基于强化学习的无人机基站多播通信系统的飞行路线在线优化

张广驰[①] 严雨琳[①] 崔 苗*[①] 陈 伟[②] 张 景[③]

[①](广东工业大学信息工程学院 广州 510006)

[②](广东省环境地质勘查院 广州 510080)

[③](中国电子科学研究院 北京 100043)

摘 要：针对无人机(UAV)基站(BS)多播通信系统的通信时延最小化问题，该文提出飞行路线在线优化算法。在该系统中无人机基站向多个地面用户同时发送公共信息，其中每次通信任务中地面用户位置是随机的。为了保证地面用户能够接收完整的公共信息以及考虑到无人机的能量有限性，该文以最小化无人机基站完成通信任务的平均时间为目标。首先将问题转化成一个马尔可夫决策过程(MDP)；然后把通信时延引入到动作价值函数中；最后提出使用Q-Learning算法对无人机飞行路线进行学习和在线优化，从而实现平均通信时延最小化。仿真结果显示，与其他基准方案相比，该文所提方案能够有效地为无人机多播通信系统飞行路线实现在线优化，并有效降低通信任务的完成时间。

关键词：无人机基站；飞行路线在线优化；强化学习

中图分类号：TN915 **文献标识码**：A **文章编号**：1009-5896(2022)03-0969-07

DOI: 10.11999/JEIT210429

Online Trajectory Optimization for the UAV-Enabled Base Station Multicasting System Based on Reinforcement Learning

ZHANG Guangchi[①] YAN Yulin[①] CUI Miao[①] CHEN Wei[②] ZHANG Jing[③]

[①](*School of Information Engineering*, *Guangdong University of Technology*, *Guangzhou* 510006, *China*)

[②](*Institute of Environmental Geology Exploration of Guangdong Province*, *Guangzhou* 510080, *China*)

[③](*China Academic of Electronics and Information Technology*, *Beijing* 100043, *China*)

Abstract: In order to deal with the communication delay problem in an Unmanned Aerial Vehicle (UAV) enabled Base Station (BS) multicasting communication system, the online trajectory design for the UAV BS is investigated. A UAV BS is dispatched to disseminate common information to multiple ground users simultaneously in this system, where the locations of the ground users are random in each multicasting communication task. To ensure that the ground users can receive the complete multicasting information and considering the limited energy of the UAV, this paper focuses on minimizing the average duration for the UAV BS to complete the multicasting task. First, the considered problem is casted as a Markov Decision Process (MDP), and then the communication delay is introduced into the action value function. Finally, an online trajectory optimization algorithm based on the Q-Learning algorithm is proposed to minimize the average duration for the UAV BS to complete the multicasting task. Simulation results show that the proposed algorithm can effectively optimize the trajectory of the UAV BS for its multicasting task in an online manner and can effectively reduce the duration of the multicast task, as compared to other benchmark schemes.

Key words: Unmanned Aerial Vehicle (UAV) Base Station (BS); Online trajectory optimization; Reinforcement learning

收稿日期：2021-05-19；改回日期：2021-09-16；网络出版：2021-12-25

*通信作者： 崔苗 cuimiao@gdut.edu.cn

基金项目：广东省科技计划项目(2020A050515010, 2021A0505030015, 2019B010119001)，广东特支计划项目(2019TQ05X409)

Foundation Items: The Science and Technology Plan Project of Guangdong Province (2020A050515010, 2021A0505030015, 2019B010119001), The Special Support Plan for High-Level Talents of Guangdong Province (2019TQ05X409)

1 引言

无人驾驶飞机(Unmanned Aerial Vehicle, UAV)简称无人机，在近十年内得到巨大的发展，其商业价值预计在2025年飙升到45.8亿美元[1]。无人机自身具有高移动性、机动性、体积小以及成本低等特点，使其在无线通信方面引起了广泛的关注。无人机融入无线通信网络的方式分为以下3类。第一，无人机作为空中基站为无线蜂窝网络补充覆盖和提升容量[2]，或者在发生大范围自然灾害时快速适应环境为地面用户提供应急通信[3]。第二，无人机作为辅助中继改善地面无线设备的连接，极大地扩宽通信范围以及提高通信质量[4]。第三，将无人机接入到物联网中提供可靠和节能的物联网上行通信链路，物联网网络的连通性和能源效率可以显著提高[5]。

本文主要研究上述第1类应用方式，即无人机作为空中基站为地面用户提供无线通信服务。通常地面基站的部署建设是根据长期通信行为来统筹规划的，可能无法满足短时间人群聚集(例如演唱会等)的通信需求和无法适应未来的通信环境变化。相比于传统地面的基站，无人机基站的机动性带来明显的优势，能够灵活便捷地调整位置适应通信需求和为流量热点区域提供额外的网络负载能力[6]。为了充分发挥无人机的机动性潜能，适当的轨迹优化可以减小无人机基站与地面用户之间的距离从而改善信道质量，这对提高通信网络的性能至关重要。文献[7]通过优化无人机的飞行轨迹以及资源分配从而实现高效节能的通信。文献[8]从多无人机的角度出发，考虑了各个无人机与多地面用户之间的干扰，通过优化多无人机的飞行轨迹实现吞吐量最大化。文献[9]研究了在无人机支持的多链路中继系统中，联合优化无人机的3维飞行轨迹和发射功率，抑制链路中的干扰以达到下界吞吐量最大化。以上文献中无人机飞行轨迹优化采用的算法都是属于离线优化算法，即根据通信环境的完美假设，在无人机起飞之前通过复杂的计算、优化设计得到无人机的飞行轨迹，并且起飞之后无法改变飞行轨迹。然而在实际中，通信环境是不断变化的，无法提前预测的，通信环境的完美假设是不切实际的。离线优化算法首先需要建立精确的通信模型，建模之后的参数配置也是难以获取的，即使模型和相关参数是已知的，大多数无线通信的优化问题都是非凸的，通常需要复杂的运算和推导将其转化成凸问题[10]。

为了克服这些局限性，文献[11,12]分别讨论了将强化学习中的算法应用于无人机通信方面的可能性，将无人机的飞行轨迹优化看作路径规划问题，其目标是在随机的飞行环境中最大化特定的累计奖励指标[13]。文献[11]研究了多无人机基站协作通信的场景，以最大化地面用户的通信速率之和为目标，提出了基于强化学习Q-Learning算法的多无人机飞行轨迹优化。文献[12]提出了一种基于体验质量(Quality of Experience, QoE)驱动的多无人机3维部署与飞行轨迹设计新框架。目前已有研究工作开始将强化学习算法应用于解决无人机的飞行路线优化问题，但是关注无人机基站的通信时延和能效问题的研究不多。同时多播通信方式能够在公共安全、应急响应以及智能交通等应用方面减轻无线通信网络的负载和提高通信效率，因此研究无人机基站多播通信系统很有必要[14]。

本文研究了无人机基站多播通信系统中通信时延问题，在该系统中无人机基站向多个地面用户同时发送公共信息，其中每次通信任务中地面用户的位置是随机的。首先建立系统模型，为了保证地面用户能够接收到完整的信息以及减少无人机的能量消耗，以最小化通信任务平均完成时间为目标，对无人机基站飞行路线在线优化问题进行数学描述。然后将问题转化成马尔可夫决策过程，采用强化学习中的Q-Learning算法实现飞行路线在线优化。最后通过仿真验证本文提出的飞行路线在线优化算法的有效性。

2 系统模型

如图1所示，本文考虑一个无人机基站多播通信系统，其中包括一个无人机和K个地面用户[1)]。无人机作为空中通信基站为矩形区域内的K个地面

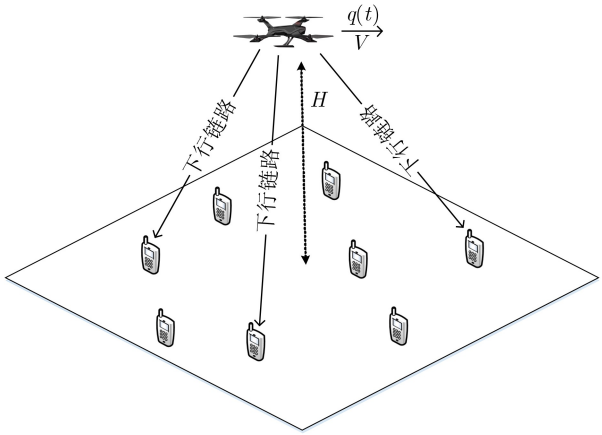

图 1 无人机基站多播通信系统

[1)]本文主要研究无人机基站的飞行路线在线优化对多播通信性能的影响，为简单起见，暂时没有考虑无人机基站的能耗问题，考虑能耗的无人机基站飞行路线在线优化将是未来的研究方向之一。另外，本文考虑的系统模型可以扩展到多个无人机协作多播通信的场景。

用户提供多播通信服务和发送公共信息。无人机地面用户可表示集合$\mathcal{K}=\{1,2,\cdots,K\}$，地面用户的位置可用2维坐标表示为$q_k=(x_k,y_k)$。假设无人机的飞行高度固定在$H$ m，无人机在t时刻的地面投影位置的坐标可表示为$q(t)=(x(t),y(t))$，无人机在飞行过程的飞行速度固定在V m/s。本文采用FDMA (Frequency Division Multiple Access)通信方式，无人机基站与地面用户的信道数量为J，即无人机基站可同时向J个地面用户发送公共信息，且无人机基站与各个地面用户的通信链路之间不存在干扰。同时通信的J个地面用户可用集合表示为$\mathcal{J}=\{1,2,\cdots,J\},\mathcal{J}\in\mathcal{K}$，其位置坐标为$q_j=(x_j,y_j),j\in\mathcal{J}$。因此无人机基站与地面用户$j$的距离可表示为

$$d_j(t)=\sqrt{H^2+\|q(t)-q_j\|^2} \tag{1}$$

假设地面用户和无人机基站之间的信道由视距信道[15]主导[2)]，无人机的移动性所导致的多普勒效应能够被地面用户的接收机有效补偿，因此无人机基站与地面用户j的信道增益为

$$h_j(t)=\beta_0 d_j^{-2}(t)=\frac{\beta_0}{H^2+\|q(t)-q_j\|^2} \tag{2}$$

其中，β_0为参考距离为1 m的信道增益。假设无人机基站与每个地面用户的信道带宽为B，噪声为σ^2，无人机基站与地面用户j的瞬时通信传输速率可表示为

$$R_j(t)=B\log_2\left(1+\frac{\gamma_0}{H^2+\|q(t)-q_j\|^2}\right) \tag{3}$$

其中，$\gamma_0=\dfrac{P_j\beta_0}{\sigma^2}$, P_j为无人机基站与地面用户j的通信的发射功率。

3　飞行路线在线优化算法

3.1　问题描述

为了保证地面用户能够接收到完整的文件信息以及考虑到无人机的能量有限性，本文以最小化无人机基站完成通信任务的平均时间为目标。无人机每次通信任务中所服务的地面用户是随机的，因此在线优化无人机的飞行路线很有必要。本文主要考察飞行路线对通信性能的影响，因此暂时不考虑无人机基站的能耗，假设飞行时间足够长。无人机基站在第m次通信任务中需同时给J个地面用户传输文件信息，且与每个地面用户传输文件信息量为L bit。当无人机基站完成第m次通信任务中所有地面用户所需的信息量之后，才能开始进行第$m+1$次通信任务，为另外J个地面用户发送公共文件信息。换句话说，无人机基站同时与J个地面用户通信，其中通信时延最大的地面用户的通信任务完成时，其他的地面用户的通信任务已完成。将无人机基站完成第m次通信任务的时间表示为$T_m=\max\{T_{m,1},T_{m,2},\cdots,T_{m,j}\}$，$T_{m,j}$表示无人机基站第$m$次通信任务中与第$j$个地面用户的通信时延。无人机基站在第$m$次通信任务中与第$j$个地面用户的通信速率可用$R_{m,j}$表示，在第$m$次通信任务中，无人机基站需与每个地面用户传输$L$ bit信息量可表示为

$$\int_0^{T_{m,j}} R_{m,j}(t)\,\mathrm{d}t\geq L,\forall j\in J,m\in\{1,2,\cdots,M\} \tag{4}$$

无人机基站多播通信系统中的飞行路线在线优化问题可表示成P1

$$\mathrm{P1}:\min_{\{q_m(t)\}} D=\frac{\sum_{m=1}^{M}T_m}{M} \tag{5}$$

$$\text{s.t.}\int_0^{T_{m,j}} R_{m,j}(t)\,\mathrm{d}t\geq L,\forall j\in J,m\in\{1,2,\cdots,M\},$$

$$X_{\min}\leq x_m(t)\leq X_{\max},\forall m\in\{1,2,\cdots,M\}, \tag{6}$$

$$Y_{\min}\leq y_m(t)\leq Y_{\max},\forall m\in\{1,2,\cdots,M\}, \tag{7}$$

$$0\leq\|q'_m(t)\|\leq V_{\max},\forall m\in\{1,2,\cdots,M\} \tag{8}$$

式(5)为目标函数，表示无人机基站完成M次通信任务的平均完成时间最小化；式(6)和式(7)为无人机基站的飞行范围约束；式(8)表示对无人机飞行速率的约束，其中$q'_m(t)$表示无人机在第m次通信任务中t时刻的飞行速率。

3.2　强化学习概述

强化学习具有高效的自我学习能力，可用于解决无人机通信网络中的优化问题。因此本文将采用强化学习中的算法对无人机基站的飞行路线进行在线优化，接下来将介绍强化学习的理论知识。强化学习以交互目标为导向，将智能体置身于环境中并与其进行交互，在此情境中，给智能体所选择的动作赋予奖赏，以智能体在交互过程中所得到的累计奖赏最大化为目标从而指导其行为[16]。强化学习中的大多数问题都可以转化成马尔可夫决策过程(Markov Decision Process, MDP)，因此马尔可夫决策过程是强化学习的基础理论。MDP的基本框架为$(\mathcal{S},\mathcal{A},\mathcal{R})$，每个离散时刻$t$可以观察到智能体的状态为$S_t\in\mathcal{S}$，然后在此状态上选择并执行一个动作$A_t\in\mathcal{A}(s)$。环境会对智能体所选择的动作进行

2)为了便于完善在线优化算法理论和检验算法的性能，本文采用了LoS空地（地空）信道模型。在未来的工作中，可以将本文提出的算法直接扩展到其他更准确的信道模型上。

反馈，然后智能体会接收到一个数值化的即时奖赏$R_{t+1}\in\mathcal{R}\subset\mathbb{R}$，并进入一个新的状态$S_{t+1}$。由这一系列状态和动作构成了智能体的策略$\pi$[17]。强化学习的目标是最大化长期交互过程中累计奖赏$G_t=\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}$，将$G_t$定义为回报，其中$\gamma$是一个参数，$0\le\gamma\le1$，称为折扣率。折扣率表示智能体此时采取的动作对未来时刻的奖赏的影响效果，当$\gamma=0$时表示智能体的动作只能影响当前的奖赏，不能对接下来的交互过程所得到奖赏产生影响，相反，γ越接近1，表示智能体此时选择的动作对接下来的交互过程的影响越大。为了使回报G_t最大化，可采用动作价值函数对当前采取策略下的“状态-动作”估计其“价值”。动作价值函数定义为智能体根据策略π，从状态s开始，执行动作a之后，所有可能的决策序列的回报期望值，可表示为

$$\begin{aligned}Q_\pi(s,a)&=\mathbb{E}_\pi\left[G_t\,|S_t=s,A_t=a\right]\\&=\mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}\,|S_t=s,A_t=a\right]\end{aligned}\tag{9}$$

解决一个强化学习问题意味对其找到一个最优策略，其对应的最优动作价值函数可写成

$$\begin{aligned}Q_*(s,a)&\doteq\max_\pi q_\pi(s,a)\\&=\mathbb{E}\left[R_{t+1}+\gamma\max_{a'}q_*(S_{t+1},a')\,|S_t=s,A_t=a\right]\end{aligned}\tag{10}$$

接下来把P1问题进行离散化，将其转化成一个MDP。首先将无人机基站在第m次通信任务中与第j个地面用户通信的完成通信任务的时间$T_{m,j}$进行离散化，可表示为$T_{m,j}=N_{m,j}\delta_t$，则无人机基站完成第m次通信任务的时间离散为$T_m=N_m\times\delta_t$，其中$N_m=\max\{N_{m,1},N_{m,2},\cdots,N_{m,j}\}$。假设时隙$\delta_t$需要足够小，使其在这段时间内可以将无人机基站的通信速率$R_{j,m}[n]$看作保持不变。无人机的飞行路线$q_m(t)$可离散成$q_m[n]=q_m(n\delta_t)$，以及飞行速度$q'_m(t)$可离散表示为$q'_m[n]=q'_m(n\delta_t)$。因此问题P1可改写为P2

$$\mathrm{P2}:\min_{\{q_m[n]\}}\ N=\frac{\sum_{m=1}^{M}N_m}{M}\tag{11}$$

$$\text{s.t.}\sum_{n=1}^{N_{m,j}}R_{m,j}[n]\cdot\delta_t\ge L,\forall j\in\mathcal{J},m\in\{1,2,\cdots,M\}\tag{12}$$

$$X_{\min}\le x_m[n]\le X_{\max},\forall m\in\{1,2,\cdots,M\}\tag{13}$$

$$Y_{\min}\le y_m[n]\le Y_{\max},\forall m\in\{1,2,\cdots,M\}\tag{14}$$

$$0\le\|q'_m[n]\|\le V_{\max},\forall m\in\{1,2,\cdots,M\}\tag{15}$$

式(11)—式(15)为问题P1的离散形式。问题P2所对应的MDP的描述如下：

状态：把无人机基站的位置坐标设置为状态。无人机基站的可飞行的矩形范围$[X_{\min},X_{\max}]\times[Y_{\min},Y_{\max}]$分割成$I\times I$个格子，其中$I=\dfrac{X_{\max}-X_{\min}}{V_{\max}\delta_t}$，保证无人机基站在每个格子内通信传输速率可以看作保持不变。将状态的位置坐标用格子中心来表示，则在X轴上第k_1时隙、Y轴上第k_2时隙的格子位置坐标可表示为$(x_{k_1},y_{k_2})=\left(X_{\min}+\dfrac{(X_{\max}-X_{\min})}{I}\times(k_1-1),Y_{\min}+\dfrac{(Y_{\max}-Y_{\min})}{I}\times(k_2-1)\right)$。

动作：无人机基站的动作集合包括5个动作：向东、向西、向南、向北以及保持当前位置不动。

奖赏：定义为无人机基站在进行通信任务时，当前J个地面用户的通信传输速率中的最小值[3)]，可表示为

$$r_m[n]=\min\left\{\left.\frac{\sum_{n=1}R_{m,j}[n]\cdot\delta_t}{L}\cdot R_{m,j}[n]\right|j\in\mathcal{J}\right\}\tag{16}$$

3.3 基于Q-Learning的飞行路线在线优化算法

本文所提出的无人机飞行路线在线优化问题中，无人机的每个动作不仅影响当前的性能，还会对接下来的状态产生影响。因此本文采用强化学习中的Q-Learning算法对问题进行求解。Q-Learning是一种典型的强化学习中离轨策略下的时序差分算法，可以在每个动作结束之后估计动作价值函数并更新改进策略。Q-Learning中采取的动作策略为ε - greedy策略，是对贪婪策略的改进。ε - greedy策略具有ε的概率探索环境寻找更优的策略，$1-\varepsilon$的概率按照贪婪思想选择动作价值函数最大的动作。动作价值函数定义为：$Q(s_n,a_n)=Q(s_n,a_n)+\alpha[r_{n+1}+\gamma\max_a Q(s_{n+1},a)-Q(s_n,a_n)]$。基于Q-Learning的在线优化算法具体如下：

步骤1　初始化探索参数N_ε，设置折扣率γ和学习率α，无人机的通信次数M，最大训练幕数N_{epi}，每幕中最大步数N_{step}，动作价值函数$Q(s,a)=0,\forall s\in S,a\in A$；

步骤2　$M=M-1$；

步骤3　随机J个地面用户发送通信请求，获取J个用户的位置坐标；$n_{\text{epi}}=0$；

步骤4　$n_{\text{epi}}=n_{\text{epi}}+1$；

步骤5　$\varepsilon=N_\varepsilon{}^{n_{\text{epi}}}$，其中$N_\varepsilon$为探索参数的初始

3)未来的研究中，若考虑多个无人机协作多播通信的场景，可以考虑在奖赏函数中设置干扰项以及设置防碰撞约束。

值，根据无人空中通信平台的位置坐标初始化状态s_0; $n_{\text{step}}=0$;

步骤6　$n_{\text{step}}=n_{\text{step}}+1$;

步骤7　根据ε - greedy策略选择动作a_t，根据公式(15)得到奖赏r_{n+1}，观察到新的状态s_{t+1}；更新动作价值函数

$$Q(s_n,a_n)=Q(s_n,a_n)+\alpha\left[r_{n+1}+\gamma\max_a Q(s_{n+1},a)-Q(s_n,a_n)\right];$$

步骤8　重复步骤6和步骤7，直到$n_{\text{step}}=N_{\text{step}}$结束；

步骤9　重复步骤4—步骤8，直到$n_{\text{epi}}=N_{\text{epi}}$结束；

步骤10　重复步骤2—步骤9，直到$M=0$结束。

步骤5中的训练参数ε随着步骤4中的训练幕数n_{epi}的增大而减小，这让训练前期具有较多的探索学习的机会，使得智能体在训练过程中能够找到最优策略。

4　仿真结果

在本部分中，利用仿真平台对所提出的飞行路线在线优化算法进行验证，将基于Q-Learning算法的在线优化算法表示为Scheme A，并与另外3种方案进行对比。

Scheme B：无人机基站总是向着当前J个地面用户中通信传输速率最大的地面用户的方向飞行。无人机基站完成了该地面用户的文件信息传输之后，在当前位置再向着通信传输速率第二大的地面用户飞行。依次类推，直到完成所有地面用户的文件信息传输。

Scheme C：与Scheme B相反，无人机基站向着当前J个地面用户中通信传输速率最小的地面用户的方向飞行。无人机基站完成了该地面用户的文件信息传输之后，从当前位置向着通信传输速率第二小的地面用户飞行。依次类推，直到完成所有地面用户的文件信息传输。

Scheme D：在接收到J个地面用户的通信请求之后，无人机基站在每个状态位置上，贪婪地向通信传输速率最小的地面用户飞行，直到完成所有地面用户的文件信息传输。

无人机基站多播通信系统仿真参数设置如下：无人机基站可飞行的矩形范围为400m × 400 m，地面用户随机分布在此范围内。矩形范围对应的位置坐标为$[X_{\min},X_{\max}]\times[Y_{\min},Y_{\max}]=[0,400]\times[0,400]$，将矩形范围分割成$I\times I=2500$个状态。无人机的飞行高度$H=100$ m，最大飞行速度$V_{\max}=20$ m/s。无人机基站与地面用户的子信道数量$J=3$，其子信道的带宽$B=1$ MHz，地面用户的通信请求信息量$L=10^7$ bit，参考距离1 m的信噪比$\gamma_{\text{dB}}=40$ dB。假设无人机基站的通信任务次数$M=100$，其他参数：$N_{\text{epi}}=7\times10^5$, $N_{\text{step}}=120$, $\alpha=0.8$, $\gamma=0.5$, $N_\varepsilon=0.9999$。

图2展示了无人机基站两次完成通信任务的训练过程，其中完成通信任务的时间随着训练次数增大而变化。在这两次训练过程中，完成通信任务中的服务对象是不同的地面用户，其位置是随机的。与其他的方案对比，基于Q-Learning算法的在线优化算法能够有效地收敛，并且收敛之后完成通信任务的时间更小。整体来看，随着训练幕数的增加，无人机基站完成通信任务的时间越小；在训练前期可以看到无人机基站完成通信任务的时间大范围震荡，这是因为前期的探索参数ε较大，具有更大的概率探索新的动作；在训练后期，完成通信任务的时间趋向稳定，这是因为探索参数ε较小且已找到最优的飞行路线。

图3和图4展示了基于Q-Learning算法的在线优化设计算法与其他3种方案的无人机基站飞行路线对比图，飞行路线所需的时间与图2中完成两次通信任务时间相对应。图3中图例“Scheme A：n”表示Scheme A方案下无人机基站第n次完成通信任务的飞行路线，“Scheme B：n”等图例与“Scheme A：n”类似。图4是在图3的基础上完成

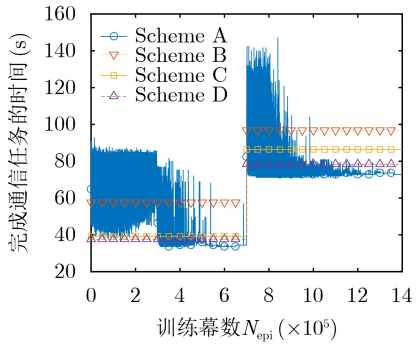

图 2 基于Q-Learning算法的在线优化设计算法的训练过程

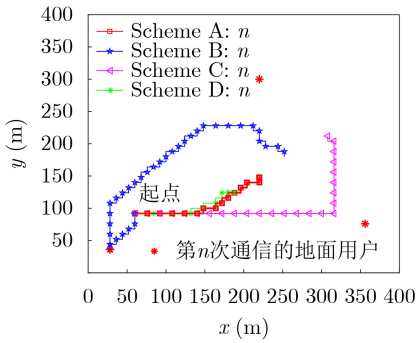

图 3 不同方案下的无人机基站飞行路线对比图

的，其中“Scheme A：n+1”表示Scheme A方案下无人机基站第n+1次完成通信任务的飞行路线。可以看出基于Q-Learning算法的在线优化设计算法比其他方案更加集中于3个地面用户的中央。Scheme A和Scheme D的飞行路线类似，但是对比图2中Scheme A与Scheme D的两次完成通信任务的时间，可以看出Scheme A比Scheme D的完成通信任务的时间更短，因此Scheme A的飞行路线更佳。

图5展示了无人机基站采用不同方案、完成不同通信信息量任务的平均时间对比图，其中完成通信任务的次数为100次。为了更好地对比效果，图5中无人机在不同方案中是完成相同的多个地面用户的通信任务，这是因为不同的地面用户位置可能导致通信任务的完成时间不同。可以看出本文提出的Scheme A方案始终优于其他3种方案，通信任务的信息量越大，Scheme A方案的性能越好。

图6展示了不同方案下的无人机基站完成100个地面用户的通信任务的平均时间，其中每次通信任务的3个地面用户是随机的，对应的通信任务的信息量为$L=10^7$ bit。可以看出Scheme A方案下完成通信任务的时间明显比其他3种方案的更小。因

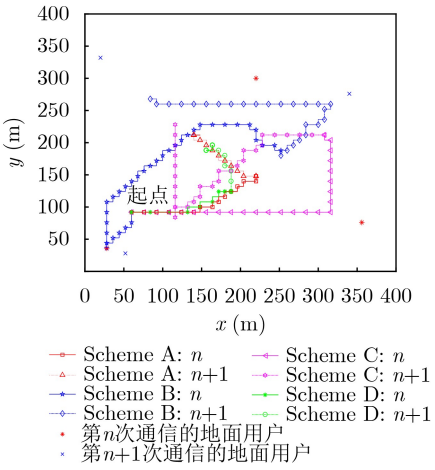

图 4 不同方案下的无人机基站飞行路线对比图

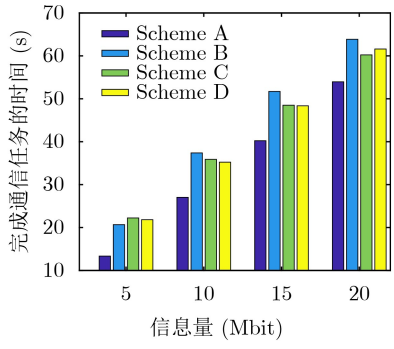

图 5 不同方案下的无人机基站完成不同信息量的多播任务时的平均时间对比图

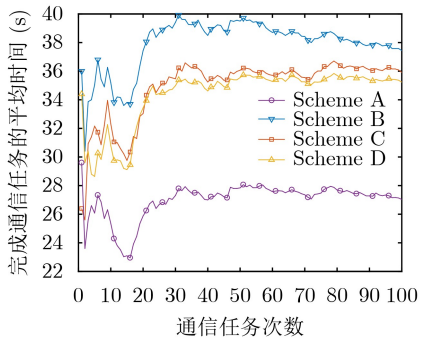

图 6 不同方案下的无人机基站完成100次通信任务的平均时间

为每次通信的地面用户是随机的，由此可以说明Scheme A方案可适应动态的、随机的地面用户的通信请求。

5 结束语

本文针对于无人机基站多播通信系统，为了保证地面用户能够接收到完整的信息以及减少无人机的能量消耗，以最小化完成多播通信任务的时间为目标，提出了基于Q-Learning的无人机飞行路线在线优化算法。仿真结果显示了与其他几种方案对比，所提出的算法能够有效实现无人机基站的飞行路线在线优化。本文的研究证实了强化学习能有效解决无人机基站飞行路线的在线优化问题，加深了我们对在线优化研究的认识。在未来的研究中，有待于将本文考虑的单无人机系统扩展到多个无人机协作多播通信的场景，并将无人机的飞行能耗纳入优化的考虑因素。

参 考 文 献

[1] WU Qingqing, XU Jie, ZENG Yong, *et al*. A comprehensive overview on 5G-and-beyond networks with UAVs: From communications to sensing and intelligence[J]. *IEEE Journal on Selected Areas in Communications*, 2021, 39(10): 2912–2945. doi: 10.1109/JSAC.2021.3088681.

[2] LYU Jiangbin, ZENG Yong, and ZHANG Rui. UAV-aided offloading for cellular hotspot[J]. *IEEE Transactions on Wireless Communications*, 2018, 17(6): 3988–4001. doi: 10.1109/TWC.2018.2818734.

[3] FENG Wanmei, TANG Jie, ZHAO Nan, *et al*. NOMA-based UAV-aided networks for emergency communications[J]. *China Communications*, 2020, 17(11): 54–66. doi: 10.23919/JCC.2020.11.005.

[4] ZENG Yong, ZHANG Rui, and LIM T J. Throughput maximization for UAV-enabled mobile relaying systems[J]. *IEEE Transactions on Communications*, 2016, 64(12): 4983–4996. doi: 10.1109/TCOMM.2016.2611512.

[5] MOZAFFARI M, SAAD W, BENNIS M, *et al*. Mobile Unmanned Aerial Vehicles (UAVs) for energy-efficient

internet of things communications[J]. *IEEE Transactions on Wireless Communications*, 2017, 16(11): 7574–7589. doi: 10.1109/TWC.2017.2751045.

[6] WANG Zhe, DUAN Lingjie, and ZHANG Rui. Adaptive deployment for UAV-aided communication networks[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(9): 4531–4543. doi: 10.1109/TWC.2019.2926279.

[7] ZENG Yong, XU Jie, and ZHANG Rui. Energy minimization for wireless communication with rotary-wing UAV[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(4): 2329–2345. doi: 10.1109/TWC.2019.2902559.

[8] WU Qingqing, ZENG Yong, and ZHANG Rui. Joint trajectory and communication design for multi-UAV enabled wireless networks[J]. *IEEE Transactions on Wireless Communications*, 2017, 17(3): 2109–2121. doi: 10.1109/TWC.2017.2789293.

[9] LIU Tianyu, CUI Miao, ZHANG Guangchi, *et al*. 3D trajectory and transmit power optimization for UAV-enabled multi-link relaying systems[J]. *IEEE Transactions on Green Communications and Networking*, 2021, 5(1): 392–405. doi: 10.1109/TGCN.2020.3048135.

[10] ZENG Yong and XU Xiaoli. Path design for cellular-connected UAV with reinforcement learning[C]. 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, USA, 2019: 1–6. doi: 10.1109/GLOBECOM38437.2019.9014041.

[11] KHAMIDEHI B and SOUSA E S. Reinforcement learning-based trajectory design for the aerial base stations[C]. The 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Istanbul, Turkey, 2019: 1–6. doi: 10.1109/PIMRC.2019.8904880.

[12] LIU Xiao, LIU Yuanwei, and CHEN Yue. Reinforcement learning in multiple-UAV networks: Deployment and movement design[J]. *IEEE Transactions on Vehicular Technology*, 2019, 68(8): 8036–8049. doi: 10.1109/TVT.2019.2922849.

[13] SAXENA V, JALDÉN J, and KLESSIG H. Optimal UAV base station trajectories using flow-level models for reinforcement learning[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2019, 5(4): 1101–1112. doi: 10.1109/TCCN.2019.2948324.

[14] ZENG Yong, XU Xiaoli, and ZHANG Rui. Trajectory design for completion time minimization in UAV-enabled multicasting[J]. *IEEE Transactions on Wireless Communications*, 2018, 17(4): 2233–2246. doi: 10.1109/TWC.2018.2790401.

[15] GOLDSMITH A. Wireless Communications[M]. Cambridge: Cambridge University Press, 2005: 26–27.

[16] SUTTON R S and BARTO A G. Reinforcement Learning: An Introduction[M]. Cambridge: MIT Press, 2018: 1–130.

[17] BELLMAN R. A markovian decision process[J]. *Journal of Mathematics and Mechanics*, 1957, 6(5): 679–684. doi: 10.1512/iumj.1957.6.56038.

张广驰：男，1982年生，教授，研究方向为新一代无线通信技术.

严雨琳：女，1996年生，硕士生，研究方向为无人机通信、强化学习.

崔　苗：女，1978年生，讲师，研究方向为新一代无线通信技术.

陈　伟：男，1979年生，高级工程师，研究方向为地质灾害监测与预警.

张　景：男，1974年生，研究员级高工，研究方向为新一代信息通信技术.

责任编辑：马秀强