2T用户吞吐量，价值网络对两者分别评价，利用权重因子β为每个智能体定制化加权时序差分(TD)误差，以此来灵活优化神经网络参数。最后，智能体根据训练好的模型联合选出最佳的频谱资源和功率水平。仿真结果表明，该算法相较于A2C算法和深度Q网络(DQN)算法，在收敛速度、T2T成功接入率、吞吐量等方面均有明显提升。

关键词：城市轨道交通；资源分配；T2T通信；多智能体深度强化学习；A2C-ac算法
中图分类号：TN929.5 **文献标识码**：A **文章编号**：1009-5896(2024)04-1306-08
DOI: 10.11999/JEIT230623

Resource Allocation Algorithm of Urban Rail Train-to-Train Communication with A2C-ac

WANG Ruifeng[①] ZHANG Ming[①] HUANG Ziheng[①] HE Tao[②]
[①](*School of Automation and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou* 730070, *China*)
[②](*Automatic Control Institute, Lanzhou Jiaotong University, Lanzhou* 730070, *China*)

Abstract: In the train control system of urban rail transit, Train-to-Train (T2T) communication, a new train communication mode, use direct communication between trains to reduce communication delay and improve train operation efficiency. In the scenario of the coexistence of T2T communication and Train to Ground (T2G) communication, an improved Advantage Actor-Critic-ac (A2C-ac) resource allocation algorithm based on Multi-Agent Deep Reinforcement Learning (MADRL) is proposed to solve the interference problem caused by multiplexing T2G links, and under the premise of ensuring the quality of user communication. Firstly, taking the system throughput as the optimization goal and the T2T communication transmitter as the agent, the policy network adopts a hierarchical output structure to guide the agent in selecting the spectrum resources and power level to be reused. Then the agent makes corresponding actions and interacts with the communication environment to obtain the throughput of T2G users and T2T users in the time slot. The value network evaluates the two separately and uses the weight factor β to customize the weighted Temporal Difference (TD) error for each agent to optimize the neural network parameters flexibly. Finally, the agents jointly select the best spectral resources and power levels according to the trained model. The simulation results show that compared with the A2C and Deep Q-Networks (DQN) algorithms, the proposed algorithm has significantly improved the convergence speed, T2T successful access rate, and the throughput.

Key words: Urban rail transit system; Resource allocation; Train-to-Train (T2T); Multi-Agent Deep Reinforcement Learning (MADRL); Advantage Actor-Critic-ac (A2C-ac) algorithm

收稿日期：2023-06-25；改回日期：2023-09-28；网络出版：2023-10-11
*通信作者： 张明 1520792671@qq.com
基金项目：国家自然科学基金铁路基础研究联合基金(U2268206)
Foundation Item: The National Natural Science Foundation of China Railway Basic Research Joint Fund (U2268206)

1　引言

目前，城市轨道交通普遍采用基于通信的列车控制系统(Communication Based Train Control system, CBTC)，实现车地(Train to Ground, T2G)信息双向传输。随着CBTC系统大量上线运行，其轨旁设备繁多、通信时延较高等问题逐渐暴露出来。终端直通(Device to Device, D2D)技术的快速发展为列车间通信提供了理论基础，文献[1]提出将列车间直接通信(Train to Train, T2T)技术纳入下一代高速铁路通信方法。文献[2]将T2T通信技术引入城市轨道交通列控系统，实现列车碰撞防护。城市轨道交通T2T通信模式通过复用车地通信上行链路的频谱资源，实现列车间信息直通，获取列车位置和状态等信息。但频谱复用必然会带来干扰，因此在保证系统通信性能的前提下，如何减少复用干扰以提高系统吞吐量成为目前T2T通信研究的难点和热点。

文献[3,4]应用图论的方法，将信道分配转化为二分图求解最大匹配问题，最大化系统吞吐量。文献[5,6]联合信道选择和功率控制，提出重叠联盟形成博弈模型，最大限度提升D2D链路吞吐量。文献[7]采用分布思想将非凸模型拆分为信道匹配和功率分配两个子问题，提出了基于二分法结合匈牙利算法的城市轨道交通资源分配算法。文献[8]应用群智能算法协调T2T用户复用T2G链路产生的干扰。文献[9]采用拉格朗日对偶函数法求解不同通信模式下列车的最优发射功率，明显提升系统吞吐量。文献[10,11]提出了基于多智能体深度强化学习(Multi-Agent Deep Reinforcement Learning, MADRL)的资源分配方案，采用深度Q网络(Deep Q-Networks, DQN)算法实现了发射功率和复用信道的自主选择。但传统DQN算法在多智能体训练中存在不稳定性，同时，策略梯度法在多智能体协作时会产生较高的方差。

本文针对T2T通信复用T2G通信链路产生的干扰问题，提出一种基于改进的优势演员-评论家(Advantage Actor-Critic-ac, A2C-ac)算法的多智能体深度强化学习资源分配策略。以T2T通信发送端为智能体，采用分层输出结构改进策略网络，以减小动作空间，加快算法收敛速度；价值网络对T2G用户和T2T用户吞吐量分别评价，利用权重因子β为每个智能体定制化加权时序差分(Temporal Difference, TD)误差，提高训练的灵活性和准确性。最后通过迭代优化神经网络参数，使智能体能够联合选出最佳传输功率和频谱资源，最大化系统吞吐量。

2　系统模型

如图1所示，列车A, B, C与基站间存在T2G通信，同时列车A, B, C之间进行T2T通信，此时，系统产生的干扰主要分为两部分：一部分是T2G通信发送端对T2T通信接收端的干扰；另一部分是T2T通信发送端对其复用的T2G通信基站的干扰。通过调整用户发送功率以及合理配置用户之间的复用关系，降低系统干扰，优化系统整体吞吐量。

考虑在T2T与T2G并存场景中，基站相比于列车有较强的抗干扰能力，因此，T2T通信用户复用T2G用户上行链路进行通信。在单个小区内，列车数量有限，最多同时存在M个T2G用户和N个T2T用户，分别用集合$\mathcal{M}\in\{1,2,\cdots,M\}$, $\mathcal{N}\in\{1,2,\cdots,N\}$表示。可用带宽均分为$M$个相互正交的资源块供T2G通信链路使用[12]。

在城轨综合车地通信系统(Long Term Evolution-Metro, LTE-M)网络中，将时间划分为离散的时隙，记为t，$t\in\{1,2,\cdots,T\}$。假设基站在每个时隙t都会为每个T2G用户分配一个车地上行链路的信道资源，即资源块(Resource Block, RB)。当多个T2T用户复用同一个T2G用户的RB时，会严重影响该T2G用户的通信质量，为此在模型中做出以下限制：在同一个时隙t下，1个RB仅允许被1个T2T用户复用，1个T2T用户也只能复用1个RB。为了便于深度强化学习算法训练，T2G用户m在时

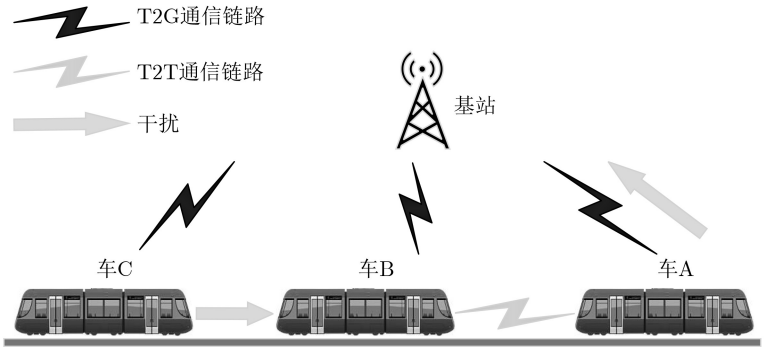

图 1 T2T通信示意图

隙t的信干噪比(Signal to Interference plus Noise Ratio, SINR)如式(1)所示

$$\mathrm{SINR}_t^m=\frac{p_t^m g_t^{m,B}}{\delta^2+\sum\limits_{n=1}^{N}\rho_{m,n}p_t^n g_t^{n,B}} \tag{1}$$

T2T用户n在时隙t的信干噪比如式(2)所示

$$\mathrm{SINR}_t^n=\frac{p_t^{n,l} g_t^n}{\delta^2+\sum\limits_{n=1}^{N}\rho_{m,n}p_t^m g_t^{m,n}} \tag{2}$$

其中，p_t^m和$p_t^{n,l}$分别表示第m个T2G用户与第n个T2T用户的发射功率；δ^2表示噪声功率。$\rho_{m,n}$代表用户间的复用关系，当$\rho_{m,n}=1$时，表示第n个T2T用户复用第m个RB；反之，$\rho_{m,n}=0$时，表示不存在复用关系。

根据香农公式可得第m个T2G用户与第n个T2T用户吞吐量如式(3)和式(4)所示

$$R_t^m=B_f\log_2\left(1+\mathrm{SINR}_t^m\right) \tag{3}$$

$$R_t^n=B_f\log_2\left(1+\mathrm{SINR}_t^n\right) \tag{4}$$

优化目标为系统中T2T用户和T2G用户的总体吞吐量R_{sum}，目标函数如式(5)所示

$$\begin{aligned}\max R_t^{\mathrm{sum}}=&\sum_{m=1}^{M}\sum_{n=1}^{N}B_f[\log_2\left(1+\mathrm{SINR}_t^m\right)\\&+\log_2\left(1+\mathrm{SINR}_t^n\right)]\end{aligned} \tag{5}$$

约束条件如式(6)–式(9)所示

$$\mathrm{SINR}_t^m\geq\mathrm{SINR}_m^{\mathrm{req}},1\leq m\leq M \tag{6}$$

$$\mathrm{SINR}_t^n\geq\mathrm{SINR}_n^{\mathrm{req}},0\leq n\leq N \tag{7}$$

$$0\leq P_t^{n,l}\leq P_n^{\max},0\leq n\leq N \tag{8}$$

$$\left.\begin{aligned}0\leq\sum_{m=1}^{M}\rho_{m,n}\leq 1,\rho_{m,n}\in\{0,1\}\\0\leq\sum_{n=1}^{N}\rho_{m,n}\leq 1,\rho_{m,n}\in\{0,1\}\end{aligned}\right\} \tag{9}$$

式(6)–式(9)中依次对T2G用户和T2T用户信干噪比、T2T用户发射功率、频谱复用作出限制。其中，B_f表示带宽，$\mathrm{SINR}_n^{\mathrm{req}}$表示T2T用户最低信干噪比阈值，$P_n^{\max}$表示T2T用户$n$可提供的最大发射功率。

3 马尔可夫决策过程(MDP)

城市轨道交通车车通信资源分配过程可以建模为马尔可夫决策过程(Markov Decision Process, MDP)，其状态空间、动作空间、奖励函数等相关设置如下：

状态空间：在时隙t的系统状态为$\boldsymbol{s}_t$，由T2T用户的吞吐量决定，即$\boldsymbol{s}_t=[s_t^1,\cdots,s_t^n,\cdots,s_t^N]$，其中$s_t^n=R_t^n$。

动作空间：在T2T通信资源分配中，影响系统吞吐量的动作是T2T通信发射功率和RB的选择。每个智能体在时隙t的动作可以表示为：$\boldsymbol{a}_t^{m,l}=[\mathrm{RB}_t^{n,m},p_t^{n,l}]$，其中$\mathrm{RB}_t^{n,m}$, $p_t^{n,l}$分别为智能体n选择复用的RB和发射功率。考虑网络复杂度和算法训练速度，将功率离散为L个水平，$p_t^{n,l}=lP_n^{\max}/L$，$l\in[1,2,\cdots,L]$。智能体n的动作空间可以表示为

$$\boldsymbol{\Lambda}=[\boldsymbol{a}_t^{1,1},\boldsymbol{a}_t^{1,2},\cdots,\boldsymbol{a}_t^{m,l},\cdots,\boldsymbol{a}_t^{M,L}] \tag{10}$$

奖励函数：当智能体执行动作$\boldsymbol{a}_t^{n,l}$后，环境将会做出对应改变，环境的状态由$\boldsymbol{s}_t\sim\boldsymbol{s}_{t+1}$，同时环境给出瞬时奖励$r_t$。为了评价智能体动作对整个系统的影响，将系统整体吞吐量设为奖励函数，此外，当约束条件式(6)–式(9)不满足时，对智能体n做出惩罚，促使算法快速收敛。智能体n在时隙t选择第m个RB时的奖励函数如式(11)所示

$$r_t^n=\begin{cases}-1,\text{式(6)}-\text{式(9)不满足}\\\sum\limits_{m=1}^{M}B_f[\log_2\left(1+\mathrm{SINR}_t^m\right)\\\quad+\log_2\left(1+\mathrm{SINR}_t^n\right)],\text{其他}\end{cases} \tag{11}$$

折扣因子：在强化学习中，出于系统长期稳定性的要求，还需要考虑一个完整回合的累积总奖励。T2T通信资源分配场景不同于围棋比赛，环境没有终止状态，为了避免产生奖励无限叠加，加入折扣因子控制长期奖励。折扣奖励如式(12)所示

$$R_t=\sum_{t=0}^{T}\gamma r_t^n \tag{12}$$

其中，折扣因子$\gamma\in[0,1]$，γ越接近1代表智能体越趋向于长远利益。

4 基于A2C-ac的资源分配算法设计

A2C算法基于Actor-Critic框架，包含两个神经网络：策略网络和价值网络，策略网络称为Actor，是对策略函数$\pi_\varphi(\boldsymbol{a}|\boldsymbol{s}_t)$的近似，控制智能体动作；价值网络称为Critic，是对状态价值函数$V_\pi(\boldsymbol{s}_t)$的近似，用来评估当前状态$\boldsymbol{s}_t$的好坏。在训练价值网络时，采用状态价值$V_\pi(\boldsymbol{s})$替代AC算法中的动作状态价值$Q_\pi(\boldsymbol{s},\boldsymbol{a})$；在训练策略网络时，采用优势函数(advantage function)代替价值网络中的原始回报，快速、稳定地实现策略优化和值函数优化。T2T

通信资源分配问题即为多智能体的RB选择和功率控制问题，相比于A2C算法，A2C-ac算法能够更好地处理多智能体之间的相互作用与协作，具有更高的训练效率。

4.1 A2C-ac资源分配算法

在基于A2C的资源分配算法中，智能体动作空间由RB数量M和功率水平数L决定，动作空间的大小直接影响算法训练速度，为此在策略网络中通过分层输出结构指导智能体动作；式(11)中由于T2T用户和T2G用户吞吐量所占比重相差过大，影响算法学习质量，因此设置额外的价值网络来评价T2G用户吞吐量，利用权重因子β为每个智能体定制化加权TD误差，灵活优化神经网络参数。如图2为基于A2C-ac的资源分配算法示意图。

4.2 策略网络改进

由式(10)可知，每个智能体的动作空间由两个独立的动作组成，即RB选择和功率水平选择，为确保智能体可以做出所有组合，Actor网络的输出单元数必须为$M\times L$。虽然深度神经网络可以处理高维计算，但会导致在每个时隙更新大量训练参数。

考虑智能体两种动作的独立性，采用两个单独的输出层代替原始策略网络的输出层，分别提供RB选择和功率水平选择的概率分布。例如，当m=8, L=15，输出层前面的隐藏层包含64个单元时，在不考虑偏置项的情况下，最后一层的权重数将由原先的8×15×64=7 680变为(8+15)×64=1 472，大大减少了神经网络计算量，从而加快算法训练速度。

采用分层输出结构后，智能体动作空间为$\boldsymbol{\Lambda}_{\mathrm{RB}}=[\mathrm{RB}_t^{n,1},\mathrm{RB}_t^{n,2},\cdots,\mathrm{RB}_t^{n,M}]$和$\boldsymbol{\Lambda}_p=[p^{n,1},p^{n,2},\cdots,p^{n,L}]$，策略网络结构如图3所示。

4.3 价值网络改进

由奖励函数式(5)可知，奖励由两部分组成，前者是所有T2G用户吞吐量之和，后者为第n个T2T用户吞吐量。不难发现，前者所占权重要远大于后者，这意味着奖励会严重偏向T2G用户吞吐量。为了解决这一问题，额外设置一个用来评价T2G用户吞吐量的价值网络，当式(6)–式(9)满足时，该网络的奖励r_t^{T2G}为时隙t下T2G用户总体吞吐量，奖励函数如式(13)所示

$$r_t^{\mathrm{T2G}}=c\begin{cases}-1, & \text{式(6)}-\text{式(9)不满}\\ \sum_{m=1}^{M}B_{\mathrm{f}}\log_2\left(1+\mathrm{SINR}_t^m\right), & \text{其他}\end{cases}\tag{13}$$

而对于智能体n，保留原价值网络，当式(6)–式(9)满足时，奖励$r_t^{n,\mathrm{T2T}}$为时隙t下T2T用户n的吞吐量，奖励函数如式(14)所示

$$r_t^{n,\mathrm{T2T}}=\begin{cases}-1, & \text{式(6)}-\text{式(9)不满足}\\ B_{\mathrm{f}}\log_2\left(1+\mathrm{SINR}_t^n\right), & \text{其他}\end{cases}\tag{14}$$

4.4 A2C-ac资源分配算法学习过程

将每个T2T用户发射端都视为一个智能体，所有智能体构成集合N。策略网络由参数向量$\boldsymbol{\theta}_t=(\theta_t^1,\theta_t^2,\cdots,\theta_t^N)$构成，状态值函数$V_{\boldsymbol{w}}^{\mathrm{T2G}}(\boldsymbol{s}_t)$, $V_{\boldsymbol{\psi}}^{n,\mathrm{T2T}}$分别由参数向量$\boldsymbol{w}_t=(w_t^1,w_t^2,\cdots,w_t^N)$和$\boldsymbol{\psi}_t=(\psi_t^1,\psi_t^2,\cdots,\psi_t^N)$构成。

在一个时隙t，策略网络通过当前状态$\boldsymbol{s}_t$计算动作RB_t^n, $p_t^{n,l}$的概率分布，根据策略函数

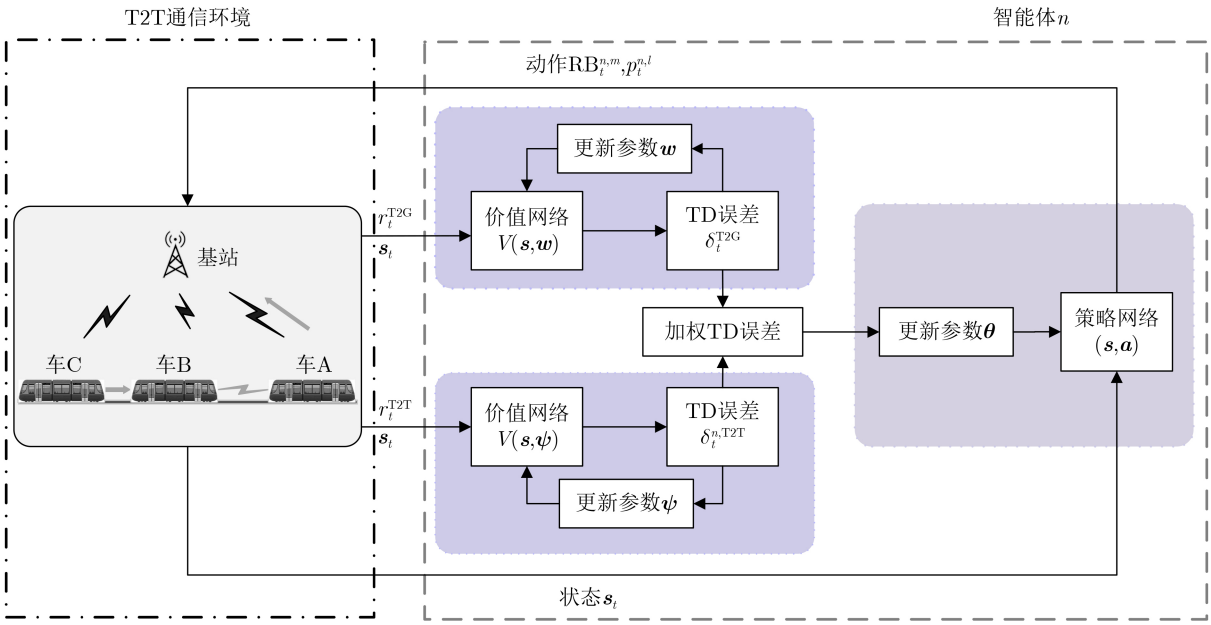

图 2 A2C-ac资源分配算法示意图

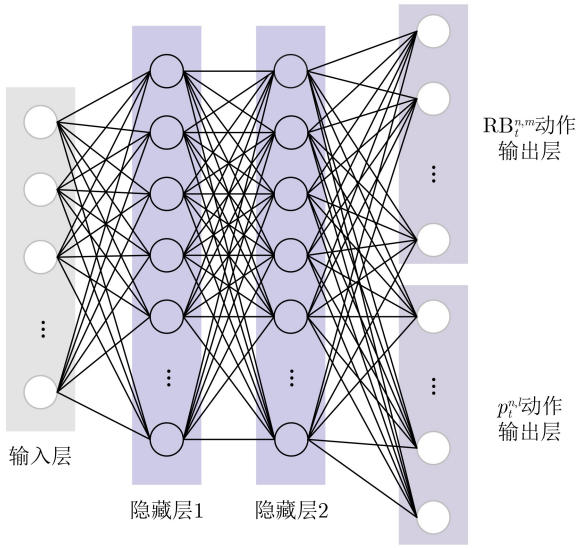

图 3 双输出层Actor网络结构

$\pi_{\boldsymbol{\theta}}(\mathrm{RB}_t^n|\boldsymbol{s}_t)$，$\pi_{\boldsymbol{\theta}}(p_t^{n,l}|\boldsymbol{s}_t)$分别采样一个动作，即$\mathrm{RB}_t^n\sim\pi_{\boldsymbol{\theta}}(\mathrm{RB}_t^n|\boldsymbol{s}_t)$，$p_t^{n,l}\sim\pi_{\boldsymbol{\theta}}(p_t^{n,l}|\boldsymbol{s}_t)$，然后智能体执行动作$\mathrm{RB}_t^n$和$p_t^{n,l}$，与环境交互，环境给出新的状态$\boldsymbol{s}_{t+1}$，并根据式(17)和式(18)给出奖励$r_t^{\mathrm{T2G}}$，$r_t^{n,\mathrm{T2T}}$。

为了降低AC算法在梯度更新过程中带来的高方差[13]，A2C算法引入了优势函数$A_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t)$，如式(15)所示

$$\begin{aligned}A_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t)&=Q_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t)-V_\pi(\boldsymbol{s}_t)\\&\approx r_t+\gamma V_\pi(\boldsymbol{s}_{t+1})-V_\pi(\boldsymbol{s}_t)\end{aligned}\tag{15}$$

根据时间差分误差的定义，TD误差是对优势函数$A_\pi(\boldsymbol{s}_t,\boldsymbol{a}_t)$的无偏估计，可以得到时隙$t$下第$n$个智能体的TD误差为

$$\delta_t=r_t+\gamma V(\boldsymbol{s}_{t+1})-V(\boldsymbol{s}_t)\tag{16}$$

其中，$V(\boldsymbol{s}_{t+1})$, $V(\boldsymbol{s}_t)$分别表示$t+1$时刻和t时刻第n个智能体的动作价值函数。

价值网络将动作、瞬时奖励、当前状态和下一状态作为输入得到TD误差。将式(13)、式(14)分别代入式(16)，得到两个价值网络对应的TD误差，如式(17)、式(18)所示

$$\delta_t^{\mathrm{T2G}}=r_t^{\mathrm{T2G}}+\gamma V_w^{\mathrm{T2G}}(\boldsymbol{s}_{t+1})-V_w^{\mathrm{T2G}}(\boldsymbol{s}_t)\tag{17}$$

$$\delta_t^{n,\mathrm{T2T}}=r_t^{n,\mathrm{T2T}}+\gamma V_\psi^{n,\mathrm{T2T}}(\boldsymbol{s}_{t+1})-V_\psi^{n,\mathrm{T2T}}(\boldsymbol{s}_t)\tag{18}$$

因每个智能体的策略网络只能由单个TD误差来更新，故需要引入权重因子β获得加权TD误差，为每个智能体定制TD误差，如式(19)所示

$$\delta_t^n=\beta\delta_t^{\mathrm{T2G}}+(1-\beta)\delta_t^{n,\mathrm{T2T}}\tag{19}$$

策略网络采用分层输出结构，为衡量策略网络输出的动作策略相对于实际策略的偏差，需要对两个独立的动作分别设置损失函数，如式(20)和式(21)所示

$$L(\pi)=-\ln\pi_{\boldsymbol{\theta}}^n(\mathrm{RB}_t^n|\boldsymbol{s}_t)\cdot\delta_t^n\tag{20}$$

$$L(\pi)=-\ln\pi_{\boldsymbol{\theta}}^n(p_t^{n,l}|\boldsymbol{s}_t)\cdot\delta_t^n\tag{21}$$

智能体n的价值网络采用均方误差(MSE)方法[14]基于加权TD误差更新参数ψ，损失函数如式(22)所示

$$L(\psi)=\frac{1}{2}[\beta\delta_t^{\mathrm{T2G}}+(1-\beta)\delta_t^{n,\mathrm{T2T}}]^2\tag{22}$$

在每个时隙t，策略网络分层输出结构会选出两个独立动作，同时两个输出层共享网络参数θ，因此在一个时隙t，θ需要更新两次，其更新如式(23)和式(24)所示

$$\boldsymbol{\theta}_{t+1}^n=\boldsymbol{\theta}_t^n+\alpha_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ln\pi_{\boldsymbol{\theta}}^n(\mathrm{RB}_t^n|\boldsymbol{s}_t)\delta_t^n\tag{23}$$

$$\boldsymbol{\theta}_{t+1}^n=\boldsymbol{\theta}_t^n+\alpha_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ln\pi_{\boldsymbol{\theta}}^n(p_t^{n,l}|\boldsymbol{s}_t)\delta_t^n\tag{24}$$

价值网络参数$\boldsymbol{w}$与$\boldsymbol{\psi}$更新如式(25)和式(26)所示

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t+\alpha_{\boldsymbol{w}}\nabla_{\boldsymbol{w}}V_w^{\mathrm{T2G}}(\boldsymbol{s}_t)\delta_t^{\mathrm{T2G}}\tag{25}$$

$$\boldsymbol{\psi}_{t+1}^n=\boldsymbol{\psi}_t^n+\alpha_{\boldsymbol{\psi}}\nabla_{\boldsymbol{\psi}}V_{\boldsymbol{\psi}}^{n,\mathrm{T2T}}(\boldsymbol{s}_t)\delta_t^{n,\mathrm{T2T}}\tag{26}$$

其中，$\alpha_{\boldsymbol{\theta}}$为策略网络学习率，取值为0.001，$\alpha_{\boldsymbol{w}}$，$\alpha_{\boldsymbol{\psi}}$为价值网络学习率，取值为0.01。

基于A2C-ac的T2T通信资源分配算法的伪代码如算法1所示。

接下来对算法1在执行时的复杂度进行分析：定义策略网络第i层神经元个数为X_i，价值网络第j层神经元个数为X_j，因此，双输出层策略网络的计算复杂度为：$O(2\sum_{i=3}^{I-2}X_{i-1}X_{i-2}+MX_{I-1}+LX_{I-1})$，价值网络计算复杂度为：$O(\sum_{j=2}^{J-1}X_{j-1}X_j)$。以T2T发送端为智能体，所提算法采用双价值网络对T2G用户和T2T用户吞吐量分别评价，故每个智能体都拥有1个策略网络和2个价值网络，同时价值网络需要估计t和$t+1$时刻的状态价值，因此对于每个智能体其计算复杂度为：$O(2\sum_{i=3}^{I-2}X_{i-1}X_{i-2}+4\sum_{j=2}^{J-1}X_{j-1}X_j+MX_{I-1}+LX_{I-1})\sqrt{b^2-4ac}$。价值网络和策略网络参数更新是线性的，因此参数更新复杂度为N+2N，即3个网络的参数个数之和。

5 仿真结果及分析

以python3.9为仿真平台，采用tensorflow2.5深度学习库，实现基于A2C-ac的T2T通信资源分配算法仿真。系统带宽为10 MHz，基站覆盖半径为1.5 km, T2G用户数为8，相邻列车间距为600～800 m, T2G与T2T用户最大发射功率为23 dBm。训练过程中通过SoftMax激活函数获得动作的概率

算法1　基于A2C-ac的T2T通信资源分配算法

(1) 初始化：初始化超参数γ, β,$\alpha_{\boldsymbol{\theta}}$, $\alpha_{\boldsymbol{w}}$, $\alpha_{\boldsymbol{\psi}}$；初始环境状态$\boldsymbol{s}_0$；初始化神经网络参数$\boldsymbol{\theta}$, $\boldsymbol{w}$,$\boldsymbol{\psi}$；
(2) For t=0: T do
(3)　For n=0: N do
(4)　　根据策略$\pi_{\boldsymbol{\theta}}(\mathrm{RB}_t^n|\boldsymbol{s})$与$\pi_{\boldsymbol{\theta}}(p_t^{n,l}|\boldsymbol{s})$各采样一个动作$\mathrm{RB}_t^n$,$p_t^{n,l}$
(5)　End
(6)　执行动作RB_t^n, $p_t^{n,l}$，得到T2G用户吞吐量奖励r_t^{T2G}和T2T用户吞吐量奖励$r_t^{n,\mathrm{T2T}}$，并得到新的观测状态$\boldsymbol{s}_{t+1}$；
(7)　计算T2G用户吞吐量TD误差：$\delta_t^{\mathrm{T2G}}=r_t^{\mathrm{T2G}}+\gamma V_{\boldsymbol{w}}^{\mathrm{T2G}}(\boldsymbol{s}_{t+1})-V_{\boldsymbol{w}}^{\mathrm{T2G}}(\boldsymbol{s}_t)$；
(8)　更新T2G用户价值网络参数: $\boldsymbol{w}_{t+1}=\boldsymbol{w}_t+\alpha_w\nabla_w V_w^{\mathrm{T2G}}(\boldsymbol{s}_t)\delta_t^{\mathrm{T2G}}$；
(9)　For n=0: N do
(10)　　计算T2T用户n吞吐量TD误差：$\delta_t^{n,\mathrm{T2T}}=r_t^{n,\mathrm{T2T}}+\gamma V_{\boldsymbol{\psi}}^{n,\mathrm{T2T}}(\boldsymbol{s}_{t+1})-V_{\boldsymbol{\psi}}^{n,\mathrm{T2T}}(\boldsymbol{s}_t)$；
(11)　　更新智能体n的价值网络参数：$\boldsymbol{\psi}_{t+1}^n=\boldsymbol{\psi}_t^n+\alpha_{\boldsymbol{\psi}}\nabla_{\boldsymbol{\psi}}V_{\boldsymbol{\psi}}^{n,\mathrm{T2T}}(\boldsymbol{s}_t)\delta_t^{n,\mathrm{T2T}}$；
(12)　　计算加权TD误差：$\delta_t^n=\beta\delta_t^{\mathrm{T2G}}+(1-\beta)\delta_t^{n,\mathrm{T2T}}$；
(13)　　更新智能体n的策略网络参数：$\boldsymbol{\theta}_{t+1}^n=\boldsymbol{\theta}_t^n+\alpha_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ln\pi_{\boldsymbol{\theta}}^n(\mathrm{RB}_t^n|\boldsymbol{s}_t)\delta_t^n$ $\boldsymbol{\theta}_{t+1}^n=\boldsymbol{\theta}_t^n+\alpha_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\ln\pi_{\boldsymbol{\theta}}^n(p_t^{n,l}|\boldsymbol{s}_t)\delta_t^n$
(14)　　End
(15)　　更新所有智能体状态：$\boldsymbol{s}_t=\boldsymbol{s}_{t+1}$
(16) End

分布，梯度优化器选用Adam算法，Actor与Critic网络设置两个隐藏层，每层64个单元，激活函数为ELU，折扣因子为0.99。图4为加权参数β与功率水平等级L关于系统吞吐量的色块图，根据图示可得在同一加权参数下，当功率水平等级L≥15时，系统吞吐量基本不变，而L取值越大，智能体动作空间越大，因此在综合考虑后，加权参数β取0.3，L取15。城市轨道交通T2T通信场景采用Winner II路径损耗模型[15]，其中T2G路径损耗模型为148+40lg(d)，T2T路径损耗模型为128.1+37.6lg(d)。考虑单小区资源分配场景，假定基站处于场景中心位置，且列车不受多普勒频移影响。

设置以下对比实验：(1)A2C算法；(2)A2C-a算法：仅改进Actor网络的A2C算法；(3)A2C-ac算法：同时改进Actor网络和Critic网络的A2C算法；(4)DQN算法。T2T通信模拟场景中列车位置等数据是随机生成的，为了保证实验有效性，在场景仿真中设置随机种子，使4种算法在同一个模拟场景中训练。当M=8, N=4时，通过奖励、RB碰撞概率、T2T成功访问率[16]和系统吞吐量等方面验证所提算法性能。

如图5所示，随着训练次数的增加，4种算法的累计奖励都在不断提高，最终达到收敛。在训练前期，智能体处于探索阶段，以随机动作为主，同时在奖励函数中添加了惩罚，使智能体犯错后能快速修正，因此奖励波动较大，甚至出现负奖励。从中可以看出，智能体在探索过程中受到惩罚时能在短

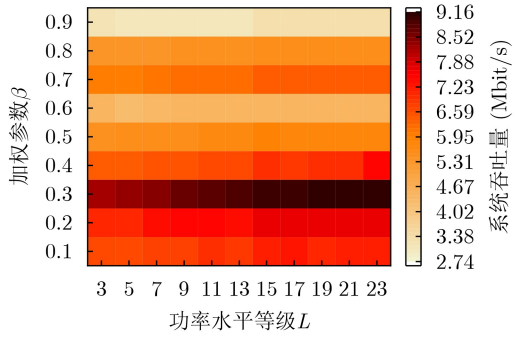

图 4 系统吞吐量色块图

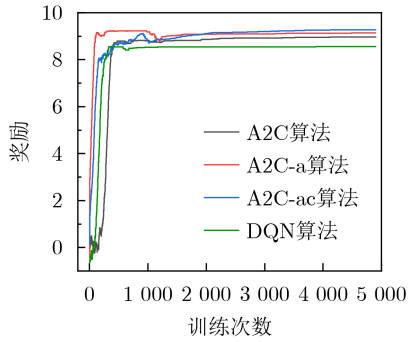

图 5 训练次数及获取平均奖励

时间内做出调整，迅速恢复正常水平。A2C-a算法和A2C-ac算法在Actor网络中采用了分层输出结构，相比于A2C算法和DQN算法，收敛速度得到有效提升。

如图6所示，随着训练次数不断增加，RB碰撞概率逐渐减小，并趋向收敛。在训练初期，智能体

随机选择RB和功率水平，发生碰撞的概率较高，但随着训练次数增加，奖励函数发挥作用，碰撞概率明显下降并收敛。但A2C-ac算法收敛速度并不如A2C-a算法，因为在训练过程中需要为每个智能体定制TD误差，提升了算法复杂度。A2C-a算法的价值网络未作改动，可以看出其收敛速度明显快于其余3种算法，但在相同训练次数内，A2C-ac算法训练结果明显优于DQN算法。

为了评估算法性能，定义T2T用户接入率为时间范围T内满足最低SINR的T2T用户数与总T2T用户数的比值。如图7所示，T2T用户接入率随训练次数的增加而不断增长，并趋向收敛。实验结果与碰撞概率实验结果相近，A2C-ac算法略逊于A2C-a算法，但其性能明显优于A2C算法和DQN算法。这是由于策略网络的分层输出结构加快了算法收敛速度，但对价值网络的改动，使A2C-ac算法的收敛速度减慢。

如图8所示，4种算法吞吐量都随着训练次数的增加不断增长，逐渐趋于稳定。A2C-ac算法收敛速度不如A2C-a算法，但是吞吐量性能最优。对每个智能体定制化TD误差，必然会提高算法复杂度，但定制化TD误差的灵活性和准确性使得系统吞吐量得到了有效提升。

如图9所示，随着T2T通信对数的增加，系统吞吐量也增加，并且所提算法具有明显优势。从仿真结果中还可以看出，系统吞吐量在T2T通信对数为3～4对时斜率最大，T2T通信对数大于4对时，系统吞吐量增长速度变慢。当T2T通信对数小于4对时，系统内干扰链路较少，通过功率控制和RB选择可以有效控制干扰；当通信对数大于4对时，T2T通信链路增加，会产生更多的复用干扰，导致系统吞吐量增长缓慢。

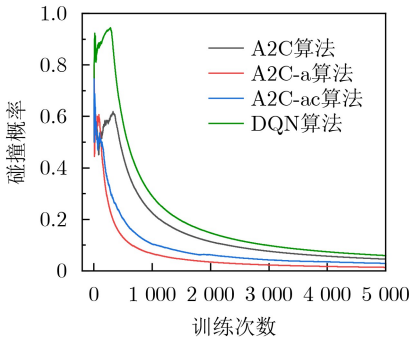

图 6 训练次数及碰撞概率

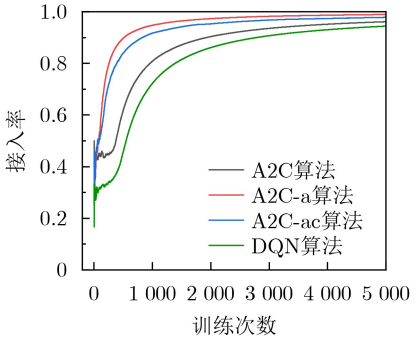

图 7 训练次数及T2T用户接入率

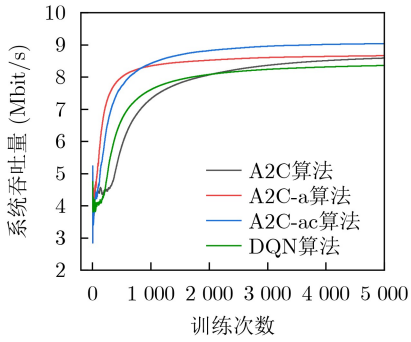

图 8 训练次数及系统吞吐量

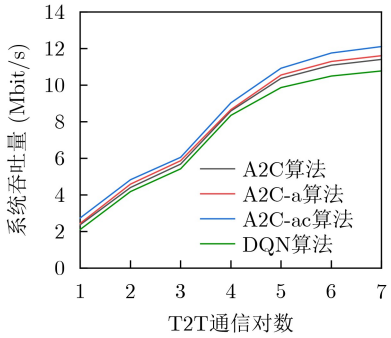

图 9 系统吞吐量及T2T通信对数

6 结论

本文针对城市轨道交通T2T通信过程中由于频谱复用带来的干扰问题进行分析，综合考虑系统整体吞吐量、发射功率限制、频谱复用要求以及用户通信质量，将T2T通信发送端视为智能体，策略网络通过分层输出结构，价值网络为每个智能体定制化TD误差，提出一种基于A2C-ac算法的多智能体资源分配策略。仿真结果表明，策略网络采用的分层输出结构明显加快了算法收敛速度，而双价值网络结构虽然提升了算法的复杂度，但相较于A2C算法和DQN算法，系统吞吐量分别提升了5.88%和11.02%。

参 考 文 献

[1] AI Bo, CHENG Xiang, KÜRNER T, *et al*. Challenges toward wireless communications for high-speed railway[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2014, 15(5): 2143–2158. doi: 10.1109/TITS.2014.2310771.

[2] 林俊亭, 王晓明, 党垚, 等. 城市轨道交通列车碰撞防护系统设

计与研究[J]. 铁道科学与工程学报, 2015, 12(2): 407–413. doi: 10.19713/j.cnki.43-1423/u.2015.02.028.
LIN Junting, WANG Xiaoming, DANG Yao, *et al*. Design and research of the improved train control system with collision avoidance system for urban mass transit[J]. *Journal of Railway Science and Engineering*, 2015, 12(2): 407–413. doi: 10.19713/j.cnki.43-1423/u.2015.02.028.

[3] 胡雪旸, 周庆华. 基于D2D的列控系统车车通信资源分配算法[J]. 铁道标准设计, 2019, 63(3): 153–157. doi: 10.13238/j.issn.1004-2954.201804080010.
HU Xueyang and ZHOU Qinghua. T2T (Train to Train) communication resource allocation algorithm based on D2D[J]. *Railway Standard Design*, 2019, 63(3): 153–157. doi: 10.13238/j.issn.1004-2954.201804080010.

[4] 申滨, 孙万平, 张楠, 等. 基于加权二部图及贪婪策略的蜂窝网络D2D通信资源分配[J]. 电子与信息学报, 2023, 45(3): 1055–1064. doi: 10.11999/JEIT220029.
SHEN Bin, SUN Wanping, ZHANG Nan, *et al*. Resource allocation based on weighted bipartite graph and greedy strategy for D2D communication in cellular networks[J]. *Journal of Electronics & Information Technology*, 2023, 45(3): 1055–1064. doi: 10.11999/JEIT220029.

[5] HU Jinming, HENG Wei, ZHU Yaping, *et al*. Overlapping coalition formation games for joint interference management and resource allocation in D2D communications[J]. *IEEE Access*, 2018, 6: 6341–6349. doi: 10.1109/ACCESS.2018.2800159.

[6] XIAO Yong, CHEN K C, YUEN C, *et al*. A Bayesian overlapping coalition formation game for device-to-device spectrum sharing in cellular networks[J]. *IEEE Transactions on Wireless Communications*, 2015, 14(7): 4034–4051. doi: 10.1109/TWC.2015.2416178.

[7] 高云波, 程璇, 李翠然, 等. T2T和T2G混合网络中的功率分配算法[J/OL]. 西南交通大学学报, 2022, 1–9. doi: 10.3969/j.issn.0258-2724.20210992.
GAO Yunbo, CHENG Xuan, LI Cuiran, *et al*. Power allocation algorithm in T2T and T2G hybrid network[J/OL]. *Journal of Southwest Jiaotong University*, 2022, 1–9. doi: 10.3969/j.issn.0258-2724.20210992.

[8] 吕宏志. 城市轨道交通车车通信资源分配算法研究[D]. [硕士论文], 兰州交通大学, 2021. doi: 10.27205/d.cnki.gltec.2021.000552.
LV Hongzhi. Research on train to train communication resource allocation algorithm of urban rail transit[D]. [Master dissertation], Lanzhou Jiaotong University, 2021. doi: 10.27205/d.cnki.gltec.2021.000552.

[9] 陈垚, 赵军辉, 张青苗, 等. 车车通信中通信模式选择与资源分配算法[J]. 计算机工程与应用, 2022, 58(10): 93–100. doi: 10.3778/j.issn.1002-8331.2012-0104.
CHEN Yao, ZHAO Junhui, ZHANG Qingmiao, *et al*. Communication mode selection and resource allocation algorithm for train-to-train communication[J]. *Computer Engineering and Applications*, 2022, 58(10): 93–100. doi: 10.3778/j.issn.1002-8331.2012-0104.

[10] TAN Junjie, LIANG Yingchang, ZHANG Lin, *et al*. Deep reinforcement learning for joint channel selection and power control in D2D networks[J]. *IEEE Transactions on Wireless Communications*, 2021, 20(2): 1363–1378. doi: 10.1109/TWC.2020.3032991.

[11] ZHAO Junhui, ZHANG Yang, NIE Yiwen, *et al*. Intelligent resource allocation for train-to-train communication: A multi-agent deep reinforcement learning approach[J]. *IEEE Access*, 2020, 8: 8032–8040. doi: 10.1109/ACCESS.2019.2963751.

[12] 赵军辉, 陈垚, 张青苗. 基于深度强化学习的车车通信智能频谱共享[J]. 铁道科学与工程学报, 2022, 19(3): 841–848. doi: 10.19713/j.cnki.43-1423/u.t20210364.
ZHAO Junhui, CHEN Yao, and ZHANG Qingmiao. Intelligent spectrum sharing for train-to-train communication based on deep reinforcement learning[J]. *Journal of Railway Science and Engineering*, 2022, 19(3): 841–848. doi: 10.19713/j.cnki.43-1423/u.t20210364.

[13] 唐伦, 贺小雨, 王晓, 等. 基于异步优势演员-评论家学习的服务功能链资源分配算法[J]. 电子与信息学报, 2021, 43(6): 1733–1741. doi: 10.11999/JEIT200287.
TANG Lun, HE Xiaoyu, WANG Xiao, *et al*. Resource allocation algorithm of service function chain based on asynchronous advantage actor-critic learning[J]. *Journal of Electronics & Information Technology*, 2021, 43(6): 1733–1741. doi: 10.11999/JEIT200287.

[14] 刘伟, 郑润泽, 张磊, 等. 基于A2C算法的低轨星座动态波束资源调度研究[J]. 中国空间科学技术, 2023, 43(3): 123–133. doi: 10.16708/j.cnki.1000-758X.2023.0045.
LIU Wei, ZHENG Runze, ZHANG Lei, *et al*. Research of dynamic beam resource scheduling of LEO constellation based on A2C algorithm[J]. *Chinese Space Science and Technology*, 2023, 43(3): 123–133. doi: 10.16708/j.cnki.1000-758X.2023.0045.

[15] WANG Xiaoxuan, LIU Liangjia, TANG Tao, *et al*. Enhancing communication-based train control systems through train-to-train communications[J]. *IEEE Transactions on Intelligent Transportation Systems*, 2019, 20(4): 1544–1561. doi: 10.1109/TITS.2018.2856635.

[16] SUN Zhenfeng and NAKHAI M R. Channel selection and power control for D2D communication via online reinforcement learning[C]. 2021 IEEE International Conference on Communications, Montreal, Canada, 2021: 1–6. doi: 10.1109/ICC42927.2021.9501055.

王瑞峰：女，教授，研究方向为城市轨道交通车车通信技术.
张　明：男，硕士生，研究方向为城市轨道交通车车通信技术.
黄子恒：男，硕士生，研究方向为交通信息工程及控制.
何　涛：男，教授，研究方向为交通信息工程及控制.

责任编辑：余　蓉