无人机基站的飞行路线在线优化设计

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：针对离线的无人机(UAV)基站飞行路线设计无法满足随机的、动态的地面用户通信请求难题，该文研究了飞行路线在线优化设计算法。考虑单个无人机空中基站为两个地面用户提供无线通信服务，通过在线实时优化无人机的飞行路线实现最小化与地面用户的平均通信时延。首先，由于系统的无人机的状态和动作是连续的，将问题转化成一个马尔可夫决策过程(MDP)；然后，把单次通信时延引入到动作价值函数中；最后分别采用强化学习中蒙特卡罗和Q-Learning算法来实现无人机的飞行路线在线优化。仿真结果表明，所提出的在线优化的平均时延性能优于“固定位置”和“贪婪算法”的时延计算结果。

关键词：无人机通信；飞行路线在线优化；平均时延最小化；强化学习

中图分类号：TN915 **文献标识码**：A **文章编号**：1009-5896(2021)12-3605-07

DOI: 10.11999/JEIT200525

Online Trajectory Optimization for the UAV-Mounted Base Stations

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: Considering dealing with the problem of random and dynamic communication requests of ground users in a UAV(Unmanned Aerial Vehicle) mounted base station communication system, which can not be tackled by an offline trajectory design scheme, an online trajectory optimization algorithm is proposed for the UAV-mounted base station. In the considered system, a single UAV is utilized as an aerial base station to provide wireless communication service to two ground users. The problem of minimizing the average communication delay of the ground users via optimizing the UAV's trajectory is considered. First, it is shown that the problem can be casted as a Markov Decision Process (MDP), and then the delay of one single communication is introduced into the action value function. Finally, the Monte Carlo and Q-Learning algorithms from the reinforcement learning technology are respectively adopted to realize the online trajectory optimization. Simulation results show that the proposed algorithm outperforms the “fixed position” and “greedy algorithm” schemes.

Key words: Unmanned Aerial Vehicle (UAV) communication; Online trajectory optimization; Average delay minimization; Reinforcement learning

1 引言

在过去的十年中，由于无人机(Unmanned Aerial Vehicle, UAV)移动性高、成本低等特点，无人机在无线通信领域引起了广泛的关注[1]。无人机在即将到来的5G通信时代也会发挥重要的作用，主要可以分为两类。第1类，无人机可以作为空中移动的通信平台辅助地面基站的通信。无人机基站可以为超密集网络的地面基站提供补充覆盖，也可以利用空地信道增益大的优势为毫米波通信提供高增益的无线连接。在应对自然灾害紧急救援情况时，无人机基站可以临时代替被损坏的地面基站提供应急通信。在基础设施覆盖不足的地区，无人机基站可以作为中继为地面用户提供无线通信服务。在热点

收稿日期：2020-06-29；改回日期：2021-06-07；网络出版：2021-07-13

*通信作者： 崔苗 cuimiao@gdut.edu.cn

基金项目：广东省科技计划(2017B090909006, 2019B010119001, 2020A050515010, 2021A0505030015)，广东特支计划(2019TQ05X409)

Foundation Items: The Science and Technology Plan Project of Guangdong Province (2017B090909006, 2019B010119001, 2020A050515010, 2021A0505030015), The Special Support Plan for High-Level Talents of Guangdong Province (2019TQ05X409)

区域或者需要处理临时事件期间，无人机基站能够为网络流量拥塞的地面基站提供有效的业务分流。第2类，无人机作为空中的用户，接入到地面的蜂窝通信网络。未来5G蜂窝网络能够为无人机提供更加可靠、更加安全和超低时延的连接从而实现其更多功能[2–6]。文献[3]研究了无人机通信网络中物理层安全性的问题；文献[4]研究了多无人机基站通信网络中的轨迹优化和功率分配问题；文献[5]研究了无人机中继系统中通过轨迹优化和功率分配来最大化吞吐量的问题。

本文主要研究的是上述第1类应用场景，即无人机作为空中基站为地面用户提供无线通信服务。相比于传统的地面基站，无人机基站的飞行高度比较高，能够与用户建立更加可靠的通信链路[4]，能够适应突发情况下的通信场景，例如救援、搜索、热点区域覆盖等。目前大多数的研究是通过优化无人机的位置部署、无人机的飞行轨迹或者资源分配来达到更佳的通信质量。例如文献[7]研究了最小化无人机基站的数量以及部署无人机的位置来覆盖给定数量的地面用户；文献[8]研究了无人机辅助的无线传感网中的数据采集；文献[9]研究了多跳无人机中继通信系统的轨迹优化以及功率分配。上述文献中采用的算法都属于离线优化算法，建立在对通信环境的完美假设的基础上，在无人机起飞之前规划好无人机的轨迹。然而在实际中，通信环境是不断变化的，无法提前预测，通信环境的完美假设无法实现[10]，因此无法解决地面用户随机的通信请求问题。与离线优化算法不同，在线优化算法不需要在无人机起飞前提前设计好无人机的飞行路线，而是能够在飞行过程中根据通信环境的变化动态、连续地规划无人机的飞行路线。文献[11]提出了一种近似动态规划的无人机机动决策方法；文献[12]提出一种基于动态规划的在线优化算法，而动态规划需要一个环境模型，且计算复杂度较高。

为了让无人机基站具有动态、实时规划飞行路线的能力，从而实现飞行路线能够实时适应地面用户随机的通信请求，并且考虑到无人机基站与地面用户进行时效性较强的决策信息通信时，减小通信时延尤为重要。本文将从平均通信时延最小化的角度出发，提出基于强化学习的飞行路线在线优化设计算法。在线优化算法能够在飞行过程中根据通信环境的变化动态地规划无人机的飞行路线，不需要完备的通信环境参数且计算复杂度比动态规划的低。强化学习方法包含蒙特卡罗和Q-Learning两类算法，用于解决马尔可夫决策问题[13]。实践证明，强化学习算法具有解决无人机无线网络在线优化问题的能力，例如文献[14]研究了多无人机辅助蜂窝网络中的用户体验质量(Quality of Experience, QoE)；文献[15]研究了多无人机辅助蜂窝网络中所有用户的通信速率和最大化的问题。

本文研究一个无人机空中基站为两个地面用户提供无线通信服务，其中地面用户的通信请求是随机的，以平均通信时延最小化为目的来设计无人机的在线飞行路线。首先，把飞行路线设计问题转化成一个马尔可夫决策过程，根据地面用户的通信请求情况和无人机的位置把无人机的状态分为通信状态和等待状态，然后定义不同状态下的动作集，将无人机完成单次任务的通信时延设置为回报，采用强化学习的蒙特卡罗算法以及Q-Learning算法实现在线优化无人机飞行路线。最后，通过计算机仿真验证本文提出的算法的有效性。

2 系统模型

如图1所示，本文考虑一个无人机基站通信系统，其中包括有一个无人机和两个地面用户[1)]。无人机作为一个空中通信基站，为两个地面用户提供无线通信服务。假定两个地面用户的通信请求是随机的，它们独立同分布，服从均值为$\lambda/2$的泊松过程($\lambda/2$次请求/秒)，每次通信请求的传输信息量为L bit，地面用户1位置坐标为$(-a,0,0)$和地面用户2的位置坐标$(a,0,0)$。假设无人机的飞行高度固定为H，最大飞行速度为$V_{\max}$，无人机在两个地面用户所连接的线段间移动。定义无人机单次完成传输任务的时间为T，称为单次通信时延，在时刻t无人机的位置坐标为$q(t)=(x(t),y(t),H),x(t)\in[-a,a],y(t)=0$。因为地面用户的位置都在$X$轴上，无人机为了获得更好的通信链路需要尽可能靠近地面用户，所以无人机的飞行路线也在X轴。

无人机收到地面用户$r\in\{1,2\}$的通信请求之后，进入通信状态，此时无人机为地面用户提供无线通信服务，其他地面用户的通信请求会被忽略。在完成数据传输之后，无人机进入等待状态，开始等待下一次通信请求，这个过程一直重复。

由于无人机的飞行高度比较高，与地面用户的链路与视距信道相似，所以本文假设无人机与地面用户之间的通信链路为视距信道[16]，无人机的发射功率固定为P，在t时刻，无人机与地面用户之间的瞬时通信速率为

1) 本文主要研究无人机基站的飞行路线在线优化，主要考察飞行路线对通信性能的影响，没有考虑无人机基站的能耗问题。另外，本文考虑的系统模型同样适用于多个无人机基站分别在不同频段上与地面用户通信的场景，并且后文提到的优化算法可以直接扩展到多个地面用户处在一条直线上的场景。

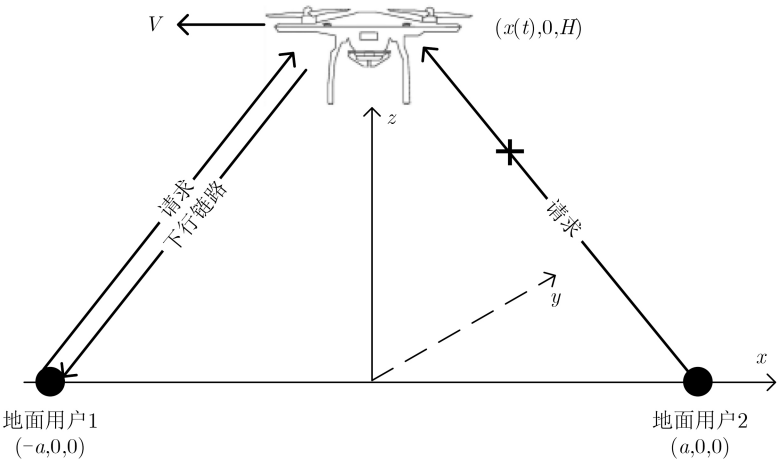

图 1 无人机基站通信系统

$$R_r^U\left(q\left(t\right)\right)=B\log_2\left(1+\frac{\gamma_0}{H^2+\left(x\left(t\right)-x_r\right)^2}\right) \tag{1}$$

其中，x_r为地面用户的X轴坐标位置，$r\in\{1,2\}$。$H^2+\left(x\left(t\right)-x_r\right)^2$为无人机与地面用户$r$的距离的平方，$B$为信道带宽，$\gamma_0$为参考距离为1 m时的信噪比。

3　飞行路线在线优化设计算法

3.1　问题描述

本文的目标是通过设计无人机的飞行路线达到最小化所有地面用户的平均通信时延，将平均通信时延定义为$D=\dfrac{\sum_{m=1}^{M}T_m}{M}$，其中$T_m$为无人机第$m$次完成任务的通信时延，$M$为完成的总通信次数。假设无人机接收到地面用户1的通信请求，立即为地面用户1提供服务，为了减小与地面用户1的通信时延，无人机会在信息传输过程中，尽可能接近发出通信请求的地面用户1以获得更佳的链路质量，从而获得最小通信时延。但从平均通信时延的角度来看，若下一次的服务对象为地面用户2，此时无人机距离地面用户2较远，与地面用户2的通信时延可能很大，从而导致平均通信时延变大。平均通信时延最小化的问题可以表述为(P1)

$$\min_{\{q(t)\}}\quad D$$

$$\text{s.t.}\quad\int_0^{T_m}R_r^U\left(q\left(t\right)\right)\mathrm{d}t\geq L,\ \forall m\in[0,M] \tag{2a}$$

$$0\leq\left\|q'\left(t\right)\right\|\leq V_{\max},\forall t \tag{2b}$$

$$-a\leq\left\|q\left(t\right)\right\|\leq a,\forall t \tag{2c}$$

其中，$q'\left(t\right)$表示无人机的速度V。式(2a)是保证在第m次通信任务中，T_m时间内能够传输L bit的信息；式(2b)表示无人机的飞行速度的约束，无人机的速度可取0或者$V_{\max}$；式(2c)是对无人机的位置约束。很明显(P1)是一个非凸的问题，同时为了让无人机基站具有动态、实时规划飞行路线的能力，从而适应不可预测、随机的地面用户通信请求，所以本文提出无人机飞行路线在线优化设计算法。所提算法基于强化学习的思想。

3.2　强化学习概述

强化学习可用于解决马尔可夫决策过程问题，属于机器学习的一类，使智能体在与环境的交互过程中获得的奖赏指导行为以达成回报最大化。马尔可夫决策过程的基本框架为$(\mathcal{S},\mathcal{A},\mathcal{R})$，在每个离散时刻$t$，观察到智能体在状态$S_t\in\mathcal{S}$，并且在此基础上选择一个动作$A_t\in\mathcal{A}\left(s\right)$。作为其动作的结果，智能体接收到一个数值化的即时奖励$R_{t+1}\in\mathcal{R}\subset\mathbb{R}$，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$\pi$。强化学习的目标不是最大化即时奖励，而是最大化长期回报$G_t=\sum_{k=0}^{\infty}\gamma^kR_{t+k+1}$，其中$\gamma$是一个参数，$0\leq\gamma\leq1$，称为折扣率。动作价值函数是强化学习算法中非常重要的一部分，其定义为策略π下在状态s时采取的动作a的价值记为$Q_\pi\left(s,a\right)=\mathbb{E}_\pi\left[G_t\left|S_t=s,A_t=a\right.\right]$，表示根据策略$\pi$，从状态$s$开始，执行动作$a$之后，所有可能的决策序列的期望回报。解决一个强化学习问题意味找到一个最优策略，使其能够在长期过程中获得最大回报。最优策略对应的最优动作价值函数$Q_*\left(s,a\right)=\max_\pi Q_\pi\left(s,a\right)$[13]。

将(P1)问题中无人机的轨迹离散化重新表述成一个马尔可夫决策过程。(P1)对应的马尔可夫决策过程如下：

状态：将两个地面用户连接的X轴区间$[-a,a]$离散成$2N$个小区间，位置索引序列为$\mathcal{I}=\{-N,-N+1,\cdots,N-1,N\}$，其对应的位置集合是$\mathcal{Q}=\left\{q_i=\dfrac{i}{N}a,\forall a\in\mathcal{I}\right\}$。地面用户请求可分为

3个状态，用$r\in\{0,1,2\}$来表示，0表示没有请求，1表示接收到来自地面用户1的请求，2表示接收到来自地面用户2的请求。本文将无人机的状态分为通信状态$\mathcal{S}_{\text{comm}}=(i,r),\ r=1,2\ i\in\mathcal{I}$和等待状态$\mathcal{S}_{\text{wait}}=(i,0),i\in\mathcal{I}$。

动作：将无人机的动作分为通信状态和等待状态的动作。等待状态的动作集合定义为$A_0=\{-1,0,1\}$，$A_0=-1$表示向左移动一个小区间，$A_0=0$表示悬停，$A_0=1$表示向右移动一个小区间。每个动作所花费的时间，即无人机在两个相邻的离散点所需时间为$\Delta=\dfrac{a}{NV}$。通信状态的动作集合定义为$A_r(q_i)=\cup_{q_j}A_r(q_i\to q_j),\ \forall i,j\in\mathcal{I}$，表示无人机通信状态的起始位置$q_i$到结束位置$q_j$的可飞行的轨迹集合，$A_r(q_i\to q_j)$指为地面用户$r$服务，起始位置为$q_i$，结束位置为$q_j$的可行的轨迹集合。

奖励：定义为$\mathcal{R}=-T_m^r(i,j),\ \forall(i,r)\in\mathcal{S}_{\text{comm}}$，表示无人机第$m$次完成通信任务的时间，服务对象为地面用户$r$，通信状态的起始位置$q_i$和结束位置$q_j$。

3.3 最小单次通信时延

根据平均通信时延$D=\dfrac{\sum_{m=1}^{M}T_m}{M}$的定义，需要先对第$m$次通信请求的通信时延$T_m$进行求解。假设无人机进入通信状态$(i,r)$的起始位置为$q_i,i\in\mathcal{I}$，则会存在$2N+1$个可能的结束位置$q_j,j\in\mathcal{I}$，给定通信状态的起始位置$q_i$和结束位置$q_j$的情况下，$A_r(q_i\to q_j)$总会存在一条单次通信时延最小的飞行路线$A_r^*(q_i\to q_j)$，因此对于任意一个通信状态$(i,r)$都存在$2N+1$条单次通信时延最小的飞行路线，从而可以将通信状态的动作集合放缩成$A_r^*(q_i)=\cup_{q_j}A_r^*(q_i\to q_j),\forall i,j\in\mathcal{I}$。将单次通信时延最小的飞行路线$A_r^*(q_i\to q_j)$的通信时延定义为$T_m^{r*}(i,j)=\min T_m^r(i,j),\forall(i,r)\in S_{\text{comm}},\forall j\in\mathcal{I}$，因此可以将平均通信时延写成$D=\dfrac{\sum_{m=1}^{M}T_m^{r*}(i,j)}{M}$，$\forall(i,r)\in S_{\text{comm}}$，可以看出$T_m^{r*}(i,j)$是由通信状态的结束位置$q_j$决定的，因此平均通信时延只与结束位置$q_j$有关。

为了求得$T_m^{r*}(i,j)$，首先定义$t_{p_1,p_2}=\dfrac{|p_1-p_2|}{V_{\max}}$，$p_1,p_2\in[-a,a]$，表示无人机以最大速度$V_{\max}$，从$p_1$飞到$p_2$的总时间。在$p_1$到$p_2$的飞行路线中，无人机发送给地面用户$r$的信息量为

$$l_{p_1,p_2}=\int_0^{t_{p_1,p_2}}R_r\left(p_1+\frac{t}{t_{p_1,p_2}}(p_1-p_2)\right)\mathrm{d}t \tag{3}$$

很明显$l_{p_1,p_2}=l_{p_2,p_1}$，$l_{p_1,p_2}=l_{-p_1,-p_2}$。

接下来证明单次通信时延最小的飞行路线的存在，假设无人机与地面用户1通信，对于任意一条轨迹$q(\cdot)\in A_1(q_i\to q_j)$，时延为$\Delta t$，可以找到另外一条轨迹$\widehat{q}(\cdot)\in A_1(q_i\to q_j)$，时延同为$\Delta t$，满足$|q(t)-x_1|\geq|\widehat{q}(t)-x_1|,\forall t\in[0,\Delta t]$，无人机在$\widehat{q}(\cdot)$轨迹下总是比在$q(\cdot)$轨迹下更靠近地面用户1。因此在相同通信时延的情况下，无人机在$\widehat{q}(\cdot)$轨迹下总是比在$q(\cdot)$轨迹下能够传输更大的信息量。换句话说，当无人机的通信状态的起始位置和结束位置相同时，$\widehat{q}(\cdot)$更靠近地面用户1，获得更好的通信链路，从而减小通信时延。同理，无人机与地面用户2通信的情况相同。

因此本文将无人机的通信状态的飞行路线分为3种情况：第1种情况，若$l_{q_i,q_j}\geq L$，$A_r^*(q_i\to q_j)$为无人机以最大速度$V_{\max}$从q_i飞向q_j，中间过程无悬停，此时的通信时延为$T_m^{r*}(i,j)=t_{q_i.q_j}$；第2种情况，若$l_{q_i,x_r}+l_{x_r,q_j}\leq L$，$A_r^*(q_i\to q_j)$为无人机以最大速度$V_{\max}$从$q_i$飞向$x_r$，在$x_r$悬停$\delta^*$时间，然后再飞向$q_j$，此时的通信时延为$T_m^{r*}(i,j)=t_{q_i,x_r}+t_{x_r,q_j}+\delta^*$，其中$\delta^*=\dfrac{L-l_{q_i,x_r}-l_{x_r,q_j}}{R_r(x_r)}$；第3种情况，若$l_{q_i,x_r}+l_{x_r,q_j}\geq L$且$l_{q_i,q_j}\leq L$，$A_r^*(q_i\to q_j)$为无人机以最大速度$V_{\max}$从$q_i$飞向$x_r$，到达$p^*$之后，中转返回飞向$q_j$，此情况的通信时延为$T_m^{r*}(i,j)=t_{q_i,p^*}+t_{p^*,q_j}$。当$r=1$时，$p^*$是$[x_r,\min\{q_i,q_j\}]$区间内的唯一解；当$r=2$时，$p^*$是$[\max\{q_i,q_j\},x_r]$区间内的唯一解。

3.4 基于蒙特卡罗的在线优化设计算法

蒙特卡罗算法通过平均样本的回报来解决强化学习问题。为了保证能够具有良好定义的回报，本文采用用于分幕式任务的蒙特卡罗算法。智能体与环境的交互分成一系列子序列，将子序列称为幕，每幕从某个标准起始状态开始，并且无论选取怎样的动作整个幕一定会终止，下一幕的开始状态与上一幕的结束方式完全无关，具有这种分幕特性重复任务称为分幕式任务。价值估计和策略改进在整个幕结束之后进行，因此蒙特卡罗算法是逐幕做出改进的。蒙特卡罗算法是从任意的策略π_0开始交替进行完整的策略评估和策略改进，最终得到最优的策略和动作价值函数。策略评估的目标是估计动作价值函数$Q_\pi(s,a)$，即在策略π下从状态s采取动作a的期望回报。策略改进是在当前的动作价值函数上贪婪地选择动作。由于已经存在动作价值函数，所以在贪婪的时候完全不需要使用任何的模型信息。对于任意的一个动作价值函数，对于任意的贪婪策略为：对于任意一个状态$s\in\mathcal{S}$，必定选择对应动作价值函数最大的动作，$\pi(s)=\arg\max_aQ(s,a)$[13]。

在每一幕结束后，使用观测到的回报进行策略评估，然后在该幕序列访问到的每一个状态上进行策略的改进。本文采用的是基于试探性出发的蒙特卡罗算法，具体算法如下：

步骤1　初始化最大训练幕数N_{epi}，每幕中最大步数N_{step}，对于所有$s\in\mathcal{S}_{\text{comm}}$，任意初始化；对于所有$s\in\mathcal{S}_{\text{comm}}$，$a\in A_r^*(q_i)$任意初始化动作值函数$Q(s,a)\in\mathbb{R}$；

步骤2　随机选择$s_0=(i,r)$，根据π生成一幕序列：$s_0,a_0,R_1,s_1.a_1,\cdots,s_{N_{\text{step}}-1}.a_{N_{\text{step}}-1},R_{N_{\text{step}}}$；

步骤3　$G=0$；计数变量W=0；$N_{\text{epi}}=N_{\text{epi}}-1$；

步骤4　对幕中的每一步循环，$n=N_{\text{step}}-1,N_{\text{step}}-2,\cdots,0$：

$$G=\gamma G+R_{n+1};$$

$$W=W+1;$$

$$Q(s_n,a_n)=Q(s_n,a_n)+\frac{1}{W}\left[G-Q(s_n,a_n)\right];$$

$$\pi(s_n)=\arg\max_a Q(s_n,a);$$

当二元组s_n,a_n在$s_0.a_0,s_1.a_1,\cdots,s_{N_{\text{step}}-1}.a_{N_{\text{step}}-1}$中出现过，退出幕中循环。

步骤5　重复步骤2—步骤4，直到N_{epi}=0。

3.5 基于Q-Learning的在线优化设计算法

与前面提到的蒙特卡罗算法一样，时序差分算法也可以直接从环境互动的经验中学习策略。时序差分算法与蒙特卡罗方法不同，不用等到交互的最终结果，而是在已得到的其他的状态估计值来更新当前状态的价值函数，其价值估计和策略改进是逐步的。与蒙特卡罗算法相比，时序差分算法的优势在于它运用了一种在线的、完全递增的方法来实现。蒙特卡罗算法必须等到一幕的结束才能知道确切的回报值，而时序差分算法只要等到下一时刻即可[13]。因为本文系统模型中的地面用户的通信请求是逐个发送的，然后由无人机逐个完成的，所以采用时序差分算法更加适合。本文采用的Q-Learning是一种典型的强化学习时序差分算法，动作价值函数定义为$Q(s_n,a_n)=Q(s_n,a_n)+\alpha\left[r_{n+1}+\gamma\max_a Q(s_{n+1},a)-Q(s_n,a_n)\right]$，其中$\alpha\in(0,1]$为步长，权衡上一次学习的结果和这一次学习的结果；$\gamma\in[0,1]$为折扣率，是考虑未来回报对现在影响的因子，$\gamma$的值越大，说明未来回报对现在影响越大；具体如下：

步骤1　初始化探索参数ε，最大训练幕数N_{epi}，每幕中最大步数N_{step}，动作价值函数$Q(s,a)=0,\forall s\in\mathcal{S}_{\text{comm}},\forall a\in A_r^*(q_i)$；

步骤2　随机给定初始状态$s_0=(i,r)$；

步骤3　$N_{\text{epi}}=N_{\text{epi}}-1$；

步骤4　对幕中的每一步循环，$n=1,2,\cdots,N_{\text{step}}$：

根据$\varepsilon-$greedy来选择通信状态的动作a_n，即通信状态的结束位置索引j；

采取动作a_n，得到回报$r_{n+1}=-T_m^{r*}(s_n,a_n)$，根据动作得到下一个状态$s_{n+1}$；

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5　重复步骤2—步骤4，直到N_{epi}=0。

4 仿真结果

本节利用计算机仿真对上述所提的飞行路线在线优化算法进行验证，并对比两种基准方案“固定位置”和“贪婪算法”。“固定位置”为无人机悬停在两个地面用户的连线的中点，“贪婪算法”的描述为在无人机每次进入通信状态(i,r)时，选择最小单次通信时延的飞行路线，即通信状态的结束位置q_j。

仿真中采用的系统参数为：地面用户1和地面用户2相距800 m，其中地面用户1的位置坐标为$(-400\text{ m},0,0)$，地面用户2的位置坐标$(400\text{ m},0,0)$；两个地面用户的通信请求到达率λ=1 次/秒，传输的信息量$L=2$ Mbit，无人机的飞行高度$H=100$ m，最大飞行速度$V_{\max}=20$ m/s，信道带宽$B=1$ MHz，参考距离1 m时的信噪比$\gamma_{\text{dB}}=40$ dB。将无人机的可飞行区域分为$2N+1=101$个位置状态，其他参数：$N_{\text{epi}}=120$, $N_{\text{step}}=10000$, ε=0.1, α=1, γ=0.2。

图2为无人机在等待状态时采取不同的动作策略对通信状态下的平均通信时延产生的影响对比。其中等待状态动作策略1：等待状态时无人机向地面用户1的方向移动(与地面用户2对称，情况相同)；等待状态动作策略2：等待状态时无人机向地面用户1和地面用户2的中点移动；等待状态动作策略3：等待状态时无人机悬停在原地不动。从图2可以看出等待状态采取动作策略2的情况下的平均通信时延最小，因此等待状态的最佳动作策略$\pi^*(\mathcal{S}_{\text{wait}})$为无人机向地面用户1和地面用户2的中点移动：

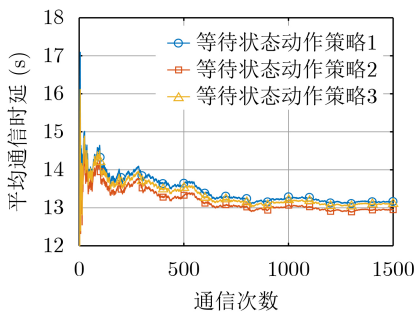

图 2 等待状态时采取不同动作策略的平均通信时延

$$\pi^*(\mathcal{S}_{\text{wait}})=\begin{cases} A_0=1, & i\in\{-N,-N+1,\cdots,-1\} \\ A_0=0, & i=0 \\ A_0=-1, & i\in\{1,2,\cdots,N\} \end{cases},$$
$$\mathcal{S}_{\text{wait}}=(i,0),i\in\mathcal{I} \tag{4}$$

在以下的仿真中，无人机的等待状态都采取等待状态动作策略2。

图3展示了不同算法下的无人机平均通信时延，可以看出相比于“固定位置”，其他3种算法下的平均通信时延都比较小，其中基于Q-Learning的在线优化设计算法的平均通信时延最小。这是因为“贪婪算法”只考虑当前的回报，没考虑长期回报；而蒙特卡罗算法把学习推迟到整幕结束之后，必须等到每一幕的结束，才能知道确切的回报值，而Q-Learning算法在每个动作结束就能够知道回报值从而进行学习，本文中的问题是一个持续性任务，不适合分幕式任务的蒙特卡罗算法，更加适合用Q-Learning算法来解决。

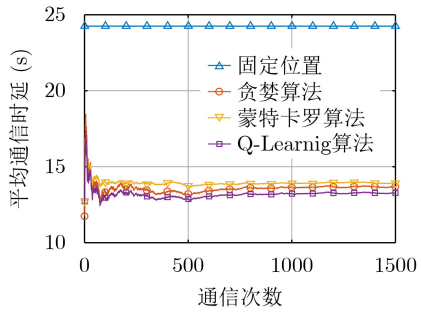

图 3 不同算法下的无人机平均通信时延

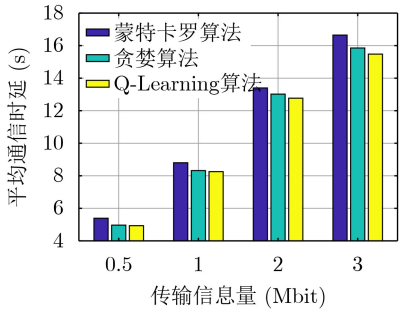

图 4 不同传输信息量以及不同算法下的无人机平均通信时延

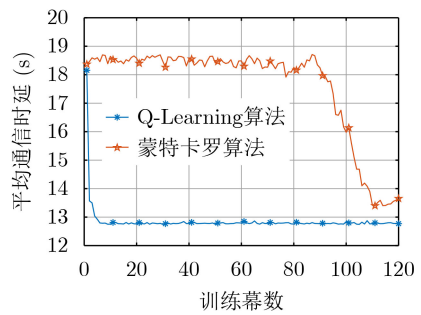

图 5 不同算法的收敛程度

图4展示在不同的传输信息量下，采用不同算法得到的平均通信时延。可以看出基于Q-Learning的在线优化设计算法的平均通信时延始终优于基于蒙特卡罗算法和“贪婪算法”的在线优化设计算法。传输信息量越大，基于Q-Learning的在线优化设计算法的时延性能越好。

图5展示了基于Q-Learning的在线优化设计算法和基于蒙特卡罗算法的在线优化设计算法下的平均通信时延随着训练幕数增大而逐渐收敛，可以看出基于Q-Learning的在线优化设计算法要比基于蒙特卡罗算法收敛得更快且稳定。

图6和图7分别展示了基于Q-Learning算法和基于蒙特卡罗算法的在线优化设计算法下无人机连续的10个状态下的飞行路线，其中无人机的起点是随机的，无人机的状态是根据地面用户的通信请求变化的，两个地面用户通信请求分别服从均值为1/2的泊松过程。图6和图7中的无人机状态：0表示无人机处于等待状态，1表示接收到地面用户1的通信请求，2表示接收到地面用户2的通信请求；时长

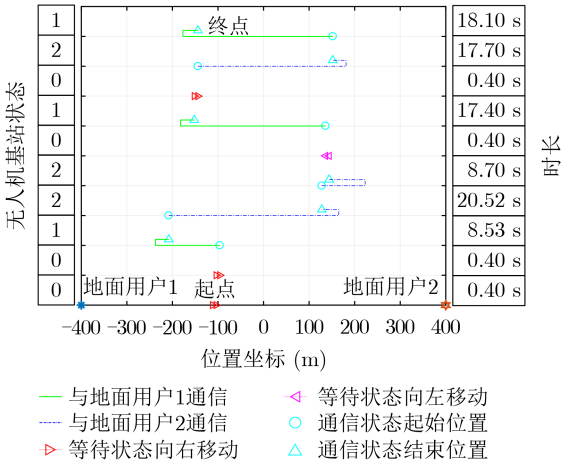

图 6 基于Q-Learning的在线优化设计算法下无人机飞行路线

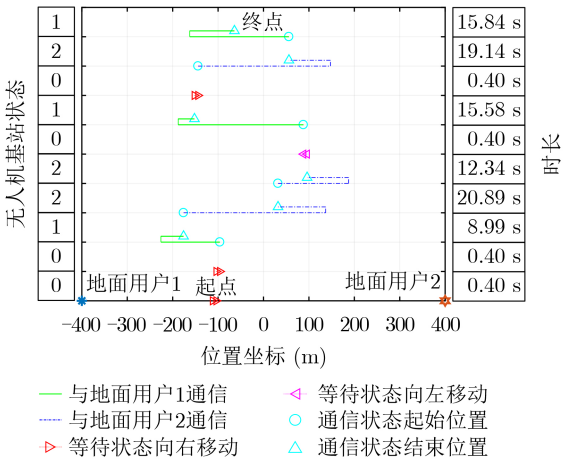

图 7 基于蒙特卡罗的在线优化设计算法下无人机飞行路线

表示无人机处于当前状态采取的动作所耗费的时间。可以对比看出基于Q-Learning的在线优化设计算法的无人机飞行路线更加集中于两个地面用户的中点，说明基于Q-Learning的在线优化设计算法更加能适应随机的地面用户请求，从而得到更小的平均通信时延。

5　结束语

本文针对无人机基站通信系统，提出了两种基于强化学习的无人机飞行路线在线优化设计算法。分别采用了强化学习的蒙特卡罗算法和Q-Learning算法来最小化无人机的平均通信时延。仿真结果显示了与无人机在固定位置相比，提出的算法具有较好的性能。基于Q-Learning的在线优化设计算法比基于蒙特卡罗的在线优化设计算法的训练结果更快收敛且稳定，能够更好地适应随机的地面用户请求从而达到更小的平均通信时延。

参 考 文 献

[1] ZENG Yong, ZHANG Rui, and LIM T J. Wireless communications with unmanned aerial vehicles: Opportunities and challenges[J]. *IEEE Communications Magazine*, 2016, 54(5): 36–42. doi: 10.1109/MCOM.2016.7470933.

[2] Paving the path to 5G: Optimizing commercial LTE networks for drone communication[EB/OL]. https://www.qualcomm.com/news/onq/2016/09/06/paving-path-5g-optimizing-commercial-lte-networks-drone-communication, 2016.

[3] ZHANG Guangchi, WU Qingqing, CUI Miao, *et al*. Securing UAV communications via joint trajectory and power control[J]. *IEEE Transactions on Wireless Communications*, 2019, 18(2): 1376–1389. doi: 10.1109/TWC.2019.2892461.

[4] WU Qingqing, ZENG Yong, and ZHANG Rui. Joint trajectory and communication design for multi-UAV enabled wireless networks[J]. *IEEE Transactions on Wireless Communications*, 2018, 17(3): 2109–2121. doi: 10.1109/TWC.2017.2789293.

[5] 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.

[6] ZENG Yong, LYU Jiangbin, and ZHANG Rui. Cellular-connected UAV: Potential, challenges, and promising technologies[J]. *IEEE Wireless Communications*, 2019, 26(1): 120–127. doi: 10.1109/MWC.2018.1800023.

[7] LYU Jiangbin, ZENG Yong, ZHANG Rui, *et al*. Placement optimization of UAV-mounted mobile base stations[J]. *IEEE Communications Letters*, 2017, 21(3): 604–607. doi: 10.1109/LCOMM.2016.2633248.

[8] ZHAN Cheng, ZENG Yong, and ZHANG Rui. Energy-efficient data collection in UAV enabled wireless sensor network[J]. *IEEE Wireless Communications Letters*, 2018, 7(3): 328–331. doi: 10.1109/LWC.2017.2776922.

[9] ZHANG Guangchi, YAN Haiqiang, ZENG Yong, *et al*. Trajectory optimization and power allocation for multi-hop UAV relaying communications[J]. *IEEE Access*, 2018, 6: 48566–48576. doi: 10.1109/ACCESS.2018.2868117.

[10] ZENG Yong and XU Xiaoli. Path design for cellular-connected UAV with reinforcement learning[EB/OL]. http://arxiv.org/abs/1905.03440, 2019.

[11] 黄长强, 赵克新, 韩邦杰, 等. 一种近似动态规划的无人机机动决策方法[J]. 电子与信息学报, 2018, 40(10): 2447–2452. doi: 10.11999/JEIT180068.

HUANG Changqiang, ZHAO Kexin, HAN Bangjie, *et al*. Maneuvering decision-making method of UAV based on approximate dynamic programming[J]. *Journal of Electronics & Information Technology*, 2018, 40(10): 2447–2452. doi: 10.11999/JEIT180068.

[12] BLISS M and MICHELUSI N. Trajectory optimization for rotary-wing UAVs in wireless networks with random requests[EB/OL]. http://arxiv.org/abs/1905.01755, 2019.

[13] SUTTON R S and BARTO A G. Reinforcement Learning: An Introduction[M]. 2nd ed. Cambridge: MIT Press, 2018: 1–130.

[14] 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.

[15] KHAMIDEHI B and SOUSA E S. Reinforcement learning-based trajectory design for the aerial base stations[EB/OL]. https://arxiv.org/abs/1906.09550, 2019.

[16] ZENG Yong and ZHANG Rui. Energy-efficient UAV communication with trajectory optimization[J]. *IEEE Transactions on Wireless Communications*, 2017, 16(6): 3747–3760. doi: 10.1109/TWC.2017.2688328.

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