# Exploiting Deep Reinforcement Learning for Stochastic AoI Minimization in NOMA-aided Multi-UAV-assisted Wireless Networks

Yusi Long*†, Jialin Zhuang*, Shimin Gong*†, Bo Gu*, Jing Xu‡, Jing Deng§,
*School of Intelligent Systems Engineering, Sun Yat-sen University, China
†Guangdong Provincial Key Laboratory of Fire Science and Intelligent Emergency Technology
‡School of Electronic Information and Communications, Huazhong University of Science and Technology, China
§Department of Computer Science, UNC Greensboro, U.S.A.

***Abstract*—In this paper, we consider a multiple unmanned aerial vehicles (UAVs) network, where low-power ground users (GUs) periodically sense the environmental information and need to upload the recent sensing information to a base station (BS). The GUs firstly backscatter their information to the UAVs and then the UAVs transmit the information to the BS by the non-orthogonal multiple access (NOMA) transmissions. Our goal is to minimize the long-term time-averaged age-of-information (AoI) by jointly optimizing the UAV's sensing scheduling, transmission control, and trajectories under the constraints of the GUs' maximum expected AoI. To solve this problem, we propose the Lyapunov-driven hierarchical proximal policy optimization framework, named Lya-HPPO, to decouple the multi-stage AoI minimization problem into several control subproblems. In each control subproblem, the UAVs' sensing scheduling and transmission control are firstly determined by the outer-loop DRL approach, and then the inner-loop optimization module is to optimize the UAVs' trajectories. Simulation results verify that the proposed Lya-HPPO framework converges very fast to a stable value and can make online decisions in real time, while guaranteeing the long-term data buffer and AoI stability.**

***Index Terms*—Unmanned aerial vehicle (UAV), backscatter, non-orthogonal multiple access (NOMA), trajectory planning, Lyapunov optimization, deep reinforcement learning (DRL).**

## I. Introduction

The emerging applications of the future Internet of Things (IoT) have a higher requirement on efficient, reliable, and real-time information sensing, such as in autonomous driving, virtual reality, and so on [1]. As such, the information freshness is very important to make accurate control decision. The age-of-information (AoI) as a novel metric of the information freshness was proposed in [2], which is defined as the time elapsed since the most recent data update event. A lower AoI indicates that the sensing information is more recent and reflects the current state accurately, while a higher AoI implies a larger temporal time delay since the generation of the sensing data, resulting in inconsistencies with the current state of the environment. Delayed delivery of sensing information may potentially result in erroneous control or even catastrophic consequences [3]. However, due to the stochastic nature of wireless environment and limited channel capacity, the timely delivery of the sensing information is challenging.

To address the challenges posed by wireless environment, the unmanned aerial vehicle (UAV) is considered as a promising solution to improve the sensing and transmitting capacity across expansive geographical areas, such as air quality monitoring [4], [5], intelligent transportation system [6], [7], and disaster rescue [8], [9]. Owning to the fast development, high mobility, and ultra-reliability, the UAVs can provide strong line-of-sight links in real-time communication, which can maintain the sensing information fresh at the destination by frequent information sensing and transmitting.

However, in the UAV sensing phase, the channel competition among different GUs will cause channel congestion, resulting in untimely information sensing. Delayed information sensing is detrimental to reducing overall AoI. Hence, the UAVs' sensing scheduling is very vital to keep the GUs' information fresh. Specifically, the GU with small AoI can receive more accurate decision when it frequently upload its information to the UAVs. However, such a sensing scheduling will result in a large overall AoI due to the continuous increase of other unscheduled GUs' AoIs. Another idea is to give priority to the GU with large AoI to upload its information, which can suppress excessive peak AoI. Many research work mainly focus on sensing scheduling in single-hop UAV-assisted network for data collection. However, the effective processing of the GUs' information requires the UAVs' frequent information sensing and transmitting to the BS. Therefore, our considered UAV-assisted network involves two-hop scheduling: $i$) the UAVs' sensing scheduling for the GUs', followed by $(ii)$ the information transmission control from the UAVs to the BS. Once the UAVs complete the information sensing, we hope that the UAVs can transmit the information to the BS with the minimum latency. Hence, we employ NOMA transmission with successive interference cancellatio (SIC) mechanism to minimize information transmission delay and improve communication capacity.

In particular, the UAVs' two-hop scheduling is closely related to the UAVs' trajectories. If the UAV flies closer to a GU with urgent uploading demand, its data can be timely processed, otherwise the GU's data will get older and lose its meaning. When the UAVs are positioned at a considerable distance from each other, they could be scheduled simultaneously to enhance the communication capacity owing to the reduced interference, otherwise they will experience strong interference. The above analysis motivates us to jointly

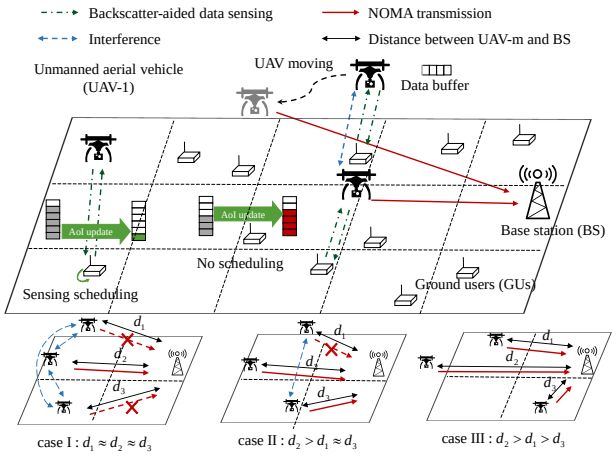



**Fig. 1:** A NOMA-aided multi-UAV-assisted wireless network

optimize the UAVs' sensing scheduling, transmission control and trajectories to reduce the overall AoI.

In this paper, we aim to minimize the long-term time-averaged AoI in a NOMA-aided multi-UAV-assisted wireless network, consisting of a BS, multiple GUs and multiple UAVs, sensing the environmental information from different physical processes. We formulate the AoI minimization as a multi-stage stochastic optimization problem, subject to the UAVs' queue stability constraints and the GUs' maximum AoI requirement. We devise a Lya-HPPO framework to solve the problem. The Lya-HPPO framework first use the Lyapunov optimization framework to convert the multi-stage problem into a series of per-slot subproblems. Then, we use a hierarchical-PPO algorithm to solve each subproblem. Specifically, we adapt the UAVs' sensing scheduling and transmission control policies by the outer-loop DRL algorithm (e.g., PPO algorithm [10]) and use the inner-loop optimization module (e.g., successive convex approximation (SCA) approach) to optimize the UAVs' trajectories. Simulation results are presented to validate that, the Lya-HPPO framework can make online sensing scheduling and transmission control decisions in real time, while guaranteeing the long-term data buffer and AoI stability. Compared to the baseline schemes, the Lya-HPPO framework converges very fast to a stable value.

## II. System Model

### A. Network Architecture

Consider a NOMA-aided multi-UAV-assisted wireless network in Fig. 1, which consists of one BS, $K$ GUs and $M$ UAVs. The UAVs are employed to collect sensing data form $K$ GUs, randomly distributed in an open area, and then transmit them to the BS for information update. The sets of GUs and UAVs are denoted as $\mathcal{K} \triangleq \{1, 2, ..., K\}$ and $\mathcal{M} \triangleq \{1, 2, ..., M\}$, respectively. The UAVs can be deployed as the relays to assist data transmissions from the GUs to the BS. We consider a 3-dimensional coordinate, where the locations of the UAV-$m$ and the GU-$k$ in the $i$-th time slot are denoted by $\boldsymbol{\ell}_m(i) = (x_m(i), y_m(i), z_m(i))$ and $\boldsymbol{q}_k = (x_k, y_k, 0)$, respectively. Without loss of generality, assume that all UAVs fly at a fixed altitude, i.e., $z_m(i) = H$. Denote $\boldsymbol{q}_0 = (x_0, y_0, 0)$ as the BS's location.

We consider a time-slotted multi-access protocol. Each frame is divided into multiple time slots with equal duration. The set of all time slots is denoted by $\mathcal{I} \triangleq \{1, 2, ..., I\}$. Each time slot $\boldsymbol{t}_m(i)$ includes two parts, i.e., sensing duration $t_m^s(i)$ and transmitting duration $t_m^f(i)$. Thus, the feasible time allocation in the $i$-th time slot is determined as follows:

$$t_m^s(i) + t_m^f(i) \leq 1, \quad \forall m \in \mathcal{M}. \tag{1}$$

The length of each time slot is sufficiently small such that each UAV's location is considered as approximately unchanged within each time slot even at the maximum speed. The UAVs must adhere to collision avoidance conditions and maximum speed constraints $V_{\max}$ in each time slot [11] as follows:

$$\|\boldsymbol{\ell}_m(i) - \boldsymbol{\ell}_{m'}(i)\| \geq d_{\min}, \quad \forall m, m' \in \mathcal{M}, \ m \neq m', \tag{2a}$$

$$\|\boldsymbol{\ell}_m(i) - \boldsymbol{\ell}_m(i-1)\| \leq V_{\max}, \quad \forall m \in \mathcal{M}, \tag{2b}$$

where $d_{\min}$ is the minimum distance between any two UAVs to ensure safety. Let the channel vector between the UAV-$m$ and GU-$k$ be modelled as $h_{m,k}(i) = \sqrt{\rho}\|\boldsymbol{\ell}_m(i) - \boldsymbol{q}_k\|^{-1}$, where $\rho$ represents the channel power gain at the reference distance of 1 meter.

### B. Data Sensing and Transmission Model

*1) Dynamic Sensing Scheduling and Transmission Control Decision:* At each time slot, our considered NOMA-aided multi-UAV-assisted wireless network involves a two-hop scheduling process: $(i)$ sensing information from a GU by each UAV in the sensing phase; $(ii)$ the multi-UAV transmission control in the transmitting phase. In the UAV sensing phase, we define a binary variable $\beta_{m,k}(i) \in \{0, 1\}$ to characterize the UAVs' sensing scheduling. We have $\beta_{m,k}(i) = 1$ if the GU-$k$ is scheduled by the UAV-$m$ in the $i$-th time slot, and $\beta_{m,k}(i) = 0$ otherwise. We assume that each UAV can select at most one GU in a time slot, which results in the following sensing scheduling constraint:

$$\sum_{m \in \mathcal{M}} \beta_{m,k}(i) \leq 1, \sum_{k \in \mathcal{K}} \beta_{m,k}(i) \leq 1, \quad \forall m \in \mathcal{M}. \tag{3}$$

Once completing sensing the GUs' information, the UAVs transmit the sensing information to the BS via NOMA transmissions to reduce the communication delay. Specifically, we define a binary variable $\alpha_m(i) \in \{0, 1\}$ to indicate that the UAV-$m$ is scheduled to transmit the cached information to the BS in the $i$-th time slot if $\alpha_m(i) = 1$, otherwise $\alpha_m(i) = 0$. Denote the distance between the UAV-$m$ and the BS in the $i$-th time slot as $d_m(i)$. The UAVs' transmission control policies and trajectory planning are intricately coupled, as demonstrated in the three cases in Fig. 1. Specifically, the UAVs' location changes can lead to various transmission control policies. In case I, the UAVs are positioned at nearly equidistant distances from the BS, i.e., $d_1 \approx d_2 \approx d_3$, resulting in small channel differences. In this case, the UAVs are less inclined to transmit information concurrently due to the potential interference. In contrast, the UAVs are positioned at varying distances from the BS in case III, i.e., $d_2 > d_1 > d_3$. Furthermore, different transmission control policies can modify the UAVs' flying paths. From case I to case III, it can be seen that when all

UAVs opt to transmit information simultaneously to the BS, they tend to navigate to different locations away from the BS to exploit substantial channel variations, thereby achieving higher transmission rates.

*2) Multi-UAV-aided Data Sensing and Transmitting:* Similar to [12], each GU is equipped with the passive backscatter communication module and is capable of transmitting its sensing information to the UAV by backscattering the incident RF signal from the UAV. Given the UAVs' hovering positions, the GUs can be selected to upload their sensing data via backscatter communications. The received signal-to-noise-ratio of the UAV-$m$ from the GU-$k$ can be represented as $\gamma_{m,k}(i) = p_s|\Gamma_0|^2|h_{m,k}(i)|^2|h_{m,k}(i)|^2/\sigma^2$, where $p_s$ is the UAV's transmit power. Thus, the UAV-$m$'s sensing data from the GU-$k$ in the $i$-th time slot can be represented as $o_{m,k}(i) = \beta_{m,k}(i)t_{s,m}(i)\log_2(1+\gamma_{m,k}(i))$. The UAVs collect the sensing information from the GUs and store them into the data buffers, and then transmit the cached data to the BS via NOMA transmissions. The scheduled UAVs are ordered according to their channel conditions. We can decode information first from the UAV with best channel condition. Without abuse of notations, let $\mathcal{M}$ denote the ordered set of UAVs according to their channel conditions with $h_{1,0}(i) \geq \ldots \geq h_{M,0}(i)$. Based on the discussion above, the signal-to-interference-noise-ratio of the UAV-$m$ at the BS is expressed as $\gamma_{m,0}(i) = \frac{\varsigma_{m,0}(i)}{\sum_{m'=m+1}^{M}\varsigma_{m',0}(i)+1}$, where $\varsigma_{m,0}(i) = \alpha_m(i)p_s|h_{m,0}(i)|^2/\sigma^2$. Thus, the transmitting throughput of the UAV-$m$ in the $i$-th time slot can be represented $o_m^r(i) = t_{r,m}(i)\log_2(1+\gamma_{m,0}(i))$. The queue backlog in the $i$-th time slot of the UAV-$m$'s buffer as $Q_m(i)$, which evolves as follows:

$$Q_m(i+1) = \max\left[Q_m(i) - o_m^r(i), 0\right] + o_m^s(i), \quad (4)$$

where $o_m^s(i) = \sum_{k=1\in\mathcal{K}} o_{m,k}(i)$ represents the UAV-$m$'s sensing data. Stability is an important metric to characterize buffer, which requires that the time-averaged arrival rate is smaller than the time-averaged departure rate, namely,

$$\lim_{I\to\infty}\frac{1}{I}\sum_{i\in\mathcal{I}}\mathbb{E}\left[o_m^s(i)\right] \leq \lim_{I\to\infty}\frac{1}{I}\sum_{i\in\mathcal{I}}\mathbb{E}\left[o_m^r(i)\right], \quad (5)$$

where the expectation is taken over the potential randomness of the channels and the scheduling decision.

### C. AoI Dynamics Model

We can use the number of slots in which data wait for processing in the GUs' buffers to indicate AoI and we evaluate the AoI of the GU-$k$ at the end of each time slot. Due to limited channel capacity, the GU-$k$ may not be able to upload all data successfully to the UAV-$m$ within the sensing subslot $t_{s,m}(i)$. At each time slot, the data size that could be generated by the GU-$k$ is denoted by $c_k(i)$. In this case, we denote $P_k(i) \triangleq \frac{s_k(i)}{c_k(i)}$ as the fraction of successfully uploaded data by the GU-$k$, where $s_k(i) \triangleq \sum_{m\in\mathcal{M}} o_{m,k}(i)$ denotes the uploaded data size of the GU-$k$. According to this definition, we update the GU-$k$'s AoI for its partially successfully transmitted information as follows:

$$a_k(i+1) = (1-P_k(i))(a_k(i)+1), \quad \forall k \in \mathcal{K}. \quad (6)$$

Let $a_{\max}$ represent the maximum expected time average age requirement of the GU-$k$, and we require that the expected time average cost is upper bounded as follows:

$$\lim_{I\to\infty}\frac{1}{I}\sum_{i=0}^{I-1}\mathbb{E}\left[a_k(i+1)\right] \leq a_{\max}. \quad (7)$$

The expectation is taken with respect to the random channel and the UAV's sensing scheduling and transmission control.

To reflect the age of each GU's data before uploading to the UAV, we can use a virtual AoI queue $Z_k(i)$ to represent the data timeliness as in [13], which provides a generalized method to approximate a stochastic inequality by using a virtual queue system. Hence, we show a simplified reformulation of the time-averaged constraint as follows:

$$Z_k(i+1) = \max[Z_k(i) - a_{\max}, 0] + a_k(i+1). \quad (8)$$

Please refer to our previous work [14] for detailed proof.

## III. Lyapunov Optimization for AoI Minimization

### A. Long-term time-averaged AoI Minimization

We aim to minimize the long-term time-averaged AoI of all GUs by optimizing the UAVs' sensing scheduling and transmission control $\Phi \triangleq (\beta_{m,k}(i), \alpha_m(i))_{k\in\mathcal{K},m\in\mathcal{M}}$, mobility control $(\boldsymbol{\ell}, \boldsymbol{t}) \triangleq (\boldsymbol{\ell}_m(i), t_{s,m}(i), t_{r,m}(i))_{m\in\mathcal{M}}$. The AoI performance has complicated couplings with the above control variables. For simplicity, we define the time-averaged AoI as follows:

$$\bar{A}(\Phi, \boldsymbol{\ell}, \boldsymbol{t}) = \lim_{I\to\infty}\frac{1}{IK}\mathbb{E}\left[\sum_{i\in\mathcal{I}}\sum_{k\in\mathcal{K}}a_k(i+1)\right]. \quad (9)$$

Till this point, we can formulate the AoI minimization problem as follows:

$$\min_{\Phi,\boldsymbol{\ell},\boldsymbol{t}} \ \bar{A}(\Phi, \boldsymbol{\ell}, \boldsymbol{t}), \quad \text{s.t. } (1)-(8). \quad (10)$$

Problem (10) is challenging to solve due to the following reasons. Firstly, the optimization of the UAVs' sensing scheduling and transmission control policies are combinatorial as it defines a discrete feasible set. Secondly, even with the fixed sensing scheduling and transmission control policies, the UAVs' trajectory planning and time allocation are spatial-temporally coupled in a dynamic program. To overcome these difficulties, we devise a Lya-HPPO framework for problem (10). The overall algorithm sketch is shown in Fig. 2. First, we decompose the multi-stage stochastic AoI minimization problem into a series of per-slot deterministic control subproblems via Lyapunov optimization framework. After the Lyapunov decomposition, the per-slot control subproblem is still hard to solve. Hence, we further devise a hierarchical-PPO structure for the per-slot subproblem, which mainly includes the outer-loop learning module for the UAVs' sensing scheduling and transmission control policies and the inner-loop optimization module for the UAVs' mobility control. Then, we can update the system queue states in the next time slot according to (4) and (8), respectively.

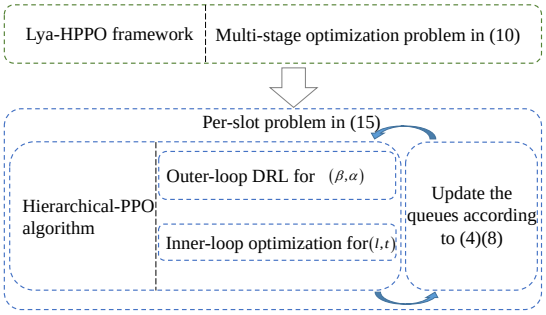



**Fig. 2:** The overall algorithm framework

### *B. Lya-HPPO framework for Per-slot Decomposition*

To solve problem (10), we define a quadratic Lyapunov function $L(\Theta(i))$ as follows:

$$L(\boldsymbol{\Theta}(i)) \triangleq \frac{1}{2}\sum_{k\in\mathcal{K}}(Z_k(i))^2 + \frac{1}{2}\sum_{m\in\mathcal{M}}(Q_m(i))^2, \tag{11}$$

where $\boldsymbol{\Theta}(i) = [Z_k(i), Q_m(i)]$ is the queue backlog vector. Therefore, we can further characterize the queue stability by using the expected change of the Lyapunov function in successive time slots, which is termed as the drift of the Lyapunov function and denoted as follows:

$$\Delta_L(\boldsymbol{\Theta}(i)) \triangleq \mathbb{E}\Big[L(\boldsymbol{\Theta}(i+1)) - L(\boldsymbol{\Theta}(i))|\boldsymbol{\Theta}(i)\Big]. \tag{12}$$

To stabilize the system queue $\boldsymbol{\Theta}(i)$, we aim to minimize the increment of the queue size, i.e., the Lyapunov drift $\Delta_L(\boldsymbol{\Theta}(i))$. Meanwhile, we need to minimize all GUs' AoI values to keep information fresh. Thus, we have the following minimization objective in each time slot:

$$T(\boldsymbol{\Theta}(i)) \triangleq \Delta_L(\boldsymbol{\Theta}(i)) + V\sum_{k\in\mathcal{K}}\mathbb{E}\Big[a_k(i+1)|\boldsymbol{\Theta}(i)\Big], \tag{13}$$

where $V$ is the constant to balance the queue backlog and overall AoI. To this point, we can replace the stochastic objective in (10) by the new minimization target in (13) and focus on the per-slot control problem with the known states of all queues. However, it is still difficult to minimize (13) directly. Instead, we can derive an upper bound to (13) and minimize the upper bound as an approximation. By a few manipulations similar to that in [14], the Lyapunov drift-plus-penalty function is upper bounded as follows:

$$T(\boldsymbol{\Theta}(i)) \leq B + U(\Phi(i), \boldsymbol{\ell}(i), \boldsymbol{t}(i)), \tag{14}$$

where $B = \frac{1}{2}\sum_{k\in\mathcal{K}}((a_k(i)+1)^2 + a_{\max}^2) - \sum_{k\in\mathcal{K}}(Z_k(i)a_{\max} + (Z_k(i)+V)(a_k(i)+1)) + \frac{1}{2}\sum_{m\in\mathcal{M}}\left[(o_m^{r,\max})^2 + (o_m^{s,\max})^2\right]$ is a finite constant and $U(\Phi(i), \boldsymbol{\ell}(i), \boldsymbol{t}(i)) = \sum_{m\in\mathcal{M}} Q_m(i)(o_m^s(i) - o_m^r(i)) - \sum_{k=1}^{K}(V + Z_k(i))P_k(i)(a_k(i)+1)$. Please refer to [14] for the detailed proof of the above upper bound.

For simplicity, we drop the time index in the per-slot control problem (13). Once we observe the queue states at the beginning of the $i$-th time slot, the minimization of $T(\boldsymbol{\Theta}(i))$ in (13) can be approximated by the following problem:

$$\min_{\Phi,\boldsymbol{\ell},\boldsymbol{t}} \; U(\Phi, \boldsymbol{\ell}, \boldsymbol{t}) \;\; \text{s.t. (1)} - \text{(3)}. \tag{15}$$

**Algorithm 1** Lya-HPPO algorithm for UAVs' sensing scheduling and transmission control in the $i$-th time slot

**Initialization:**DNN weight parameters $\theta$, policy network $\pi_{\theta_{\text{old}}}$, $t \leftarrow 0$, $\boldsymbol{a}_k^t = a_k(i-1)$, $\boldsymbol{Q}_m^t = Q_m(i-1)$, $\boldsymbol{\ell}_m^t = \boldsymbol{\ell}_m(i-1)$, $\boldsymbol{H}_m^t = \boldsymbol{H}_m(i-1)$;
**for** Episode $= 1, 2, ...,$ Max **do**
  **repeat**
    Observe the system state $(\boldsymbol{a}^t, \boldsymbol{Q}^t, \boldsymbol{\ell}^t, \boldsymbol{H}^t)$;
    Choose the outer-loop action $\Phi^t$ for joint scheduling;
    Optimize UAVs' mobility control $(\boldsymbol{\ell}^t, \boldsymbol{t}^t)$ in (17);
    Execute the action $\boldsymbol{x}^t \triangleq (\Phi^t, \boldsymbol{\ell}^t, \boldsymbol{t}^t)$;
    Execute the reward $v^t(\boldsymbol{s}^t, \boldsymbol{x}^t)$;
    Buffer the transition $(\boldsymbol{s}^t, \boldsymbol{x}^t, v^t, \boldsymbol{s}^{t+1})$;
    $t \leftarrow t+1$
  **until** $t = T$;
  Take samples from the experience replay buffer;
  Update the DNN parameters by using PPO algorithm.
**end for**

Instead of the stochastic optimization in (10), now we focus on the deterministic subproblem (15), which becomes a mixed-integer problem and still difficult to solve directly. In the following, we devise a hierarchical-PPO algorithm for problem (15), which mainly includes the outer-loop learning for the UAVs' sensing scheduling and transmission control, as well as the inner-loop optimization for the UAVs' mobility control. Thus, in each iteration, the DRL agent first determines the UAVs' sensing scheduling and transmission control based on the past observations of the UAVs' data statues and the GUs' AoI dynamics. Then, the inner-loop optimization of the UAVs' mobility control becomes much easier by using the SCA method. Finally, the BS can execute the joint action $(\Phi, \boldsymbol{\ell}, \boldsymbol{t})$ in the $t$-th step and then update the system states. The agent can accelerate the learning process since the optimization module can provide partial actions.

### *C. Outer-Loop Learning for UAV's Sensing Scheduling and Transmission Control*

The outer-loop DRL approach aims to update the UAVs' sensing scheduling and transmission control policies by continuously interacting with the network environment. We can reformulate the UAVs' sensing scheduling and and transmission control optimization problem into the Markov decision process, which can be characterized by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R})$. The state space $\mathcal{S}$ denotes the set of all system states. In the $t$-th decision epoch, the system's state $\boldsymbol{s}^t \in \mathcal{S}$ includes all GUs' AoI values, denoted as a vector $\boldsymbol{a}^t \triangleq [a_1^t, ..., a_K^t]$, the UAVs' data buffer state $\boldsymbol{Q}^t \triangleq [Q_1^t, ..., Q_M^t]$, the UAVs' position $\boldsymbol{\ell}^t \triangleq [\ell_1^t, ..., \ell_M^t]$, and the channel conditions $\boldsymbol{H}^t \triangleq [h_{1,0}^t, ..., h_{m,K}^t]$; Hence, we can define the system state in each decision epoch as $\boldsymbol{s}^t \triangleq (\boldsymbol{a}^t, \boldsymbol{Q}^t, \boldsymbol{\ell}^t, \boldsymbol{H}^t)$. The action space $\mathcal{A}$ denotes the set of all feasible decisions $\boldsymbol{x}^t$ that satisfies the UAVs' transmission constraints. The reward $\mathcal{R}$ assigns each state-action pair an immediate value, defined as follows:

$$v^t(\boldsymbol{s}^t, \boldsymbol{x}^t) = \sum_{t\in\mathcal{T}} U(\Phi^t, \boldsymbol{\ell}^t, \boldsymbol{t}^t). \tag{16}$$

The PPO algorithm, introduced in [10], offers a balance between learning efficiency, simplicity, and sample efficiency, making it a suitable choice for policy-based policy gradient tasks. We implement the PPO algorithm to adapt the UAVs' sensing scheduling and transmission control in the outer-loop learning, as listed in Algorithm 1.

At the initialization stage, we randomly initialize the deep neutral network (DNN) weight parameters $\theta$ for the policy network. In each learning episode, The BS collects observations $\boldsymbol{s}^t \triangleq (\boldsymbol{a}^t, \boldsymbol{Q}^t, \boldsymbol{\ell}^t, \boldsymbol{H}^t)$ of the system in the $t$-th decision epoch, and then executes an action $\Phi^t$ from the DRL agent according to the old policy network $\pi_{\theta_{\text{old}}}$, as shown in line 5 of Algorithm 1. Given the outer-loop decision $\Phi^t$, the BS needs to optimize the UAVs' mobility control, as shown in problem (17). This corresponds to lines 6 of Algorithm 1. Then, we will execute the decision variables $(\Phi^t, \boldsymbol{\ell}^t, \boldsymbol{t}^t)$ and the correspoding reward, as shown in lines 7 - 9 of Algorithm 1. The DNN training of the PPO algorithm is based on the mini-batch from the experience replay buffer, as shown in lines 12 - 13 of Algorithm 1.

### *D. Inner-loop Optimization for UAV's Mobility Control*

Given the UAVs' sensing scheduling and transmission control policies $\Phi$, the optimization of $(\boldsymbol{\ell}, \boldsymbol{t})$ can be solved by SCA method efficiently. The UAVs' mobility control includes the UAVs' hovering positions and time allocation for the UAVs' sensing, and transmitting phases. For simplicity, given the UAVs' sensing scheduling and transmission control, we introduce $\hat{\boldsymbol{\ell}}_{m,k} = \|\boldsymbol{\ell}_m - \boldsymbol{q}_k\|^2$ and slack variables $(\eta_{m,k}, \phi_{m,k}, \vartheta_{m,0})$ to approximate the the UAVs' mobility control subproblem as follows:

$$\min_{\boldsymbol{\ell},\boldsymbol{t}} \quad \sum_{m\in\mathcal{M}}\sum_{k\in\mathcal{K}} \beta_{m,k}\hat{Q}_{m,k}\eta_{m,k}^2 - \sum_{m\in\mathcal{M}} Q_m^u D(\vartheta_{m,0}) \quad (17a)$$

$$\text{s.t.} \quad \log_2\left(1+\frac{p_s|\Gamma_0|^2\rho^2/\sigma^2}{\phi_{m,k}}\right) \le \frac{\eta_{m,k}^2}{t_{s,m}}, \quad \text{if } \hat{Q}_{m,k}>0, \quad (17b)$$

$$\phi_{m,k} \le E(\hat{\boldsymbol{\ell}}_{m,k}), \quad \text{if } \hat{Q}_{m,k} > 0 \quad (17c)$$

$$\eta_{m,k}^2/t_{s,m} \le F(\hat{\boldsymbol{\ell}}_{m,k}), \quad \text{if } \hat{Q}_{m,k} \le 0 \quad (17d)$$

$$\vartheta_{m,0}^2/t_{r,m} \le H(\hat{\boldsymbol{\ell}}_{m,0}) - \hat{H}(\hat{\boldsymbol{\ell}}_{m,0}), \quad (17e)$$

$$(1) \text{ and } (2), \quad (17f)$$

where $\hat{Q}_{m,k}$, $D(\vartheta_{m,0})$, $E(\hat{\boldsymbol{\ell}}_{m,k})$, $F(\hat{\boldsymbol{\ell}}_{m,k})$ and $\hat{H}(\hat{\boldsymbol{\ell}}_{m,0})$ are linear approximations, detailed as follows:

$$\hat{Q}_{m,k} \triangleq Q_m - (Z_k + V)(a_k + 1)/c_k, \quad (18a)$$

$$D(\vartheta_{m,0}) \triangleq \left(\vartheta_{m,0}^{(\tau)}\right)^2 + 2\vartheta_{m,0}^{(\tau)}\left(\vartheta_{m,0} - \vartheta_{m,0}^{(\tau)}\right), \quad (18b)$$

$$E(\hat{\boldsymbol{\ell}}_{m,k}) \triangleq \left(\hat{\boldsymbol{\ell}}_{m,k}^{(\tau)}\right)^2 + 4\hat{\boldsymbol{\ell}}_{m,k}^{(\tau)}\left(\boldsymbol{\ell}_m^{(\tau)} - \boldsymbol{q}_k\right)^T\left(\boldsymbol{\ell}_m - \boldsymbol{\ell}_m^{(\tau)}\right), \quad (18c)$$

$$F(\hat{\boldsymbol{\ell}}_{m,k}) \triangleq \log_2(1+\gamma_{m,k}^{(\tau)}) - \frac{\gamma_{m,k}^{(\tau)}\left(\hat{\boldsymbol{\ell}}_{m,k}^2 - (\hat{\boldsymbol{\ell}}_{m,k}^{\tau})^2\right)}{\left(\hat{\boldsymbol{\ell}}_{m,k}^{(\tau)}\right)^4(1+\gamma_{m,k}^{(\tau)})}, \quad (18d)$$

$$\hat{H}(\hat{\boldsymbol{\ell}}_{m,0}) \triangleq \log_2\left(1 + \sum\nolimits_{j=m}^{M} \varsigma_{j,0}\right). \quad (18e)$$

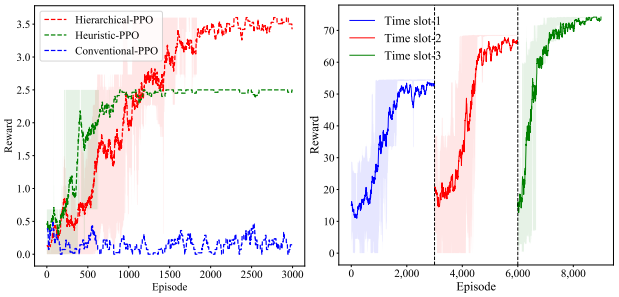
The above analysis reveals that the Lya-HPPO algorithm is to solve the problem (10) following the overall framework in Fig. 2. The purpose is to minimize the overall AoI by optimizing the UAVs' sensing scheduling, transmission control, and trajectory planning. First, the Lyapunov optimization is used to decompose the problem. Then, once the UAVs' sensing scheduling and transmission control are determined by outer-loop DRL, the UAVs' trajectories and time allocation can be optimized by solving problem (17). At the end of each time slot, the UAVs' data queues and the GUs' AoI queues can be updated according to (4) and (8).

## IV. Numerical Results

In this section, we present simulation results to verify the performance gain of the proposed Lya-HPPO framework. The BS's location in meters is given by $(550, 200, 0)$. The GUs are randomly distributed in a rectangular area with the dimension of $500 \times 500$ $m$ in the $(x, y)$-plane with $z = 10$. The default parameter settings are given as follows: $v_{\max} = 25m/s$, $d_{\min} = 30m$, $p_s = 27$ dBm, and $V = 100$.

In Fig. 3(a), we reveals the convergence of the outer-loop DRL in a time slot. We compare the reward performance of the proposed Hierarchical-PPO with the Conventional-PPO and the Heuristic-PPO algorithms in Fig. 3(a). All decision variables $(\beta_{m,k}, \alpha_m, \boldsymbol{\ell}_m, \boldsymbol{t}_m)$ are adapted simultaneously in the Conventional-PPO algorithm. In the Heuristic-PPO algorithm, we employ a heuristic method to replace the sensing scheduling $\beta_{m,k}$ learned in the Hierarchical-PPO algorithm. The heuristic method is that each UAV is connected with its closest GU. It is clear that the Conventional-PPO is unlikely to converge effectively due to a huge action space in the mixed discrete and continuous domain. The Hierarchical-PPO algorithm can reduce the action space in the outer-loop PPO framework and thus achieve a significantly higher reward performance and faster convergence guided by the inner-loop optimization. Compared with the Hierarchical-PPO algorithm, the Heuristic-PPO algorithm significantly reduces the action space by introducing a heuristic design for sensing scheduling $\beta_{m,k}$, which results in faster convergence. However, compared to learning method, heuristic approach obtain a less accurate sensing scheduling $\beta_{m,k}$, which results in a lower reward.

Figure 3(b) further reveals the learning performance of the Lya-HPPO framework in multiple time slots. At the beginning of each time slot, given the outer-loop decision, another action can be estimated by the inner-loop optimization. Then, the UAVs and the GUs can execute the joint action and further update their data buffers and AoIs. As such, the multi-slot learning problem can be transformed into sequential learning processes. The updated data buffers and AoIs are used as the initial states in the next time slot. At the beginning of each time slot, the reward firstly drops significantly due to the change of the UAVs' sensing scheduling and transmission control. However, the reward quickly increases as the UAVs' sensing scheduling and transmission control are adapted by the Hierarchical-PPO algorithm. When finding a stable joint action, each UAV will report their locations to the BS through optimization method. We can observe that the reward can be improved gradually at the end of each time slot, which validates of the Lya-HPPO framework.




**(a)** Convergence of the outer-loop in a time slot. **(b)** Convergence of Lya-HPPO framework in multiple time slots.

**Fig. 3:** The Lya-HPPO framework improves reward performance and learning efficiency.

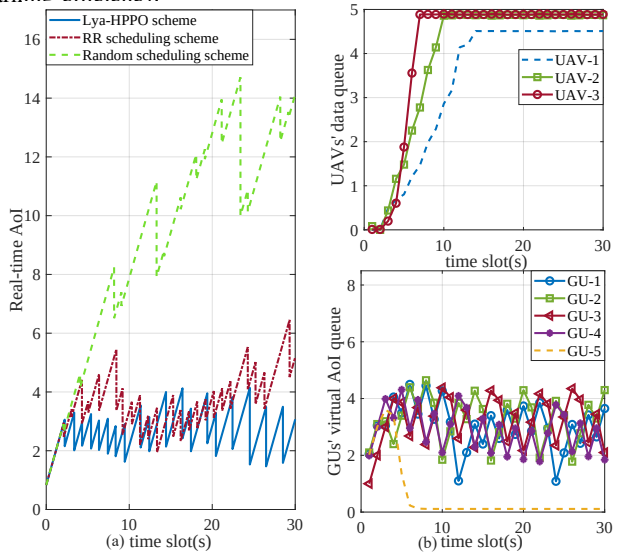



**Fig. 4:** (a) AoI dynamics with different scheduling schemes; (b) Dynamics of UAVs' data queue and GUs' AoI.

Figure 4(a) depicts the GUs' AoI dynamics in different scheduling schemes. We devise a set of baseline schemes for comparison, i.e., the round robin (RR) scheduling scheme and the Random scheduling scheme. The RR scheduling scheme means that the each UAV periodically selects one GU to collect its status-update information. The Random scheduling scheme allows the GUs to randomly access the uplink GU-UAV channels for data uploading. It is observed that the Lya-HPPO scheme can achieve the best real-time AoI and ensure the smooth AoI change of each GU. As the GUs' AoIs and the UAVs' data backlogs are taken into considerations in the Lya-HPPO scheme, it will not yield high AoI fluctuation.

In Fig.4(b), we show the dynamics of the UAVs' data queues and the GUs' virtual AoI queues of the Lya-HPPO algorithm. It is observed that the data of UAVs increases gradually and tends to be stabilized in the subsequent time slots. The reason is that the amount of collected data is relatively low due to the long distance between the UAVs and the GUs at the beginning of the sensing period. In the later time slots, the closer UAV-GU distances can bring higher sensing and transmission rates, which is beneficial for stabilizing the data queues. It is also observed that the virtual AoI queues of GU-1 to GU-4 fluctuate in a smaller range over different time slots, while the GU-5's virtual AoI queue quickly approaches zero. This is because the GU-5 is located in a remote area, and the UAVs tend to serve it more frequent after reaching it, which effectively reduces the GU-5's AoI. The above results confirm that the Lyapunov control helps maintain queues stability, which is critical to achieve stable and efficient system operation.

## V. Conclusions

In this paper, we have investigated a NOMA-aided multi-UAV-assisted wireless network for AoI minimization. We have provided a novel Lyapunov-driven hierarchical PPO (Lya-HPPO) framework to reduce the overall AoI. The Lya-HPPO framework first decomposes the multi-stage problem into several per-slot subproblems via Lyapunov optimization. Then, we use hierarchical-PPO algorithm to solve each per-slot subproblem, including the outer-loop learning and the inner-loop optimization. The Lya-HPPO framework can keep information fresh by flexibly optimizing the UAVs' sensing scheduling, transmission control, and trajectories, while maintaining the queue stability. Numerical results have demonstrated that the Lya-HPPO algorithm can efficiently achieve a faster convergence and reduce the overall AoI.

## References


[1] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[2] A. Kosta, N. Pappas, and V. Angelakis, "Age of information: A new concept, metric, and tool," *Now Foundations and Trends in Netw.*, vol. 12, no. 3, pp. 162–259, Nov. 2017.

[3] L. Cui, Y. Long, D. T. Hoang, and S. Gong, "Hierarchical learning approach for age-of-information minimization in wireless sensor networks," in *proc IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Belfast, Northern Ireland, Jun. 2022, pp. 130–136.

[4] 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, Nov. 2018.

[5] Z. Hu, Z. Bai, Y. Yang, Z. Zheng, K. Bian, and L. Song, "UAV aided aerial-ground IoT for air quality sensing in smart city: Architecture, technologies, and implementation," *IEEE Netw.*, Mar. 2019.

[6] H. Peng and X. Shen, "Multi-agent reinforcement learning based resource management in MEC- and UAV-assisted vehicular networks," *IEEE J. Sel. Area. Commun.*, Jan. 2021.

[7] M. Khabbaz, J. Antoun, and C. Assi, "Modeling and performance analysis of UAV-assisted vehicular networks," *IEEE Trans. Veh. Techn.*, vol. 68, no. 9, pp. 8384–8396, Apr. 2019.

[8] C. Luo, W. Miao, H. Ullah, S. McClean, G. Parr, and G. Min, *Unmanned Aerial Vehicles for Disaster Management*. Springer Singapore, Aug. 2019.

[9] Z. Huang, C. Chen, and M. Pan, "Multi-objective UAV path planning for emergency information collection and transmission," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6993–7009, Aug. 2020.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[11] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Jan. 2018.

[12] P. X. Nguyen, D.-H. Tran, O. Onireti, P. T. Tin, S. Q. Nguyen, S. Chatzinotas, and H. Vincent Poor, "Backscatter-assisted data offloading in OFDMA-based wireless-powered mobile edge computing for IoT networks," *IEEE Internet Things J.*, vol. 8, no. 11, pp. 9233–9243, Jan. 2021.

[13] M. Neely, "Energy optimal control for time-varying wireless networks," *IEEE Trans. Inf. Theory*, vol. 52, no. 7, pp. 2915–2934, Jul. 2006.

[14] Y. Long, W. Zhang, S. Gong, X. Luo, and D. Niyato, "AoI-aware scheduling and trajectory optimization for multi-UAV-assisted wireless networks," in *proc. IEEE GLOBECOM*, Rio de Janeiro, Brazil, Dec. 2022, pp. 2163–2168.