Reinforcement Learning for Decentralized Trajectory Design in Cellular UAV Networks With Sense-and-Send Protocol 北京大学 cite 16

Jingzhi Hu, *Student Member, IEEE*, Hongliang Zhang, *Student Member, IEEE*, and Lingyang Song, *Senior Member, IEEE*

Hu J, Zhang H, Song L. Reinforcement learning for decentralized trajectory design in cellular UAV networks with sense-and-send protocol[J]. IEEE Internet of Things Journal, 2018, 6(4): 6177-6189.

***Abstract*—Recently, the unmanned aerial vehicles (UAVs) have been widely used in real-time sensing applications over cellular networks. The performance of a UAV is determined by both its sensing and transmission processes, which are influenced by the trajectory of the UAV. However, it is challenging for the UAV to determine its trajectory, since it works in a dynamic environment, where other UAVs determine their trajectories dynamically and compete for the limited spectrum resources in the same time. To tackle this challenge, we adopt the reinforcement learning to solve the UAV trajectory design problem in a decentralized manner. To coordinate multiple UAVs performing real-time sensing tasks, we first propose a sense-and-send protocol, and analyze the probability for successful valid data transmission using nested Markov chains. Then, we propose an enhanced multi-UAV *Q*-learning algorithm to solve the decentralized UAV trajectory design problem. Simulation results show that the proposed algorithm converges faster and achieves higher utilities for the UAVs, compared to traditional single- and multi-agent *Q*-learning algorithms.**

***Index Terms*—Reinforcement learning, sense-and-send protocol, trajectory design, unmanned aerial vehicle (UAV).**

I. Introduction

In THE cellular network, the use of unmanned aerial vehicles (UAVs) to perform sensing has been of particular interests, due to their high mobility, flexible deployment, and low operational cost [1]. Specially, UAVs have been widely applied to execute critical sensing tasks, such as traffic monitoring [2], precision agriculture [3], and forest fire surveillance [4]. In these UAV sensing applications, the sensory data collected by the UAVs need to be transmitted to the base station (BS) immediately for further real-time data processing. This poses a significant challenge for the UAVs to sense the task and send the collected sensory data simultaneously with a satisfactory performance.

Manuscript received July 15, 2018; revised September 8, 2018; accepted October 8, 2018. Date of publication October 17, 2018; date of current version July 31, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61625101. *(Corresponding author: Lingyang Song.)*

The authors are with the School of Electrical Engineering and Computer Science, Peking University, Beijing 100871, China (e-mail: jingzhi.hu@pku.edu.cn; hongliang.zhang@pku.edu.cn; lingyang.song@pku.edu.cn).

Digital Object Identifier 10.1109/JIOT.2018.2876513

In order to enable the real-time sensing applications, the cellular network controlled UAV transmission is considered as one promising solution [5], [6], in which the uplink QoS is guaranteed compared to that in ad-hoc sensing networks [7]. However, it remains a challenge for UAVs to determine their trajectories in such cellular UAV networks. When a UAV is far from the task, it risks in obtaining invalid sensing data, while if it is far from the BS, the low uplink transmission quality may lead to difficulties in transmitting the sensory data back to the BS. Therefore, the UAVs need to take both the sensing accuracy and the uplink transmission quality into consideration in designing their trajectories. Moreover, it is even more challenging when the UAVs belong to different entities and are noncooperative. Since the spectrum resource is scarce, the UAVs have the incentive to compete for the limited uplink channel resources. In this regard, each UAV should also consider other UAVs that are competing for the spectrum dynamically when it determines the trajectory. Therefore, a decentralized trajectory design approach is necessary for the UAV trajectory design problem, in which the locations of both the task and the BS, as well as the behaviors of the other UAVs should be taken into consideration by each UAV.

To tackle these problems, in this paper, we consider the scenario where multiple UAVs in a cellular network perform different real-time sensing tasks. We first propose a sense-and-send protocol to coordinate the UAVs, which can be analyzed and by nested Markov chains Under this condition, the UAV trajectory design problem can be seen as a Markov decision problem, which makes the reinforcement learning the suitable and promising approach to solve the problem. To be specific, we formulate the UAV trajectory design problem under the reinforcement learning framework, and propose an enhanced multi-UAV *Q*-learning algorithm to solve the problem efficiently.

In literature, most works focused on either the sensing or the transmission in the UAV networks, instead of considering UAV sensing and transmission jointly. For example, Tisdale *et al.* [8], Maza *et al.* [9], Gu *et al.* [10], and Yang *et al.* [11] focused on the sensing part. In [8], the autonomous path planning problem was discussed for a team of UAVs, which were equipped with vision-based sensing system to search for a stationary target. In [9], an architecture was proposed to deal with the cooperation and the control of multiple UAVs with sensing and actuation capabilities for the

deployment of loads. In [10], the optimal cooperative estimation problem for a team of UAVs to estimate both the position and the velocity of a ground moving target was considered. In [11], a mobile air quality monitoring system boarded on the UAV was designed to sense the real-time air quality and generate the estimated air quality index maps.

Besides, Zhang *et al.* [12] and Bor-Yaliniz *et al.* [13] focused on the transmission part. In [12], the joint trajectory and power optimization problem was formulated to minimize the outage probability in the network, in which the UAV relayed the transmissions of mobile devices. In [13], UAVs were used as aerial BSs which assisted the BS in providing connectivity within the cellular network, and an optimization problem was formulated to maximize the revenue. In [14], both the sensing and the transmission are taken into consideration, and an iterative trajectory, sensing, and scheduling algorithm was proposed to schedule UAVs' trajectories in a centralized manner, in which the task completion time was minimized. Nevertheless, the decentralized trajectory design problem remains to be lack of discussion. This is important since in practical scenarios the UAVs may belong to different entities, and thus they have the incentive to maximize their own utilities.

The main contributions of this paper can be summarized as follows.

1) We propose a sense-and-send protocol to coordinate UAVs performing real-time sensing tasks, and solve the probability of successful valid sensory data transmission by using nested Markov chains.
2) We formulate the decentralized UAV trajectory design problem, and propose an enhanced multi-UAV Q-learning algorithm to solve the problem.
3) Simulation results show that the enhanced multi-UAV Q-learning algorithm converges faster and to higher rewards of UAVs compared to both single-agent and opponent modeling Q-learning algorithms.

The rest of this paper is organized as follows. In Section II, the system model is described. In Section III, we propose the sense-and-send protocol to coordinate the UAVs performing real-time sensing tasks. We analyze the performance of the proposed sense-and-send protocol in Section IV, and derive the probability of successful valid sensory data transmission using the nested Markov chains. Following that, the reinforcement learning framework and the enhanced multi-UAV Q-learning algorithm are given in Section V, together with the analyzes of complexity, convergence, and scalability. The simulation results are presented in Section VI. Finally, the conclusions are drawn in Section VII.

II. System Model

As illustrated in Fig. 1, we consider a single cell orthogonal frequency-division multiple access network which consists of N UAVs to perform real-time sensing tasks. We set the horizontal location of the BS as the origin, and the location of the BS and the UAVs can be specified by 3-D cartesian coordinates, i.e., the location of the ith UAV can be denoted as $\boldsymbol{s}_i = (x_i, y_i, h_i)$, and the location of the BS can be denoted

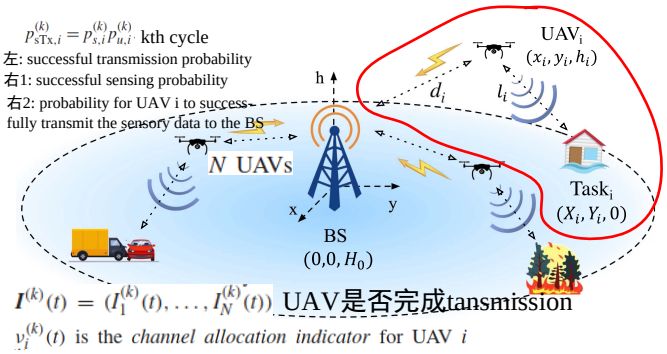

Fig. 1. Illustration on the single-cell UAV network, in which UAVs perform real-time sensing tasks.

每个UAV对应一个任务，任务不动，UAV飞到离任务近的地方先去sense（但是UAV不知道是否接收成功），sense成功后飞向BS传输信息，相当于relay

as $\boldsymbol{S}_0 = (0, 0, H_0)$ with H_0 being its height. Besides, the location of the UAV i's real-time sensing task is denoted as $\boldsymbol{S}_i = (X_i, Y_i, 0)$. To perform the real-time sensing task, each UAV continuously senses the condition of its task, and sends the collected sensory data back to the BS immediately. In this regard, the sensing process and the transmission process jointly determine the UAVs' performance on their real-time sensing tasks. The sensing and transmission models of the UAVs are described in the following.

A. UAV Sensing

To evaluate the sensing quality of the UAV, we utilize the probabilistic sensing model as introduced in [15] and [16], where the successful sensing probability is an exponential function of the distance between the UAV and its task. Supposing that UAV i senses task i for a second, the probability for the UAV to sense the condition of its task successfully, i.e., the successful sensing probability, can be expressed as

$$\mathrm{Pr}_{s,i} = e^{-\lambda l_i} \tag{1}$$

in which λ is the parameter evaluating the sensing performance, and l_i denotes the distance from UAV i to task i.

It is worth noticing that UAV i cannot figure out whether the sensing is successful or not from its collected sensory data, due to its limited on-board data processing ability. Therefore, UAV i needs to send the sensory data to the BS, and the BS will decide whether the sensory data is valid or not. Nevertheless, UAV i can evaluate its sensing performance by calculating the successful sensing probability based on (1).

B. UAV Transmission

In the UAV transmission, the UAVs transmit the sensory data to the BS over orthogonal subchannels (SCs) to avoid mutual interference. To be specific, we adopt the 3GPP channel model to evaluate the urban macro cellular support for UAVs [17], [18].

Denoting the transmit power of UAVs as P_u, we can express the received signal-to-noise ratio (SNR) at the BS of UAV i as

$$\gamma_i = \frac{P_u \|H_i\|}{N_0 10^{\mathrm{PL}_{a,i}/10}} \tag{2}$$

in which $\mathrm{PL}_{a,i}$ denotes the air-to-ground pathloss, N_0 denotes the power of noise at the receiver of the BS, and H_i is the

small-scale fading coefficient. Specifically, the pathloss $\mathrm{PL}_{a,i}$ and small-scale fading H_i are calculated in two cases separately, i.e., the case where line-of-sight component exists (LoS case), and the case where none LoS components exist (NLoS case). The probability for the UAV i-BS channel to contain an LoS component can be calculated as

$$\Pr_{\mathrm{LoS},i} = \begin{cases} 1, & r_i \le r_c \\ \frac{r_c}{r_i} + e^{-r_i/p_0 + r_c/p_0}, & r_i > r_c \end{cases} \tag{3}$$

in which $r_i = \sqrt{x_i^2 + y_i^2}$, $p_0 = 233.98\log_{10}(h_i) - 0.95$, and $r_c = \max\{294.05\log_{10}(h_i) - 432.94, 18\}$.

When the channel contains an LoS component, the pathloss from UAV i to the BS can be calculated as $\mathrm{PL}_{\mathrm{LoS},i} = 30.9 + (22.25 - 0.5\log_{10}(h_i))\log_{10}(d_i) + 20\log_{10}(f_c)$, where f_c is the carrier frequency and d_i is the distance between the BS and UAV i. In the LoS case, the small-scale fading H_i obeys Rice distribution with scale parameter $\Omega = 1$ and shape parameter $K[\mathrm{dB}] = 4.217\log_{10}(h_i) + 5.787$. On the other hand, when the channel contains none LoS components, the pathloss from UAV i to the BS can be calculated as $\mathrm{PL}_{\mathrm{NLoS},} = 32.4 + (43.2 - 7.6\log_{10}(h_i)) \times \log_{10}(d_i) + 20\log_{10}(f_c)$, and the small-scale fading H_i obeys Rayleigh distribution with zero means and unit variance.

To achieve a successful transmission, the SNR at the BS needs to be higher than the decoding threshold γ_{th}. Therefore, each UAV can evaluate its successful transmission probability by calculating the probability for the SNR at BS to be larger than γ_{th}. The successful transmission probability $\Pr_{\mathrm{Tx},i}$ for UAV i can be calculated as

$$\begin{aligned} \Pr_{\mathrm{Tx},i} = {} & \Pr_{\mathrm{Los},i}\big(1 - F_{ri}\big(\chi_{\mathrm{LoS},i}\big)\big) \\ & + \big(1 - \Pr_{\mathrm{LoS},i}\big)\big(1 - F_{ra}\big(\chi_{\mathrm{NLoS},i}\big)\big) \end{aligned} \tag{4}$$

in which $\chi_{\mathrm{NLoS},i} = N_0 10^{0.1\mathrm{PL}_{\mathrm{NLoS},i}}\gamma_{\mathrm{th}}/P_u$, $\chi_{\mathrm{LoS},i} = N_0 10^{0.1\mathrm{PL}_{\mathrm{LoS},i}}\gamma_{\mathrm{th}}/P_u$, $F_{ri}(x) = 1 - Q_1(\sqrt{2K}, x\sqrt{2(K+1)})$ is the cumulative distribution function (CDF) of the Rice distribution with $\Omega = 1$ [19], and $F_{ra}(x) = 1 - e^{-x^2/2}$ is the CDF of the Rayleigh distribution with unit variance. Here, $Q_1(x)$ denotes the Marcum Q-function of order 1 [20].

III. Sense-and-Send Protocol

In this section, we propose a sense-and-send protocol to coordinate multiple UAVs to perform the sensing tasks simultaneously. We first introduce the sense-and-send cycle, which consists of the beaconing phase, the sensing phase, and the transmission phase. After that, we describe the uplink SC allocation mechanism of the BS.

A. Sense-and-Send Cycle

In this paper, we propose that the UAVs perform the sensing tasks in a synchronized iterative manner. Specifically, the process is divided into cycles, which are indexed by k. In each cycle, each UAV senses its task and then sends the collected sensory data to the BS. In order to synchronize the transmissions of the UAVs, we further divide each cycle into T_c *frames*, which are the basic time unit for the SC allocation. The duration of the time unit frame is set to be the duration

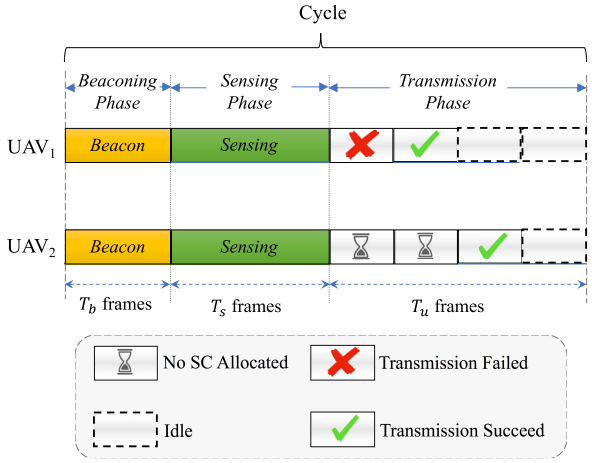

Fig. 2. Illustration on the sense-and-send protocol.

of the transmission and acknowledgment of the sensory data frame.[1]

The cycle consists of three separated phases, i.e., the *beaconing phase, sensing phase*, and the *transmission phase*, which contain T_b, T_s, and T_u frames, respectively. The duration of the beaconing phase and sensing phase are considered to be fixed and determined by the time necessary for transmitting beacons and collecting sensory data. On the other hand, the duration of the transmission phase is decided by the BS, which is related to the network conditions, such as the number of UAVs in the network. Moreover, as illustrated in Fig. 2, we consider that the sensing and transmission phases are separated to avoid the possible interference.[2]

In the beaconing phase, each UAV sends its location to the BS in its beacon through the control channel, which can be obtained by the UAV from the on-board GPS positioning. Collecting the beacons sent by the UAVs, the BS then broadcasts to inform the UAVs of the general network settings as well as the locations of all the UAVs. By this means, UAVs can obtain the locations of other UAVs in the beginning of each cycle. Based on the acquired information, each UAV then decides its trajectory in the cycle and informs the BS by another beacon.

In the sensing phase, each UAV senses the task for T_s frames continuously, during which it collects the sensory data. In each frame of the transmission phase, the UAVs attempt to transmit the collected sensory data to the BS if SCs are allocated to them by the BS. Specifically, there are four possible situations for each UAV, which are shown in Fig. 2 and can be described as follows.

1) *No SC Allocated:* In this case, no uplink SC is allocated to UAV i by the BS. Therefore, the UAV cannot transmit its collected sensory data to the BS. It will wait for the BS to assign an SC to it in order to transmit the sensory data.

[1]In this paper, we assume that the collected sensory data of each UAV in a cycle can be converted into a single sensory data frame with the same length.

[2]For example, the UAV's transmission will interfere with its sensing if the UAV tries to sense electromagnetic signals in the frequency bands which are adjacent to its transmission frequency.

2) *Transmission Failed:* In this case, an uplink SC is allocated to UAV i by the BS. However, the transmission is unsuccessful due to the low SNR at the BS, and thus, UAV i attempts to send the sensory data again to the BS in the next frame.
3) *Transmission Succeed:* In this case, an uplink SC is allocated to UAV i, and the UAV succeeds in sending its collected sensory data to the BS.
4) *Idle:* In this case, UAV i has successfully sent its sensory data in the former frames, and will keep idle in the rest of the cycle until the beginning of the next cycle.

Although in this paper we have assumed that the transmission of sensory data occupies a single frame, it can be extended to the case where the sensory data transmission takes n frames. In that case, the channel scheduling unit becomes n frames instead of a single frame.

B. Uplink Subchannel Allocation Mechanism

Since the uplink SC resources are usually scarce, in each frame of the transmission phase, there may exist more UAVs requesting to transmit the sensory data than the number of available uplink SCs. To deal with this problem, the BS adopts the following SC allocation mechanism to allocate the uplink SCs to the UAVs.

In each frame, the BS allocates the C available uplink SCs to the UAVs with uplink requirements, in order to maximize the sum of successful transmission probabilities of the UAVs. Based on the matching algorithm in [21], it is equivalent to that the BS allocates the C available SCs to the first C UAVs with the highest successful transmission probabilities. The successful transmission probabilities of UAVs can be calculated by the BS based on (4), using the information on the trajectories of the UAVs collected in the beaconing phase. Moreover, we denote the transmission states of the UAVs in the kth cycle as the vector $\boldsymbol{I}^{(k)}(t)$, in which $\boldsymbol{I}^{(k)}(t) = (I_1^{(k)}(t), \ldots, I_N^{(k)}(t))$. Here, $I_i^{(k)}(t) = 0$ if UAV i has not succeeded in transmitting its sensory data to the BS at the beginning of the tth frame, otherwise, $I_i^{(k)}(t) = 1$. Based on the above notations, the uplink SC allocation can be expressed by the channel allocation vector $\boldsymbol{v}^{(k)}(t) = (v_1^{(k)}(t), \ldots, v_N^{(k)}(t))$, in which the elements are determined by

$$v_i^{(k)}(t) = \begin{cases} 1, & \Pr_{\mathrm{Tx},i}^{(k)}(t) I_i^{(k)}(t) \geq \left(\boldsymbol{Pr}_{\mathrm{Tx}}^{k}(t) \boldsymbol{I}^{(k)}(t)\right)_C . \\ 0, & \text{o.w.} \end{cases} \tag{5}$$

Here, $v_i^{(k)}(t)$ is the *channel allocation indicator* for UAV i, i.e., $v_i^{(k)}(t) = 1$ only if an uplink SC is allocated to UAV i in the tth frame, $\Pr_{\mathrm{Tx},i}^{(k)}(t)$ denotes the successful transmission probability of UAV i in the tth frame of the kth cycle, and $(\boldsymbol{Pr}_{\mathrm{Tx}}^{(k)}(t)\boldsymbol{I}^{(k)}(t))_C$ denotes the Cth largest successful transmission probabilities among the UAVs who have not succeeded in sending the sensory data before the tth frame.

Since the trajectory of UAV i determines UAV i's distance to the BS, it influences the successful transmission probability. As the UAVs are allocated SCs only when they have the C-largest transmission probabilities, the UAVs have the incentive to compete with each other by selecting trajectories where their successful transmission probabilities are among the highest C ones. Consequently, the UAVs need to design their trajectories with the consideration of not only their distance to the BS and the task, but also of the trajectories of other UAVs.

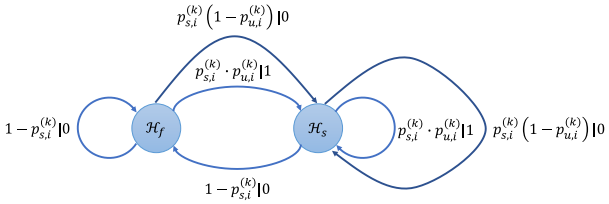

Fig. 3. Illustration on outer Markov chain of UAV sensing.

IV. Sense-and-Send Protocol Analysis

In this section, we analyze the performance of the proposed sense-and-send protocol by calculating the probability of successful valid sensory data transmission, which plays an important role in solving the UAV trajectory design problem. We first specify the state transitions of the UAVs by using nested bi-level Markov chains. The outer Markov chain depicts the state transitions in the UAV sensing process, and the inner Markov chain depicts the state transitions in the UAV transmission process, which will be elaborated on in the following parts, respectively.

A. Outer Markov Chain of UAV Sensing

In the outer Markov chain, the state transition takes place among different cycles. As shown in Fig. 3, for each UAV, it has two states in each cycle, i.e., state $\mathcal{H}_f$ to denote that the sensing is failed, and state $\mathcal{H}_s$ to denote that the sensing is successful. Supposing the successful sensing probability of UAV i in the kth cycle is $p_{s,i}^{(k)}$, UAV i transits to the $\mathcal{H}_s$ state with probability $p_{s,i}^{(k)}$ and transits to the $\mathcal{H}_f$ state with probability $(1-p_{s,i}^{(k)})$ after the kth cycle. The value at the right side of the transition probability denotes the number of valid sensory data that have been transmitted successfully to the BS in the cycle.

Besides, we denote the probability for UAV i to successfully transmit the sensory data to the BS as $p_{u,i}^{(k)}$. Therefore, UAV i successfully transmits valid sensory data to the BS with the probability $p_{s,i}^{(k)} p_{u,i}^{(k)}$. On the other hand, with probability $p_{s,i}^{(k)}(1-p_{u,i}^{(k)})$, no valid sensory data is transmitted to the BS though the sensing is successful in the kth cycle. The probability $p_{u,i}^{(k)}$ can be analyzed by the inner Markov chain of the UAV transmission in the next section, and $p_{s,i}^{(k)}$ can be calculated as follows.

Since the change of the UAVs' locations during each frame is small, we assume that the location of each UAV is fixed within each frame. Therefore, the location of UAV i in the kth cycle can be expressed as a function of the frame index t, i.e., $\boldsymbol{s}_i^{(k)}(t) = (x_i^{(k)}(t), y_i^{(k)}(t), h_i^{(k)}(t))$, $t \in [1, T_c]$. Similarly, the distance between UAV i and its task can be expressed as $l_i^{(k)}(t)$, and the distance between the UAV and the BS can be expressed as $d_i^{(k)}(t)$. Moreover, we assume that the UAVs move with uniform speed and direction in each cycle after the

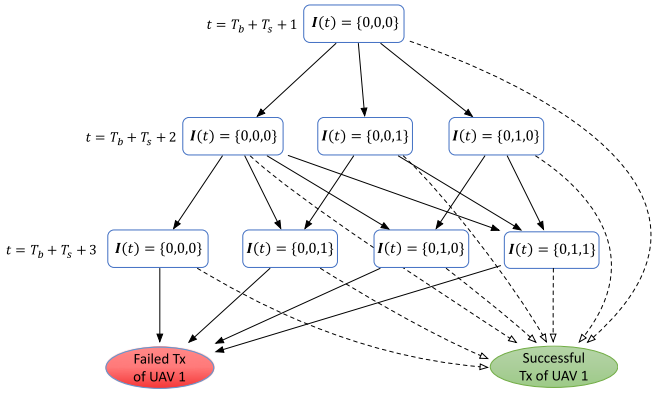

Fig. 4. Illustration on inner Markov chain of UAV 1's transmission given $C = 1$, $N = 3$, and $T_u = 3$.

beginning of the sensing phase. Based on the above assumptions, at the tth frame of the kth cycle, the location of UAV i can be expressed as

$$\boldsymbol{s}_i^{(k)}(t) = \boldsymbol{s}_i^{(k)}(T_b) + \frac{t}{T_c}\left(\boldsymbol{s}_i^{(k+1)}(1) - \boldsymbol{s}_i^{(k)}(T_b)\right), \quad t \in [T_b, T_c]. \tag{6}$$

Therefore, the successful sensing probability of UAV i in the cycle can be calculated as

$$p_{s,i}^{(k)} = \prod_{t=T_b+1}^{T_s+T_b} (\mathrm{Pr}_{s,i}^{(k)}(t))^{t_f} = \prod_{t=T_b+1}^{T_s+T_b} e^{-\lambda t_f l_i^{(k)}(t)} \tag{7}$$

in which t_f denotes the duration of a frame, and $l_i^{(k)}(t) = \|\boldsymbol{s}_i^{(k)}(t) - \boldsymbol{S}_i\|$.

B. Inner Markov Chain of UAV Transmission

For simplicity of description, we consider a general cycle and omit the superscript cycle index k. Since the general state transition diagram is rather complicated, we illustrate the inner Markov chain by giving an example where the number of available uplink SC $C = 1$, the number of UAVs $N = 3$, and the number of uplink transmission frames $T_u = 3$.

Taking UAV 1 as an example, the state transition diagram is illustrated in Fig. 4. The state of the UAVs in frame t can be represented by the transmission state vector $\boldsymbol{I}(t)$ as defined in Section III-B. Initially $t = T_b + T_s + 1$, the transmission state is $\boldsymbol{I}(T_b + T_s + 1) = \{0, 0, 0\}$, which indicates that UAVs 1, 2, and 3 have not succeeded in uplink transmission at the beginning of the transmission phase, and all of them are competing for the uplink SCs. In the next frame, the transmission state will transit to the *Successful Tx* state for UAV 1, if the sensory data of UAV 1 has been successfully transmitted to the BS. The probability for this transition equals to $\mathrm{Pr}_{Tx,1}(T_b + T_s + 1)\nu_1(T_b + T_s + 1)$, i.e., the probability for successful uplink transmission if an SC is allocated to UAV 1, otherwise, it equals to zero.

However, if UAV 1 does not succeed in uplink transmission, the transmission transits into other states, which is determined by whether other UAVs have succeeded in uplink transmission, e.g., it transits to $\boldsymbol{I}(T_b + T_s + 2) = (0, 0, 1)$ if UAV 3 succeeds in the first transmission frame. Note that when other UAVs succeed in transmitting sensory data in the previous frames, UAV 1 will face less competitors in the following frames, and thus, it have a larger probability to transmit successfully. Finally, when $t = T_c$, i.e., the last transmission frame in the cycle, UAV 1 will enter the *Failed Tx* state if it does not transmit the sensory data successfully, which means that the sensory data in this cycle is failed to be uploaded. Therefore, to obtain the $p_{u,i}$ in the outer Markov chain is equivalent to calculate the absorbing probability of successful Tx state in the inner Markov chain.

From the above example, it can be observed that the following general recursive equation holds for UAV i when $t \in [T_b + T_s + 1, T_c]$:

$$\begin{aligned}\mathrm{Pr}_{u,i}\{t|\boldsymbol{I}(t)\} = {} & \mathrm{Pr}_{\mathrm{Tx},i}(t)\nu_i(t) \\ & + \sum_{\substack{\boldsymbol{I}(t+1),\\ I_i(t+1)=0}} \Pr\{\boldsymbol{I}(t+1)|\boldsymbol{I}(t)\}\mathrm{Pr}_{u,i}\{t+1|\boldsymbol{I}(t+1)\}\end{aligned} \tag{8}$$

in which $\Pr\{\boldsymbol{I}(t+1)|\boldsymbol{I}(t)\}$ denotes the probability for the transmission state vector of the $(t+1)$th frame to be $\boldsymbol{I}(t+1)$, on the condition that the transmission state vector of the tth frame is $\boldsymbol{I}(t)$, and $\mathrm{Pr}_{u,i}\{t|\boldsymbol{I}(t)\}$ denotes the probability for UAV i to transmit sensory data successfully after the tth frame in the current cycle, given the transmission state $\boldsymbol{I}(t)$.

Since the successful uplink transmission probabilities of the UAVs are independent, we have $\Pr\{\boldsymbol{I}(t+1)|\boldsymbol{I}(t)\} = \prod_{i=1}^{N}\Pr\{I_i(t+1)|I_i(t)\}$, in which $\Pr\{I_i(t+1)|I_i(t)\}$ can be calculated as

$$\begin{cases} \Pr\{I_i(t+1) = 0|I_i(t) = 0\} = 1 - \mathrm{Pr}_{\mathrm{Tx},i}(t) \\ \Pr\{I_i(t+1) = 1|I_i(t) = 0\} = \mathrm{Pr}_{\mathrm{Tx},i}(t) \\ \Pr\{I_t(t+1) = 0|I_i(t) = 1\} = 0 \\ \Pr\{I_t(t+1) = 1|I_i(t) = 1\} = 1. \end{cases} \tag{9}$$

Here, the first two equations hold due to that the successful transmission probability of UAV i in the tth frame is $\mathrm{Pr}_{\mathrm{Tx},i}(t)$. The third and forth equations indicate that the UAVs keep idle in the rest of frames once they have successfully sent their sensory data to the BS.

Based on (8), the recursive algorithm can be proposed to solve $\mathrm{Pr}_{u,i}\{t|\boldsymbol{I}(t)\}$, as presented in Algorithm 1. Therefore, the successful transmission probability can be obtained by $p_{u,i} = \mathrm{Pr}_{u,i}\{T_b + T_s + 1|\boldsymbol{I}(T_b + T_s + 1)\}$.

In summary, the probability of successful valid sensory data transmission for UAV i in the kth cycle can be calculated as

$$p_{\mathrm{sTx},i}^{(k)} = p_{s,i}^{(k)} p_{u,i}^{(k)}. \tag{10}$$

C. Analysis on the Spectral Efficiency

In this paper, we evaluate the spectral efficiency by the average number of valid sensory data transmissions per second, which is denoted as N_{vd}. The value of N_{vd} is influenced by many factors, such as the distance between the BS and the tasks, the number of available SCs, the number of UAVs in the network, and the duration of the transmission phase.

In this paper, we analyze the influence of the duration of transmission phase T_u on N_{vd} in a simplified case: assuming

Algorithm 1 Algorithm for Successful Transmission Probability in a Cycle

Input: Frame index (t); Transmission state vector ($\boldsymbol{I}(t)$); Length of beaconing phase (T_b); Length of sensing phase (T_s); Length of transmission phase (T_u); Location of UAVs ($\boldsymbol{s}(t)$); Number of SCs (C).

Output: $\Pr_{u,i}\{t|\boldsymbol{I}(t)\},\ i=1,\ldots,N$;

if $t = T_b + T_s + 1$ **then**
  $\Pr_{u,i}\{t|\boldsymbol{I}(t)\} \coloneqq 0, i = 1,\ldots,N$;
else if $t > T_c$ **then**
  return $\Pr_{u,i}\{t|\boldsymbol{I}(t)\} = 0,\ i = 1,\ldots,N.$
end if
Calculate the successful transmission probabilities $\Pr_{\mathrm{Tx},i}(t),\ i = 1,\ldots,N$ based on (4).
Determine the SC allocation indicator $\boldsymbol{v}(t)$ based on (5).
for $i \in [1, N]$ and $I_i(t) = 0$ **do**
  $\Pr_{u,i}\{t|\boldsymbol{I}(t)\} \coloneqq \Pr_{Tx,i}(t)v_i(t)$.
end for
for all $\boldsymbol{I}(t+1)$ with $\Pr\{\boldsymbol{I}(t+1)|\boldsymbol{I}(t)\} > 0$ **do**
  Solve $\Pr_{u,i}\{t+1|\boldsymbol{I}(t+1)\}$ by calling Alg. 1, in which $t \coloneqq t+1$ and $\boldsymbol{I}(t) \coloneqq \boldsymbol{I}(t+1)$ and other parameters hold.
  $\Pr_{u,i}\{t|\boldsymbol{I}(t)\} \coloneqq \Pr_{u,i}\{t|\boldsymbol{I}(t)\} + \Pr\{\boldsymbol{I}(t+1)|\boldsymbol{I}(t)\}\Pr_{u,i}\{t+1|\boldsymbol{I}(t+1)\}$.
end for
return $\Pr_{u,i}\{t|\boldsymbol{I}(t)\},\ i = 1,\ldots,N.$

all the UAVs are equivalent, i.e., they have the same probabilities for successful uplink transmission in a frame, the same probabilities for successful sensing, and the same probabilities to be allocated SCs. Based on the above assumptions, the following proposition can be derived.

Proposition 1 (Optimal Duration of Transmission Phase): When the N UAVs are equivalent, and have the probability for successful sensing p_s, the probability for successful uplink transmission p_u, then N_{vd} first increases then decreases with the increment of T_u, and the optimal T_u^* can be calculated as

$$T_u^* = \frac{N}{C\ln(1-p_u)}\left(1 + W_{-1}\left(-\frac{(1-p_u)^{\frac{CT_u}{N}}}{e}\right)\right) - T_b - T_s \tag{11}$$

in which $W_{-1}(\cdot)$ denotes the lower branch of Lambert-W function in [22].

Proof: See the Appendix. ∎

The above proposition sheds light on the relation between the spectral efficiency and the duration of transmission phase in the general cases. In the cases where the UAVs are not equivalent, the spectral efficiency also first increases then decreases with the duration of transmission phase. This is because when $T_u = 0$, $N_{vd} = 0$, and when $T_u \to \infty$, $N_{vd} \to 0$.

V. Decentralized Trajectory Design

In this section, we first formulate the decentralized trajectory design problem of UAVs, and then analyze the problem in the reinforcement learning framework. After that, we describe the single-agent and multi-agent reinforcement learning algorithms under the framework, and propose an enhanced multi-UAV Q-learning algorithm to solve the UAV trajectory design problem efficiently.

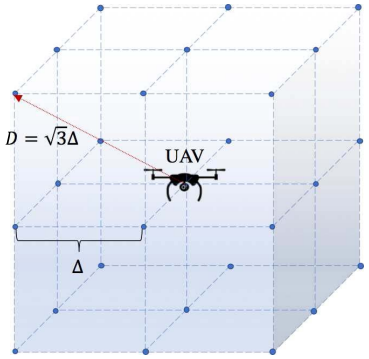

Fig. 5. Illustration on the set of available spatial points that the UAV can reach at the beginning of the next cycle.

A. UAV Trajectory Design Problem

Before the formulation of the trajectory design problem, we first set up the model to describe the UAVs' trajectories. In this paper, we focus on the cylindrical region with the maximum height $h_{\max}$ and the radius of the cross section $R_{\max}$ which satisfies $R_{\max} = \max\{R_i | R_i = \sqrt{X_i^2 + Y_i^2}, \forall i \in [1, N]\}$, since it is inefficient for the UAVs to move beyond the farthest task. Moreover, we assume that the space is divided into a finite set of discrete spatial points $\mathcal{S}_p$, which is arranged in a square lattice pattern as shown in Fig. 5. Therefore, the trajectory of UAV i starting from the kth cycle can be represented as the sequence of spatial points $\mathcal{S}_i^{(k)} = \{\boldsymbol{s}_i^{(k)}, \boldsymbol{s}_i^{(k+1)}, ...\}$ that the UAV locates at the beginning of each cycle.

To determine the trajectories, UAVs select their next spatial point at the beginning of each cycle. To be specific, UAV i locates at the spatial point $\boldsymbol{s}_i^{(k)}$ at the beginning of the kth cycle, and decides which spatial point $\boldsymbol{s}_i^{(k+1)}$ it will move to at the beginning of the $(k+1)$th cycle. After the UAV has selected its next spatial point, it will move to the point with a uniform speed and direction within this cycle. Moreover, the available spatial points that UAV i can reach is within the maximum distance it can fly in a cycle, which is denoted as D. We set the distance between two adjacent spatial points to be $\Delta = D/\sqrt{3}$, and thus, the available spatial point UAV i can fly to in the $k+1$ cycle is within a cube centered at $(x_i^{(k)}, y_i^{(k)}, h_i^{(k)})$ with side length equal to 2Δ, as illustrated in Fig 5.

In Fig. 5, there are at most 27 available spatial points that can be selected by the UAVs in each cycle. We denote the set of all the vectors from the center to the available spatial points as the available action set of the UAVs which is denoted as $\mathcal{A}$. Moreover, it is worth noticing that when the UAV is at the marginal location (e.g., flying at the minimum height), there are less available actions to be selected. To handle the differences among the available action sets at different spatial points, we denote the available action set at the spatial point $\boldsymbol{s}$ as $\mathcal{A}(\boldsymbol{s})$.

In this paper, we consider the utility of each UAV to be the total number of successful valid sensory data transmissions

for its task. Therefore, the UAVs have incentive to maximize the total amount of successful valid sensory data transmission by designing their trajectories. Besides, we assume that the UAVs have discounting valuation on the successfully transmitted valid sensory data. Specifically, for the UAVs in the kth cycle, the successfully valid sensory data transmitted in the k'th cycle is worth only $\rho^{|k'-k|}$ [$\rho \in [0, 1)$] the successful valid sensory data transmitted in the current cycle, due to the timeliness requirements of real-time sensing tasks. Therefore, at the beginning of kth cycle, the utility of UAV i is defined as the *total discounted rewards* in the future, and can be denoted as

$$U_i^{(k)} = \sum_{n=0}^{\infty} \rho^n R_i^{(k+n)} \tag{12}$$

in which $R_i^{(k)}$ denotes the reward of UAV i in the kth cycle, and $R_i^{(k)} = 1$ if valid sensory data is successfully transmitted to the BS by UAV i in the kth cycle, otherwise, $R_i^{(k)} = 0$.

Based on the above assumptions, the UAV trajectory design problem can be formulated as

$$\max_{\mathcal{S}_i^{(k)}} \quad U_i^{(k)} = \sum_{n=0}^{\infty} \rho^n R_i^{(k+n)} \tag{13}$$

$$\text{s.t.} \quad \boldsymbol{s}_i^{(k'+1)} - \boldsymbol{s}_i^{(k')} \in \mathcal{A}(\boldsymbol{s}_i^{(k')}), \ k' \in [k, \infty). \tag{13a}$$

B. Reinforcement Learning Framework

Generally, the UAV trajectory design problem (13) is hard to solve since the rewards of the UAVs in the future cycles are influenced by the trajectories of all UAVs, which are determined in a decentralized manner and hard to model. Fortunately, reinforcement learning is able to deal with the problem of agent programming in environment with deficient understanding, which removes the burden of developing accurate models and solving the optimization with respect to those models [23]. For this reason, we adopt reinforcement learning to solve the UAV trajectory design problem.

To begin with, we formulate a reinforcement learning framework for the problem. With the help of [24], the reinforcement learning framework can be given as follows, in which the superscript cycle index k is omitted for description simplicity.

Definition 1: A reinforcement learning framework for UAV trajectory design problem is described by a tuple $<\mathcal{S}_1, \ldots, \mathcal{S}_N, \mathcal{A}_1, \ldots, \mathcal{A}_N, \mathcal{T}, p_{R,1}, \ldots, p_{R,N}, \rho>$.

1) $\mathcal{S}_1, \ldots, \mathcal{S}_N$ are finite state spaces of all the possible locations of the UAVs at the beginning of each cycle, and the state space of UAV i equals to the finite spatial space, i.e., $\mathcal{S}_i = \mathcal{S}_p, \ \forall i \in [1, N]$.
2) $\mathcal{A}_1, \ldots, \mathcal{A}_N$ are the corresponding finite sets of actions available to each agents. The set $\mathcal{A}_i$ consists of all the available action of UAV i, i.e., $\mathcal{A}_i = \mathcal{A}, \ \forall i \in [1, N]$.
3) $\mathcal{T} : \prod_{i=1}^{N} \mathcal{S}_i \times \prod_{i=1}^{N} \mathcal{A}_i \to (\mathcal{S}_p)^N$ is the state transition function. It maps the location profile and the action profile of the UAVs in a certain cycle to the location profile of the UAVs in the next cycle.
4) $p_{R,i} : \prod_{j=1}^{N} \mathcal{S}_j \times \prod_{j=1}^{N} \mathcal{A}_j \to \Pi(0, 1), \ i = 1, \ldots, N$ represents the reward function for UAV i. That is to say, it maps the location profile and action profile of all the UAVs in the current cycle to the probability for UAV i to get unit reward from successful valid sensory data transmission.
5) $\rho \in [0, 1)$ is the discount factor, which indicates UAVs' evaluation of the rewards that obtained in the future or in the past.

In the framework, the rewards of the UAVs are informed by the BS. Specifically, we assume that the BS informs each UAV whether it has transmitted valid sensory data in a certain cycle by the BS beacon at the beginning of the next cycle. For each UAV, it obtains one reward if the BS informs that the valid sensory data has been received successfully. Therefore, the probability for UAV i to obtain one reward is equal to the probability for it to transmit valid sensory data successfully in the cycle, i.e., $p_{R,i} = p_{\text{sTx},i}$. Since the probability of successful valid sensory data transmission is influenced by both the successful sensing probability and the successful transmission probability, the UAV's trajectory learning process is associated with the sensing and transmission processes through the obtained reward in each cycle.

Under the reinforcement learning framework for the UAV trajectory design, the following two kinds of reinforcement learning algorithms can be adopted, which are single-agent Q-learning algorithm and multiagent Q-learning algorithm.

1) Single-Agent Q-Learning Algorithm: One of the most basic reinforcement learning algorithm is the single-agent Q-learning algorithm [25]. It is a form of model-free reinforcement learning and provides a simple way for the agent to learn how to act optimally. The agent selects actions by following its policy in each state, which is denoted as $\pi(\boldsymbol{s}), \ \boldsymbol{s} \in \mathcal{S}$ and is a mapping from states to actions. The essence of the algorithm is to find the Q-value of each state and action pairs, which is defined as the accumulated reward received when taking the action in the state and then following the policy thereafter. In its simplest form, the agent maintains a table containing its current estimated Q-values, which is denoted as $Q(\boldsymbol{s}, \boldsymbol{a})$ with $\boldsymbol{a}$ indicating the action. It observes the current state $\boldsymbol{s}$ and selects the action $\boldsymbol{a}$ that maximizes $Q(\boldsymbol{s}, \boldsymbol{a})$ with some exploration strategies. Q-learning has been studied extensively in single-agent tasks where only one agent is acting alone in a stationary environment.

In the UAV trajectory design problem, multiple UAVs take actions at the same time. When each UAV adopts the single-agent Q-learning algorithm, it assumes that the other agents are part of the environment. In the UAV trajectory design problem, the single-agent Q-learning algorithm can be adopted as follows. For UAV i, the policy of UAV i to select action is

$$\pi_i(\boldsymbol{s}_i) = \underset{\boldsymbol{a}_i' \in \mathcal{A}(\boldsymbol{s}_i)}{\text{argmax}} \, Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i). \tag{14}$$

Upon receiving a reward R_i after the end of the cycle and observing the next state $\boldsymbol{s}_i'$, it updates its table of Q-values according to the following rule:

$$Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) \leftarrow Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) + \alpha \left(R_i + \rho \max_{\boldsymbol{a}_i' \in \mathcal{A}(\boldsymbol{s}_i)} Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) \right) \tag{15}$$

Algorithm 2 Single-Agent Q-Learning Algorithm for UAV Trajectory Design Problem of UAV i

Initialize $Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) \coloneqq 0$, $\forall \boldsymbol{s}_i \in \mathcal{S}_p$, $\boldsymbol{a}_i \in \mathcal{A}_i(\boldsymbol{s}_i)$, $\pi_i(\boldsymbol{s}_i) \coloneqq \text{Rand}(\{\mathcal{A}_i(\boldsymbol{s}_i)\})$.
for $k = 1$ to max-number-of-cycles **do**
  With probability $1 - \epsilon^{(k)}$, choose action $\boldsymbol{a}_i$ from the policy at the state $\pi_i(\boldsymbol{s}_i)$, and with probability $\epsilon^{(k)}$, randomly choose an available action for exploration;
  Perform the action $\boldsymbol{a}_i$ in the k-th cycle;
  Observe the transited state $\boldsymbol{s}_i'$ and the reward R_i;
  Select $\boldsymbol{a}_i'$ in the transited state $\boldsymbol{s}_i'$ according to $\pi_i(\boldsymbol{s}_i')$;
  Update the Q-value for the former state–action pair, i.e., $Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) \coloneqq Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i) + \alpha^{(k)}(R_i + \rho Q(\boldsymbol{s}_i', \boldsymbol{a}_i') - Q_i(\boldsymbol{s}_i, \boldsymbol{a}_i))$;
  Update the policy at state $\boldsymbol{s}_i$ as $\pi_i(\boldsymbol{s}_i, \boldsymbol{a}_i') \coloneqq 1$, where $\boldsymbol{a}_i = \text{argmax}_{\boldsymbol{m} \in \mathcal{A}_i(\boldsymbol{s}_i)} Q_i(\boldsymbol{s}_i, \boldsymbol{m})$;
  Update the state $\boldsymbol{s}_i \coloneqq \boldsymbol{s}_i'$;
end for

in which $\alpha \in (0, 1)$ denotes the learning rate. With the help of [26], the single-agent Q-learning algorithm for UAV trajectory design problem can be given in Algorithm 2.

2) Multiagent Q-Learning Algorithm: Although single-agent Q-learning algorithm has many favorable properties, such as small state space and easy implementation, it lacks of consideration on the states and the strategic behaviors of other agents. Therefore, we adopt a multiagent Q-learning algorithm called opponent modeling Q-learning to solve the UAV trajectory design problem, which enables the agent to adapt to other agents' behaviors.

Opponent modeling Q-learning is an effective multiagent reinforcement learning algorithm [27], [28], in which explicit models of the other agents are learned as stationary distributions over their actions. These distributions, combined with learned joint state–action values from standard temporal differencing, are used to select an action in each cycle. Specifically, at the beginning of each cycle, UAV i selects the action a_i which maximizes the expected discounted reward according to the observed frequency distribution of other agents' action in the current state $\boldsymbol{s}$, i.e., its policy at $\boldsymbol{s}$ is

$$\pi_i(\boldsymbol{s}) = \underset{\boldsymbol{a}_i'}{\text{argmax}} \sum_{\boldsymbol{a}_{-i}'} \frac{\Phi(\boldsymbol{s}, \boldsymbol{a}_{-i}')}{n(\boldsymbol{s})} Q_i\left(\boldsymbol{s}, \left(\boldsymbol{a}_i', \boldsymbol{a}_{-i}'\right)\right) \tag{16}$$

in which the state $\boldsymbol{s} = (s_1, \ldots, s_N)$ is location profile of all the UAVs, $\Phi(\boldsymbol{s}, \boldsymbol{a}_{-i}')$ denotes the number of times for the agents other than agent i to select action profile $\boldsymbol{a}_{-i}'$ in the state $\boldsymbol{s}$, and $n(\boldsymbol{s})$ is the total number of times the state $\boldsymbol{s}$ has been visited.

After the agent i observes the transited state $\boldsymbol{s}'$, the action profile $\boldsymbol{a}_{-i}$, and the reward in the previous cycle, it updates the Q-value as follows:

$$Q_i(\boldsymbol{s}, (\boldsymbol{a}_i, \boldsymbol{a}_{-i})) = (1 - \alpha) Q_i(\boldsymbol{s}, (\boldsymbol{a}_i, \boldsymbol{a}_{-i})) + \alpha\left(R_i + \rho V_i\left(\boldsymbol{s}'\right)\right) \tag{17}$$

in which $V_i(\boldsymbol{s}') = \max_{\boldsymbol{a}_i'} \sum_{\boldsymbol{a}_{-i}'} ([\Phi(\boldsymbol{s}', \boldsymbol{a}_{-i}')]/n(\boldsymbol{s}')) Q(s, (\boldsymbol{a}_i', \boldsymbol{a}_{-i}'))$ indicating that agent i selects the action in the transited state $\boldsymbol{s}'$ to maximize the expected discounted reward based on the empirical action profile distribution then. With the help of [28], the multiagent Q-learning algorithm for UAV trajectory design can be given in Algorithm 3.

Algorithm 3 Opponent Modeling Q-Learning Algorithm for UAV Trajectory Design Problem of UAV i

Initialize $Q_i(\boldsymbol{s}, (\boldsymbol{a}_i, \boldsymbol{a}_{-i})) \coloneqq 0$, $\forall \boldsymbol{s} \in \prod_i^N \mathcal{S}_i$, $\boldsymbol{a}_i \in \mathcal{A}_i(\boldsymbol{s}_i)$, $\boldsymbol{a}_{-i} \in \prod_{j \neq i}^N \mathcal{A}_j$, $\pi_i(\boldsymbol{s}_i) \coloneqq \text{Rand}(\{\mathcal{A}_i(\boldsymbol{s}_i)\})$.
for $k = 1$ to max-number-of-cycles **do**
  With probability $1 - \epsilon^{(k)}$, choose action $\boldsymbol{a}_i$ according to the policy $\pi_i(\boldsymbol{s})$, or with probability $\epsilon^{(k)}$, randomly choose an available action for exploration;
  Perform the action $\boldsymbol{a}_i$ in the k-th cycle;
  Observe the transited state $\boldsymbol{s}'$ and the reward R_i;
  Select action $\boldsymbol{a}_i'$ in the transited state $\boldsymbol{s}'$ according to the strategy in state $\boldsymbol{s}'$ according to (16);
  Update the Q-value for the former state–action pair according to (17);
  Update the policy at state $\boldsymbol{s}$ to the action that maximizes the expected discounted reward according to (16);
  Update the state $\boldsymbol{s} \coloneqq \boldsymbol{s}'$;
end for

C. Enhanced Multi-UAV Q-Learning Algorithm for UAV Trajectory Design

In the opponent modeling multiagent reinforcement learning algorithm, UAVs need to tackle too many state–action pairs, resulting in a slow convergence speed. Therefore, we enhance the opponent modeling Q-learning algorithm in the UAV trajectory design problem by reducing the available action set and adopting an model-based reward representation. These two enhancing approaches are elaborated as follows, and the proposed enhanced multi-UAV Q-learning algorithm is given in Algorithm 4.

1) Available Action Set Reduction: It can be observed that although the UAVs are possible to reach all the spatial points in the finite location space $\mathcal{S}_p$, it makes no sense for the UAVs to move away from the vertical planes passing the BS and their tasks, i.e., the BS-task planes, which decreases the successful sensing probability as well as the successful transmitting probability. Therefore, we confine the available action set of the UAV to the actions which does not increase the horizontal distance between it and the BS-task plane, which is illustrated in Fig. 6 (the arrows).

Ideally, the UAVs should be in the BS-task plane and only move within the plane. However, since the spatial space is discrete, the UAV cannot only move within the BS-task plane, and needs to deviate from the plane in order to reach different locations on or near the plane. Therefore, we mitigate the constraint by allowing the UAVs to move to the spatial points from which the distance to their BS-task planes are within Δ, as the spots shown in Fig. 6. The reduced available action set of UAV i at state $\boldsymbol{s}_i = (x_i, y_i, h_i)$ can be defined as follows.

Definition 2 (Reduced Available Action Set of UAV): Supposing UAV i at the state $\boldsymbol{s}_i = (x_i, y_i, h_i)$, the location

Algorithm 4 Enhanced Multi-UAV Q-Learning Algorithm for Trajectory Design Problem of UAV i

1: **for** $k = 1$ to max-number-of-cycles **do**
2: Obtain the available action set $\mathcal{A}_j^+(s_j)$, $\forall j \in [1, N]$ for the current state $\boldsymbol{s}$ according to Def. 2.
3: **if** $\boldsymbol{s}$ has not been visited before **then**
4: Initialize $Q_i(\boldsymbol{s}, \boldsymbol{a}) \coloneqq p_{\mathrm{sTx},i}(\boldsymbol{s}, \boldsymbol{a})$, $\forall \boldsymbol{s} \in \prod_i^N \mathcal{S}_i$, $\boldsymbol{a} \in \prod_{j=1}^N \mathcal{A}_j^+(\boldsymbol{s}_j)$, $\pi_i(\boldsymbol{s}_i) \coloneqq \mathrm{Rand}(\{\mathcal{A}_i^+(\boldsymbol{s}_i)\})$.
5: **end if**
6: With probability $1-\epsilon^{(k)}$, choose action a_i from the strategy at the state $\pi_i(\boldsymbol{s})$, or with probability $\epsilon^{(k)}$, randomly choose an available action for exploration;
7: Perform action $\boldsymbol{a}_i$ in the k-th cycle;
8: Observe the transited state $\boldsymbol{s}'$ and the action profile $\boldsymbol{a}$ in the previous state;
9: Select action $\boldsymbol{a}_i'$ in the transited state $\boldsymbol{s}'$ according to policy $\pi_i(\boldsymbol{s}')$ defined in (16);
10: Calculate the probability of successful valid sensory data transmission in the previous cycle $p_{\mathrm{sTx},i}(\boldsymbol{s}, \boldsymbol{a})$.
11: Update the Q-function for the former state–action pair according to (17), with $R_i \coloneqq p_{\mathrm{sTx},i}(\boldsymbol{s}, \boldsymbol{a})$;
12: Update the policy at state $\boldsymbol{s}$ to the action that maximizes the expected discounted rewards according to (16);
13: Update the state $\boldsymbol{s} \coloneqq \boldsymbol{s}'$;
14: **end for**

of its task at $\boldsymbol{S}_i = (X_i, Y_i, 0)$, and the location of BS at $\boldsymbol{S}_0 = (0, 0, H_0)$, the action $\boldsymbol{a} = (a_x, a_y, a_h)$ in the reduced available action set $\mathcal{A}_i^+(\boldsymbol{s}_i)$ satisfies the following conditions.

1) $\mathrm{Dist}(\boldsymbol{s}_i + \boldsymbol{a}; \boldsymbol{S}_i, \boldsymbol{S}_0) \leq \mathrm{Dist}(\boldsymbol{s}_i; \boldsymbol{S}_i, \boldsymbol{S}_0)$ or $\mathrm{Dist}(\boldsymbol{s}_i + \boldsymbol{a}; \boldsymbol{S}_i, \boldsymbol{S}_0) \leq \Delta$.
2) $x_i + a_x \in [\min(x_i, X_i, 0), \max(x_i, X_i, 0)]$, $y_i + a_y \in [\min(y_i, Y_i, 0), \max(y_i, Y_i, 0)]$, and $h_i + a_h \in [h_{\min}, h_{\max}]$.

Here, $\mathrm{Dist}(\boldsymbol{s}_i; \boldsymbol{S}_i, \boldsymbol{S}_0)$ denotes the horizontal distance between s_i to the vertical plane passing through $\boldsymbol{S}_i$ and $\boldsymbol{S}_0$.

In Definition 2, condition (1) limits the actions to those leading the UAV to the spatial points near the BS-task plane, and condition (2) stops the UAV from moving away from the cross region between the location of its task and the BS.

Moreover, instead of initializing the Q-values for all the possible state–action pairs at the beginning, we propose that the UAVs initialize the Q-values only for the reached state and the actions within the reduced available action set. In this way, the state sets of the UAVs are reduced to much smaller sets, which makes the learning more efficient.

2) Model-Based Reward Representation: In both the single-agent Q-learning algorithm and the opponent modeling Q-learning algorithm, the UAVs update their Q-values based on the information provided by the BS, which indicates the validity of the latest transmitted sensory data. Nevertheless, since the UAVs can only observe the reward to be either 1 or 0, the Q-values converge slowly and the performance of the algorithms is likely to be poor.

Therefore, in this paper, we propose that the UAVs update their Q-values based on the probabilities of successful valid sensory data transmission. Specifically, UAV i calculates the probability $p_{\mathrm{sTx},i}$ after observing the state–action profile $(\boldsymbol{s}, (a_i, \boldsymbol{a}_{-i}))$ in the latest cycle according to (10), and takes it as the reward R_i for the kth cycle.

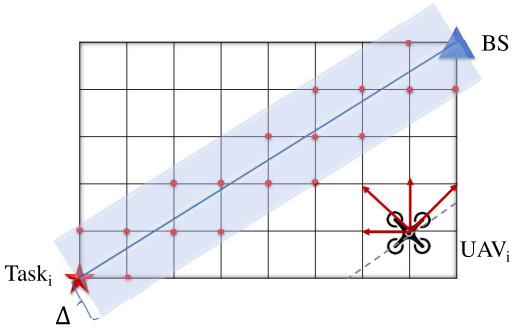

Fig. 6. Illustration on the constrained available action set of UAV i.

Moreover, to make the learning algorithm converge with higher speed, in the initialization of the enhanced multi-UAV Q-learning algorithm, we propose that UAV i initializes its $Q_i(\boldsymbol{s}, (a_i, \boldsymbol{a}_{-i}))$ with the calculated $p_{sTx,i}$ for the state–action pair. In this way, the update of the Q-values is more accurate and the learning algorithm is expected to have higher convergence speed.

Remark (Signaling in UAVs' Learning Algorithms): In the above mentioned reinforcement learning algorithms, UAVs need to know the locations profile in the beginning of each cycle, and the rewards in the last cycle associated with their actions taken. This information gathering can be done in beaconing phase of the cycle as described in Section III-A, in which the BS can include the rewards of UAVs in the last cycle in the broadcasted beacon.

D. Analysis of Reinforcement Learning Algorithms

In the final part of this section, we analyze the convergence, the complexity, and the scalability of the proposed reinforcement learning algorithms.

1) Convergence Analysis: For the convergence of the reinforcement learning algorithms, it has been proved in [29] that under certain conditions, single agent Q-learning algorithm is guaranteed to converge to the optimal Q^*. In consequence, the policy π of the agent converges to the optimal policy π^*. It can be summarized in the following Theorem 1.

Theorem 1 (Convergence of Q-Learning Algorithm): The Q-learning algorithm given by

$$Q^{(k+1)}\left(\boldsymbol{s}^{(k)}, \boldsymbol{a}^{(k)}\right) = \left(1 - \alpha^{(k)}\right) Q^{(k)}\left(\boldsymbol{s}^{(k)}, \boldsymbol{a}^{(k)}\right) + \alpha^{(k)}\left[R\left(\boldsymbol{s}^{(k)}, \boldsymbol{a}^{(k)}\right) + \gamma \max_{\boldsymbol{a}'} Q\left(\boldsymbol{s}^{(k+1)}, \boldsymbol{a}'\right)\right] \tag{18}$$

converges to the optimal Q^* values if the following conditions are satisfied.

1) The state and action spaces are finite.
2) $\sum_k \alpha^{(k)} = \infty$ and $\sum_k (\alpha^{(k)})^2 < \infty$.
3) The variance of $R(\boldsymbol{s}, \boldsymbol{a})$ is bounded.

Therefore, in the multiagent reinforcement learning cases, if other agents play, or converge to stationary strategies, the single-agent reinforcement learning algorithm also converges to the optimal policy.

However, it is generally hard to prove convergence with other agents that learning simultaneously. This is because when the agent is learning the Q-value of its actions in the presence of other agents, it faces a nonstationary environment and the convergence of Q-values is not guaranteed. The theoretical convergence of the Q-learning in multiagent cases are guaranteed only in few situations, such as in the iterated dominance solvable games and the team games [26]. Like single-agent Q-learning algorithm, the convergence of opponent modeling Q-learning is not generally guaranteed, except for in the setting of iterated dominance solvable games and team matrix game [28].

To handle this problem, in this paper, we adopt $\alpha^{(k)} = 1/k^{2/3}$ in [30] which satisfies the conditions for convergent in single-agent Q-learning, and analyze the convergence of the reinforcement learning in the multiagent case through simulation results which will be provided in Section VI.

2) Complexity Analysis: For the single-agent Q-learning algorithm, the computational complexity in each iteration is $\mathcal{O}(1)$, since the UAV does not consider the other UAVs in the learning process. For the multiagent Q-learning algorithm, the computational complexity in each iteration is $\mathcal{O}(2^N)$, due to the calculation of the expected discounted reward in (16).

As for the proposed enhanced multi-UAV Q-learning algorithm, each UAV needs to calculate the probability for successful valid data transmission based on Algorithm 1. It can be seen that the recursive Algorithm 1 runs for at most 2^{CT_u} times and each iteration has the complexity of $\mathcal{O}(N)$, which makes its overall complexity $\mathcal{O}(N)$. Therefore, the complexity of the proposed enhanced algorithm is still $\mathcal{O}(2^N)$, due to the expectation over the joint action space.

Although the computational complexity of the enhanced multi-UAV Q-learning algorithm in each iteration is in the same order with opponent modeling Q-learning algorithm, it reduces the computational complexity and speeds up the convergence by the following means.

1) Due to the available action set reduction, the available action set of each UAV is at least reduced to one-half its original size. This makes the joint action space to be 2^N times smaller.
2) The reduced available action set leads to a much smaller state space of each UAV. For example, for UAV i and its task at $(X_i, Y_i, 0)$, the original size of its state space can be estimated as $\pi R_{\max}^2 (h_{\max} - h_{\min})/\Delta^3$, and the size of its state space after available action set reduction is $2(X_i + Y_i)(h_{\max} - h_{\min})/\Delta^2$, which is $2\Delta/(\pi R_{\max})$ of the original one.
3) The proposed algorithm adopts model-based reward representation, which makes the Q-value updating to be more precise, and saves the number of iterations needed to estimate the accurate Q-values of the state–action pairs.

3) Scalability Analysis: With the growth of the number of UAVs, the state spaces of the UAVs in the multiagent Q-learning algorithm and the enhanced multi-UAV Q-learning algorithm grow exponentially. Besides, it can be seen that the enhanced multi-UAV Q-learning algorithm still has exponential computational complexity in each iteration, and thus, it is not suitable for large-scale UAV networks.

To adapt the algorithms to large-scale UAV networks, the reinforcement learning methods need to be combined with function approximation approaches in order to estimate Q-values efficiently. The function approximation approaches take examples from a desired function, Q-function in the case of reinforcement learning, and generalize from them to construct an approximation of the entire function. In this regard, it can be used to estimate the Q-values of the state–action pairs in the entire state space efficiently when the state space is large.

TABLE I
SIMULATION PARAMETERS

Parameter	Value
BS height H	25 m
Number of UAVs N	3
Noise power N_0	-85 dBm
BS decoding threshold γ_{th}	10 dB
UAV sensing parameter λ	10^{-3}/s
UAV transmit power P_u	10 dBm
Duration of frame t_f	0.1 s
Distance between adjacent spatial points Δ	25 m
UAVs' minimum flying height $h_{\min}$	50 m
UAVs' maximum flying height $h_{\max}$	150 m
Discount ratio ρ	0.9
Duration of beaconing phase in frames T_b	3
Duration of sensing phase in frames T_s	5
Duration of transmission phase in frames T_u	5

VI. SIMULATION RESULTS

In order to evaluate the performance of the proposed reinforcement learning algorithms for the UAV trajectory design problem, simulation results are presented in this section. Specifically, we use MATLAB to build a frame-level simulation of the UAV sense-and-send protocol, based on the system model described in Section II and the parameters in Table I. Besides, the learning ratio in the algorithm is set to be $\alpha^{(k)} = 1/k^{2/3}$ in order to satisfy the converge condition in Theorem 1, and the exploration ratio is set to be $\epsilon^{(k)} = 0.8e^{-0.03k}$, which approaches 0 when $k \to \infty$.

Fig. 7 shows UAV 1's probability of successful valid sensory data transmission versus UAV 1's height and its distance to the BS, given that task 1 is located at $(500, 0, 0)$, and the locations of UAV 2 and UAV 3 are fixed at $(-125, 125, 75)$, $(-125, -125, 75)$, respectively. It can be seen that the optimal point at which UAV 1 has the maximum probability of successful valid sensory data transmission is located in the region between BS and task 1. This is because when the UAV approaches the BS, its successful sensing probability drops, and when the UAV approaches the task, its successful transmission probability suffers. Besides, it is shown that the optimal point for UAV 1 to sense and send is *above*, rather than *on* the BS-task line, where UAV 1 can be closer to both the BS and its task. This is because in

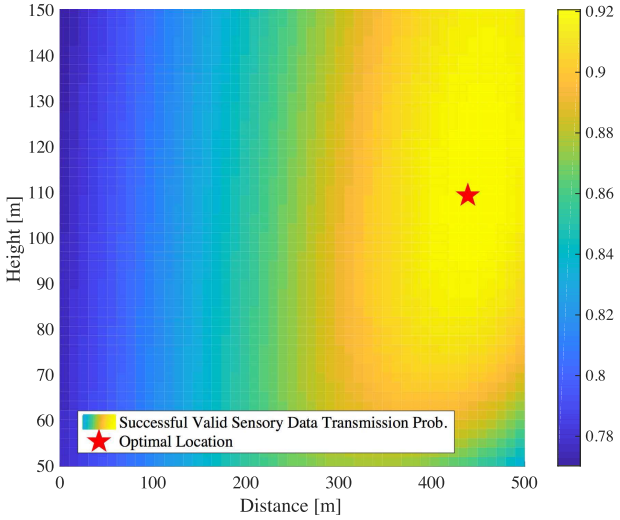

Fig. 7. Successful valid sensory data transmission probability versus the location in the task-BS surface.

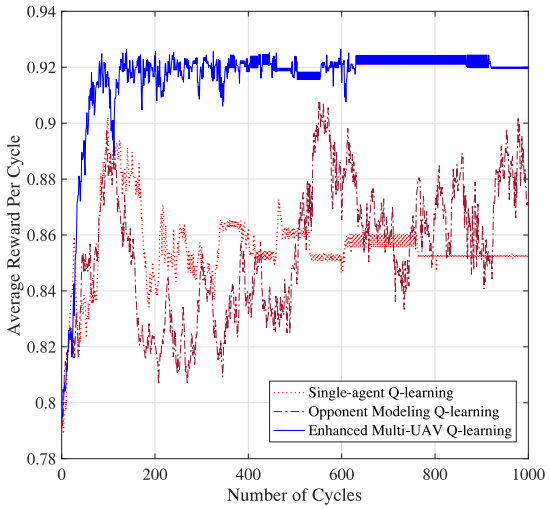

Fig. 8. UAVs' average reward per cycle versus number of cycles of different reinforcement learning algorithms.

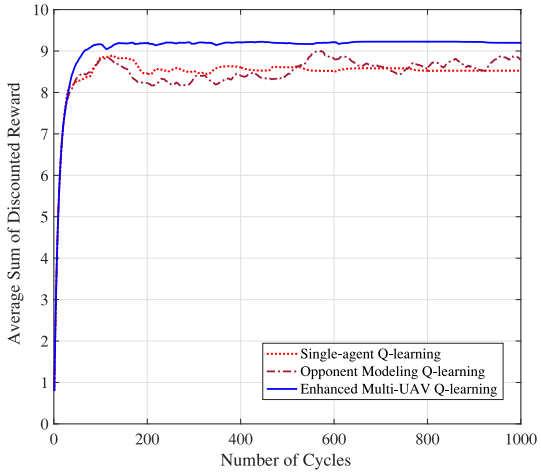

Fig. 9. UAVs' average discounted reward versus number of cycles of different reinforcement learning algorithms.

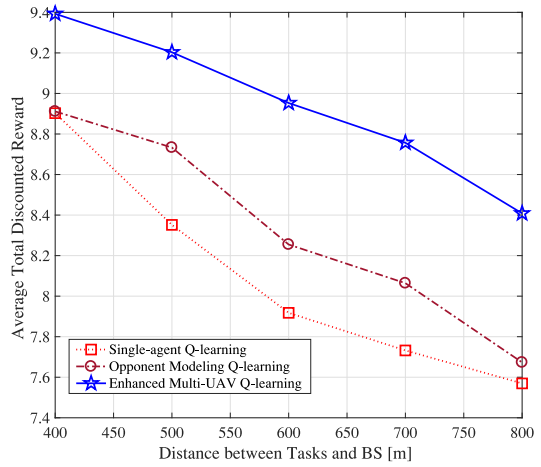

Fig. 10. UAVs' average discounted reward versus distance between tasks and BS in different reinforcement learning algorithms.

the transmission model in Section II-B, with the increment of the height of the UAV, the LoS probability increases, and thus, the successful uplink transmission probability of the UAV increases.

Figs. 8 and 9 show the average reward per cycle and the average total discounted reward of the UAVs versus the number of cycles in different reinforcement learning algorithm, in which tasks 1–3 are located at (500, 0, 0), $(-250\sqrt{2}, 250\sqrt{2}, 0)$, and $(-250\sqrt{2}, -250\sqrt{2}, 0)$, respectively. It can be seen that compared to the single-agent Q-learning algorithm, the proposed algorithm converges to higher average rewards for the UAVs. This is because the UAV in the enhanced multi-UAV Q-learning algorithm takes the states of all the UAVs into consideration, which makes the estimation of Q-values more accurate. Besides, it can also be seen that compared to the opponent modeling Q-learning algorithm, the proposed algorithm converges faster, due to the available action set reduction and the model-based reward representation.

Moreover, in Fig. 10, we can observe that for different distances between the tasks and the BS, the proposed algorithm converges to higher average total discounted rewards for UAVs after 1000 cycles compared to the other algorithms. It can be seen that the average total discounted reward in the three algorithms decreases with the increment of the distance between the BS and the tasks. Nevertheless, the decrement in the proposed algorithm is less than those in the other algorithms. This indicates that the proposed algorithm is more robust to the variance of the tasks' locations.

Fig. 11 shows the average number of successful valid sensory data transmissions versus the duration of the transmission phase T_u in the proposed algorithm, under different conditions of the distance between the tasks and the BS. It can be seen that the average number of successful valid sensory data transmissions per second first increases and then decreases with the increment of T_u. This is because when T_u is small, the successful uplink transmission probability increases rapidly with the increment of T_u. However, when T_u is large, the successful uplink transmission probability is already high and increases slightly when T_u becomes larger. Therefore, the

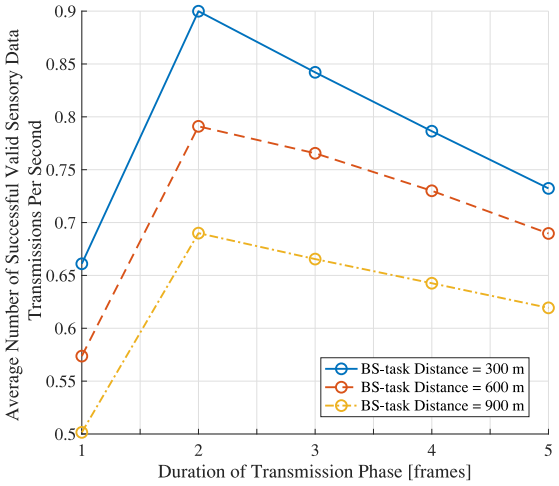

Fig. 11. Average number of successful valid sensory data transmissions per second versus duration of transmission phase T_u under different task distance conditions.

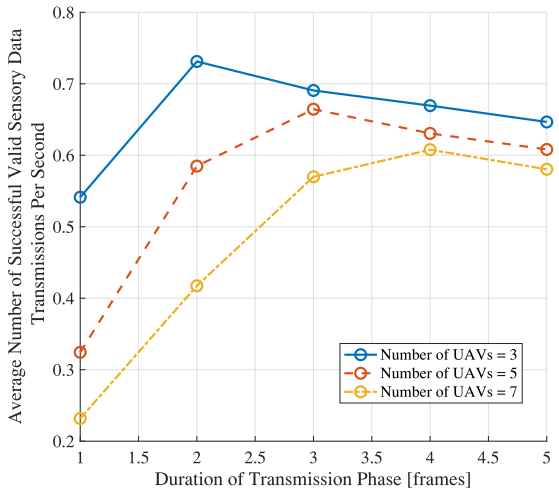

Fig. 12. Average number of successful valid sensory data transmissions per second versus duration of transmission phase T_u under different number of UAVs. Distance between the BS and the tasks = 800 m.

average number of successful valid sensory data transmissions per second drops due to the increment of cycles' duration.

Fig. 12 shows the average number of successful valid sensory data transmissions per second versus T_u, under the different conditions of the number of UAVs. It can be seen that when the number of UAVs increases, the average number of successful valid sensory data transmissions per second decreases. This is because the competition among the UAVs for the limited SCs becomes more intensive. Besides, when the number of UAVs increases, the optimal duration of the transmission phase becomes longer. This indicates that in order to achieve optimal spectral efficiency, the BS needs to increase the duration of the transmission phase when the number of UAVs in the network increases.

VII. Conclusion

In this paper, we have considered the scenarios where UAVs performs real-time sensing tasks, and solved the distributed UAV trajectory design problem by using reinforcement learning. First, we have derived a sense-and-send protocol to coordinate multiple UAVs. To evaluate the performance of the protocol, we have solved the probability of successful valid sensory data transmission in the protocol by using nested Markov chains. Then, after the UAV trajectory problem formulation under the reinforcement learning framework, we have proposed the enhanced multi-UAV Q-learning algorithm to solve it efficiently. The simulation results have shown that the proposed algorithm converged faster and achieved higher utilities for the UAVs. It has been also shown in simulation that our proposed algorithm is more robust to the increment of tasks' distance, compared to single-agent and opponent modeling Q-learning algorithms. Moreover, the simulation also have shown that the BS needs to increase the duration of the transmission phase to improve the shown efficiency when the number of UAVs increases.

Appendix
Proof of Proposition 1

Denoting the UAVs' probability for successful uplink transmission as p_u and their probability for successful sensing as p_s, we can calculate the average number of valid sensory data transmissions per second by

$$N_{vd} = N \cdot \frac{p_s\left(1-(1-p_u)^{\frac{CT_u}{N}}\right)}{(T_b+T_s+T_u)t_f}$$

in which t_f is the duration of single frame in seconds.

The partial derivative of N_{vd} with respect to T_u can be calculated as

$$\frac{\partial N_{vd}}{\partial T_u} = \frac{p_s F(T_u)}{t_f(T_b+T_s+T_u)^2}$$

in which $F(T_u) = p_f^{\frac{CT_u}{N}}(N - C(T_b+T_s+T_u)\ln p_f) - N$, and $p_f = 1-p_u$. Taking partial derivative of $F(T_u)$ with regard to T_u, we can derive that $\partial F(T_u)/\partial T_u = -C^2 p_f^{CT_u/N}(T_s + T_b + T_u)\ln p_f/N < 0$. Besides, when $T_u \to \infty$, $F(T_u) \to -N$ and $N_{vd} \to 0$, and when $T_u = 0$, $N_{vd} = 0$. Therefore, $\partial F(T_u)/\partial T_u < 0$ indicates that there is a unique maximum point for N_{vd} when $T_u \in (0, \infty)$.

The maximum of N_{vd} is reached when $F(T_u^*) = 0$, in which T_u^* can be obtained by

$$T_u^* = \frac{N}{C\ln p_f}\left(1 + W_{-1}\left(-\frac{p_f^{\frac{CT_u}{N}}}{e}\right)\right) - T_b - T_s$$

where $W_{-1}(\cdot)$ denotes the lower branch of Lambert-W function [22]. ■

References

[1] J. Wang *et al.*, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, Sep. 2017.

[2] A. Puri, K. P. Valavanis, and M. Kontitsis, "Statistical profile generation for traffic monitoring using real-time UAV based video data," in *Proc. Mediterr. Conf. Control Autom.*, Athens, Greece, Jun. 2007, pp. 1–6.

[3] B. H. Y. Alsalam, K. Morton, D. Campbell, and F. Gonzalez, "Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture," in *Proc. Aerosp. Conf.*, Big Sky, MT, USA, Mar. 2017, pp. 1–12.
[4] D. W. Casbeer, D. B. Kingston, R. W. Beard, and T. W. McLain, "Cooperative forest fire surveillance using a team of small unmanned air vehicles," *Int. J. Syst. Sci.*, vol. 37, no. 6, pp. 351–360, Feb. 2006.
[5] B. Van der Bergh, A. Chiumento, and S. Pollin, "LTE in the sky: Trading off propagation benefits with interference costs for aerial nodes," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 44–50, May 2016.
[6] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular UAV-to-X communications: Design and optimization for multi-UAV networks," *IEEE Trans. Wireless Commun.*, submitted for publication. [Online]. Available: https://arxiv.org/abs/1801.05000
[7] M. Thammawichai, S. P. Baliyarasimhuni, E. C. Kerrigan, and J. B. Sousa, "Optimizing communication and computation for multi-UAV information gathering applications," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 2, pp. 601–615, Apr. 2018.
[8] J. Tisdale, Z. Kim, and J. K. Hedrick, "Autonomous UAV path planning and estimation," *IEEE Robot. Autom. Mag.*, vol. 16, no. 2, pp. 35–42, Jun. 2009.
[9] I. Maza, K. Kondak, M. Bernard, and A. Ollero, "Multi-UAV cooperation and control for load transportation and deployment," *J. Intell. Robot. Syst.*, vol. 57, nos. 1–4, p. 417, Jan. 2010.
[10] G. Gu, P. Chandler, C. J. Schumacher, A. Sparks, and M. Pachter, "Optimal cooperative sensing using a team of UAVs," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 4, pp. 1446–1458, Oct. 2006.
[11] Y. Yang, Z. Zheng, K. Bian, L. Song, and Z. Han, "Real-time profiling of fine-grained air quality index distribution using UAV sensing," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 186–198, Feb. 2018.
[12] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 161–164, Jan. 2018.
[13] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *Proc. IEEE ICC*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.
[14] S. Zhang, H. Zhang, B. Di, and L. Song, "Cellular controlled cooperative unmanned aerial vehicle networks with sense-and-send protocol," *IEEE Internet Things J.*, to be published.
[15] V. V. Shakhov and I. Koo, "Experiment design for parameter estimation in probabilistic sensing models," *IEEE Sensors J.*, vol. 17, no. 24, pp. 8431–8437, Dec. 2017.
[16] A. Chakraborty, R. R. Rout, A. Chakrabarti, and S. K. Ghosh, "On network lifetime expectancy with realistic sensing and traffic generation model in wireless sensor networks," *IEEE Sensors J.*, vol. 13, no. 7, pp. 2771–2779, Jul. 2013.
[17] "Study on channel model for frequencies from 0.5 to 100 GHz, Release 14," 3GPP, Sophia Antipolis, France, Rep. TR 38.901, Dec. 2017.
[18] "Enhanced LTE support for aerial vehicles, Release 15," 3GPP, Sophia Antipolis, France, Rep. TR 36.777, Dec. 2017.
[19] S. O. Rice, "Mathematical analysis of random noise," *Bell Syst. Tech. J.*, vol. 23, no. 3, pp. 282–332, Jul. 1944.
[20] J. I. Marcum, *Table of Q-functions*. Santa Monica, CA, USA: Rand Corporation, Jan. 1950.
[21] G. Demange, D. Gale, and M. Sotomayor, "Multi-item auctions," *J. Polit. Econ.*, vol. 94, no. 4, pp. 863–872, Aug. 1986.
[22] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert *W* function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996. [Online]. Available: https://doi.org/10.1007/BF02124750
[23] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, Sep. 1998.
[24] E. Yang and D. Gu, "Multiagent reinforcement learning for multi-robot systems: A survey," Dept. Comput. Sci., Univ. Essex, Colchester, U.K., Rep. CSM-404, 2004.
[25] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Dept. Psychol., King's College, Cambridge Univ., Cambridge, U.K., 1989.
[26] M. Bowling, "Multiagent learning in the presence of agents with limitations," Ph.D. dissertation, Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, May 2003.
[27] W. Uther and M. Veloso, "Adversarial reinforcement learning," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Rep. CMU-CS-03-107, 1997. [Online]. Available: http://www.cs.cmu.edu/afs/cs/user/will/ www/papers/Uther97a.ps
[28] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. Conf. Innov. Appl. Artif. Intell.*, Madison, WI, USA, Jul. 1998, pp. 746–752.
[29] T. Jaakkola, M. I. Jordan, and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms," *Neural Comput.*, vol. 6, no. 6, pp. 1185–1201, Nov. 1994, doi: 10.1162/neco.1994.6.6.1185.
[30] S. Singh, M. Kearns, and Y. Mansour, "Nash convergence of gradient dynamics in general-sum games," in *Proc. Conf. Uncertainty Artif. Intell.*, Stanford, CA, USA, Jun./Jul. 2000, pp. 541–548.

Jingzhi Hu (S'16) received the B.S. degree in electronic engineering from Peking University, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree at the School of Electrical Engineering and Computer Science.

His current research interests include game theory, full-duplex Wi-Fi networks, and wireless network virtualization.

Hongliang Zhang (S'15) received the B.S. degree in electronic engineering from Peking University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree at the School of Electrical Engineering and Computer Science.

His current research interests include device-to-device communications, unmanned aerial vehicle networks, hypergraph theory, and optimization theory.

Dr. Zhang has also served as a TPC member for GlobeCom 2016, ICC 2016, ICCC 2017, ICC 2018, and GlobeCom 2018.

Lingyang Song (S'03–M'06–SM'12) received the Ph.D. degree from the University of York, York, U.K., in 2007.

He was a Research Fellow with the University of Oslo, Oslo, Norway. In 2008, he joined Philips Research, Cambridge,U.K. In 2009, he joined the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, and is currently a Boya Distinguished Professor. His current research interests include wireless communication and networks, signal processing, and machine learning.

Dr. Song was a recipient of the K. M. Stott Prize for Excellent Research from the University of York, the IEEE Leonard G. Abraham Prize in 2016, and the IEEE Asia–Pacific Young Researcher Award in 2012. He has been an IEEE Distinguished Lecturer since 2015.