Multi-UAV Navigation and Recharging for Fair and Sustainable Coverage in Wireless Networks

Li Xu
School of Electrical Engineering and Computer Science, Ningbo University
Ningbo, China
lxu3777@gmail.com

Juan Liu
School of Electrical Engineering and Computer Science, Ningbo University
Ningbo, China
eeliujuan@gmail.com

Lingfu Xie
School of Electrical Engineering and Computer Science, Ningbo University
Ningbo, China
xielingfu@nbu.edu.cn

Xiaofan He
School of Electronic Information, Wuhan University
Wuhan, China
xiaofanhe@whu.edu.cn

ABSTRACT

Unmanned aerial vehicles (UAVs) arouse considerable interest in coverage applications such as emergency communication. This paper attempts to address the multi-UAV navigation problem for fair coverage in wireless networks, where a charging station is deployed to recharge the UAVs. It is challenging to control each UAV to provide coverage collaboratively such that each point-of-interest in the target area is covered fairly for a reasonable time duration. The problem of joint multi-UAV navigation and recharging is formulated into a Markov decision process with the objective to maximize the fair coverage score per unit of energy with recharging reward. The state-of-the-art deep reinforcement learning method, coined by us as PPO-UNC (proximal strategy optimization for multi- UAV navigation and recharging), is employed to find the solution efficiently. Simulation results show the superiority of the proposed PPO-UNC strategy as compared to two baseline policies.

CCS CONCEPTS

• **Computing methodologies** → **Planning and scheduling**; **Multi-agent systems**; **Motion path planning**.

KEYWORDS

UAV, Coverage, Fairness index, Deep Reinforcement Learning, PPO

ACM Reference Format:
Li Xu, Juan Liu, Lingfu Xie, and Xiaofan He. 2021. Multi-UAV Navigation and Recharging for Fair and Sustainable Coverage in Wireless Networks. In *2021 3rd International Conference on Algorithms, Machine Learning and Signal Processing (AMLSP 2021), November 26–28, 2021, Sanya, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3503047.3503129

AMLSP 2021, November 26–28, 2021, Sanya, China

ACM ISBN 978-1-4503-8586-2/21/11…$15.00
https://doi.org/10.1145/3503047.3503129

1 INTRODUCTION

Recent years have seen a surge of interest from the research communities in unmanned aerial vehicles (UAVs) aided wireless communications. UAVs can be leveraged for various purposes, due primarily to their high mobility and deployment flexibility. For example, they can act as mobile base stations to extend the coverage of a network to users in remote areas or at natural disaster scenes, to collect fresh information from inaccessible environments , or to augment a data link via line-of-sight (LOS) transmissions. Our paper tackles the fair and sustainable coverage issue in UAV-aided networks.

Refs. [1–4] employed hovering or stationary maximize the coverage area by optimizing the height of the UAV while [2] did so by jointly optimizing the UAV's height and beam angle. [3] optimized the 3D locations of multiple UAVs to maximize their aggregate coverage areas, subject to a signal-to-noise ratio (SNR) threshold for all users. By contrast, [4] considered dividing an area into cells and associating each with a UAV to minimize the overall transmit power of the UAV; for each cell, a minimum power was needed to guarantee a capacity threshold for its users.

Besides, there are a spectrum of works, e.g., [5–7], on employing flying UAVs for network coverage. [5] was an attempt to optimize the trajectories of multiple UAVs so as to maximize the energy efficiency. [6] investigated the optimal UAV trajectory to address the tension between the in-flight energy consumption of the UAV and the aggregate downlink throughput of a ground terminal. By contrast, [7] considered two UAVs competing to serve ground users and the problem of optimal beaconing in drone small cells networks was resolved from the game theoretic perspective.

Since UAVs are energy-constrained, deploying charging stations is a way to prolong the continuous operation time of the UAVs [5][8][9]. A central question here is how to recharge the UAVs appropriately or optimally such that certain design goals (e.g., the seamless and constant coverage of an area [5], the longest network lifetime [8], etc.) could be accomplished. In [9], the UAV's trajectory and the power of the charging station were jointly optimized to maximize the overall power transmission efficiency. Our study also considers implementing a charging station for improving the multi-UAV coverage performance.

This paper is an effort to address a multi-UAV navigation problem for effective coverage in a wireless network with a charging station

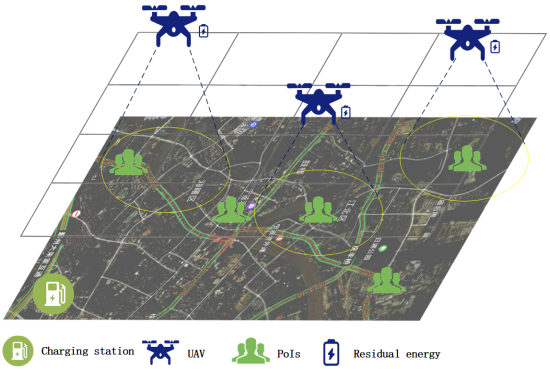

Fig. 1: The low-altitude UAV coverage scenario where multiple UAVs provide effective coverage for ground users.

deployed. Multiple UAVs are navigated to provide fair and energy-efficient aerial coverage for ground users. When necessary (e.g., their energy levels are relatively low), they can fly directly to the charging station for energy replenishment. The point is how to control each UAV to provide coverage collaboratively while getting recharged regularly such that each point-of-interest in the target area is covered for a reasonable time period.

The main contributions of this paper are summarized as follows: 1) The problem of joint multi-UAV navigation and recharging is formulated into a Markov decision process (MDP) with the objective to provide fair and energy-efficient coverage; 2) In the framework of deep reinforcement learning (DRL), proximal strategy optimization (PPO) [10] is applied to seek the multi-UAV navigation and recharging solution efficiently, referred to as PPO-UNC; 3) Simulation results show that the proposed PPO-UNC method achieves a higher fair coverage performance than the baseline greedy recharging and random control policies.

The rest of this paper is organized as follows. Section II introduces the system model. The multi-UAV navigation and recharging problem is formulated as an MDP in Section III. Section IV proposes the PPO-UNC algorithm to find the solution and Section V presents the simulation results, respectively. Concluding remarks are given in Section VI.

2 SYSTEM MODEL

2.1 Network Description

As depicted in Fig. 1, we consider a typical low-altitude aerial platform (LAP), where N multi-rotor UAVs serve as aerial base stations to provide long-term effective coverage for ground users in the target area. Following the coverage model in [11], the whole target region $\mathcal{R}$ is divided into K cells appropriately, depending on the distribution of the ground users. The center of each cell, called the Point-of-Interest (PoI), needs to be covered by the drones for a reasonable period of time. The coordinate of PoI k is denoted by $\mathbf{w}_k = (X_k, Y_k, 0)$ $(k = 1, \cdots, K)$.

The system time is assumed to be slotted and the aerial coverage task of the UAVs lasts T time slots, where T is a sufficiently large integer. The length of each time slot is equal to t_s seconds. Considering that the number of drones is limited, it is impossible for the drones to cover all the PoIs seamlessly in each slot. Hence, the drones are navigated to move around such that each PoI is effectively covered for a reasonable number of time slots. To enable a long-term communication coverage, a charging station is deployed in the target area at the location $\mathbf{w}_0 = (X_0, Y_0, 0)$. The drones can fly to the charging station for energy replenishment when they reach a low battery level. Each UAV should keep a safe distance from the other drones to avoid collision. Let d_{min} be the minimum distance between any two UAVs.

Each UAV is equipped with a battery of finite capacity E_{max}. At the very beginning, each UAV carries an amount of E_{max} (Joules) energy, takes off at a random origin, and flies at a certain altitude. At the beginning of each time slot t, each UAV i hovers at its current location $\mathbf{v}_i^t = (x_i^t, y_i^t, h_i)$ to provide coverage, or flies horizontally in a direction $\theta_i^t \in (0, 2\pi]$ for a distance d_i^t to the next cell or the charging station, or gets recharged at the charging station, according to the UAV deployment decision. When hovering or flying, each UAV i consumes E_i^t (Joules) energy at time slot t. It needs to be recharged when its residual energy is at low level. The residual energy of UAV i at the end of each time slot t is denoted by $e_i^t \in [0, E_{max}]$.

The recharging demand of UAV i is represented by $q_i^t \in \{0, 1\}$. Specifically, $q_i^t = 1$ means that UAV i needs to get recharged, and otherwise $q_i^t = 0$. Suppose that at most one UAV can get the recharging service at each time slot. Let $u_t \in \{0, 1, \cdots, N\}$ denote the recharging state at time slot t. $u_t = 0$ means that the charging station is idle and $u_t = i$ means that UAV i is getting recharged.

2.2 Air-to-Ground Channel Model and Communication Links

Similar to the model in [5], the ground users can access the UAVs via orthogonal frequency division multiplex access (OFDMA), and the interference among UAV-user links is neglected. It is also assumed that the UAVs and the charging station communicate with each other via line-of-sight (LoS) links using different frequency bands. We consider the scenario where the charging station or a cloud periodically collects the state (e.g., the locations, energy usages, etc) of the UAV network via gateways.

To analyze the UAVs' energy consumption on coverage, we consider the UAV-user downlink communications. Considering the air-to-ground (ATG) channel propagation model [12], the LoS and non-line-of-sight (NLoS) links between the UAVs and ground users occur with separate probabilities, respectively. For the ATG channel between UAV i located at $\mathbf{v}_i = (x_i, y_i, h_i)$ and a user located at PoI k, the path loss can be expressed as

$$PL_{i,k}^{LoS} = \left(\frac{4\pi f_c D_{i,k}}{c}\right)^2 \xi^{LoS}, \tag{1}$$

$$PL_{i,k}^{NLoS} = \left(\frac{4\pi f_c D_{i,k}}{c}\right)^2 \xi^{NLoS}, \tag{2}$$

where f_c is the carrier frequency, c is the speed of light, $D_{i,k} = ||\mathbf{v}_i - \mathbf{w}_k|| = \sqrt{(x_i - X_k)^2 + (y_i - Y_k)^2 + h_i^2}$ is the distance between UAV i and PoI k, and ξ^{LoS} and ξ^{NLoS} denote the excessive path losses for LoS and NLoS links, respectively.

From the channel modeling in [12], the LoS probability can be given by

$$p_{i,k}^{LoS} = \frac{1}{1 + \alpha \exp\left[-\beta\left(\phi_{i,k} - \alpha\right)\right]} \tag{3}$$

where $\phi_{i,k} = \frac{180}{\pi} \arcsin(\frac{h_i}{d_{i,k}})$ is the elevation angle between UAV i and PoI k, and α and β are constant parameters depending on the propagation environment. Hence, the average path loss can be obtained as

$$\overline{PL}_{i,k} = p_{i,k}^{LoS} PL_{i,k}^{LoS} + (1 - p_{i,k}^{LoS}) PL_{i,k}^{NLoS}, \tag{4}$$

where $1 - p_{i,k}^{LoS}$ is the probability of NLoS. The average path loss is an implicit function of the UAV-user distance $d_{i,k}$ and can be expressed as $\overline{PL}_{i,k}(d_{i,k})$.

Each UAV has a certain communication range R. Since the UAVs can fly at identical or different altitudes, their coverage ranges might be different, and less than the communication range R. When PoI k is covered by UAV i, the users located at this PoI can get satisfactory communication services. That is, the downlink data rate per channel use is equal to or greater than the threshold r_{th}, i.e.,

$$r_{i,k} = \log_2\left(1 + \frac{P_i^{tr}}{\overline{PL}_{i,k}(d_{i,k})\sigma^2}\right) \geq r_{th}, \tag{5}$$

where P_i^{tr} denotes the transmit power of UAV i, and $\overline{PL}_{i,k}$ reaches the largest value when the users located at the cell edge with $d_{i,k} = R$.

Accordingly, the UAV's transmit power satisfies $P_i^{tr} \geq \overline{PL}_{i,k}(R)\sigma^2(2^{r_{th}} - 1)$.

2.3 UAV Power Model

According to the UAV power model in [13][5], the power of a multi-rotor UAV consists of three parts: the induced power P_i, profile power P_p and parasitic power P_{par}. The induced power produces thrust by pushing air downward and can be expressed as

$$P_i(F, V_{vert}) = c_1 F\left[\frac{V_{vert}}{2} + \sqrt{\left(\frac{V_{vert}}{2}\right)^2 + \frac{F}{c_2^2}}\right], \tag{6}$$

where F is the thrust and V_{vert} is the vertical speed of the UAV. The profile power overcomes the rotational drag encountered by rotating blades and can be evaluated as

$$P_p(F, V_{hor}) = c_3 F^{3/2} + c_4(V_{hor}\cos\omega)^2 F^{1/2}, \tag{7}$$

where V_{hor} is the horizontal speed of the UAV, and ω is the angle of attack of a propeller disk when V_{hor} is not zero. The parasitic power resists the fuselage drag when there is a relative translational motion between the UAV and wind, which can be expressed as

$$P_{par}(V_{hor}) = c_5 V_{hor}^3. \tag{8}$$

In (6)-(8), c_1, c_2, c_3, c_4, and c_5 are the constant parameters related to physical properties of the UAV.

With sufficient energy, each UAV flies horizontally at a constant speed V_t from one PoI to another one, or hovers over one PoI to provide communication services for the covered users. According to the UAV power model, the power required for flying is given by

$$\begin{aligned} P_f(V_t) &= P_i(F, 0) + P_p(F, V_t) + P_{par}(V_t) \\ &= (\tfrac{c_1}{c_2} + c_3)F^{3/2} + c_4(V_t cos\alpha)^2 F^{1/2} + c_5 V_t^3 \\ &\approx (\tfrac{c_1}{c_2} + c_3)F^{3/2} + c_5 V_t^3, \end{aligned} \tag{9}$$

where the second term can be omitted since it contributes very little [13]. Substituting $V_t = 0$ into (9), the hovering power denoted by P_h can be expressed as $P_h = P_i(F, 0) + P_p(F, 0) = \left(\frac{c_1}{c_2} + c_3\right)(mg)^{3/2}$, where mg is the UAV's gravity. When the residual energy is insufficient, each UAV needs to fly to the charging station for energy supplement. It is recharged when flying into a certain short range from the charging station, like a small distance d_c. Suppose that each UAV is charged at a rate of η Joule/s by some wireless power transfer technology.

3 PROBLEM FORMULATION

To provide fair and sustainable coverage, we study the joint UAV navigation and recharging problem, and formulate it as a Markov decision process (MDP), as described below.

3.1 Coverage Fairness

If PoI k falls within the communication range of any UAV in time slot t, we say it is covered at this time slot. The coverage indicator can be expressed as

$$\psi_k^t = \begin{cases} 1, \text{if } \mathbf{v}_i^t - \mathbf{w}_k \leq R, i \in \{1, \cdots, N\}, \\ 0, \text{otherwise.} \end{cases} \tag{10}$$

Till time slot t, the coverage score of PoI k is

$$c_k^t = \frac{\sum_{j=1}^{t} \psi_k^j}{T}. \tag{11}$$

To measure the coverage fairness among all the PoIs, we use the widely used fairness metric, Jain's fairness index [11], which is defined as

$$f_t = \frac{\left(\sum_{k=1}^{K} c_k^t\right)^2}{K\sum_{k=1}^{K}\left(c_k^t\right)^2}. \tag{12}$$

It can be easily seen that the higher the fairness index, the fairer the coverage service provided by the drones.

3.2 MDP Problem

Then, we define the system state and action spaces, state update, and cost function in the MDP formulation.

- State and action spaces

The system state at time slot t, denoted by $\mathbf{s}_t$, consists of the following components: 1) $c_k^t \in \mathcal{C} = [0, 1]$: the coverage score of each PoI k in time slot t; 2) $\mathbf{v}_i^t = (x_i^t, y_i^t, h_i)$ ($(x_i^t, y_i^t) \in \mathcal{R}$): the position of UAV i in time slot t; 3) $e_i^t \in \mathcal{E} = [0, E_{max}]$: the remaining energy of UAV i in time slot t; 4) $u_t \in \mathcal{N}^+ = \{0, 1, ..., N\}$: the recharging state in time slot t. Hence, the system state can be expressed as a vector $\mathbf{s}_t = [c_1^t, ..., c_K^t, e_1^t, ..., e_N^t, \mathbf{v}_1^t, ..., \mathbf{v}_N^t, u_t] \in \mathcal{S}$, where $\mathcal{S} = \mathcal{C}^K \times \mathcal{E}^N \times \mathcal{R}^N \times \mathcal{N}^+$ is the state space.

Each UAV takes sequential actions in the aerial coverage or recharging phrase. In the coverage phrase, each UAV i hovers to provide coverage or flies to the next target cell. When necessary, it flies directly to the charging station for energy replenishment and stays in the recharging phrase till its battery is fully charged. Accordingly, the action of each UAV i can be expressed as $\mathbf{a}_i^t = [d_i^t, \theta_i^t, q_i^t]$, where $d_i^t \in \mathcal{D} = [0, d_{max}]$ is the flight distance with d_{max} denoting the UAVs' maximum flight distance, $\theta_i^t \in \mathcal{L} = (0, 2\pi]$ is the flight direction, and $q_i^t \in \mathcal{Q} = \{0, 1\}$ is the recharging demand. Similarly, the system action can be represented by a vector $\mathbf{a}_t = [\mathbf{a}_1^t, \cdots, \mathbf{a}_N^t]$ in the action space $\mathcal{A} = (\mathcal{D} \times \mathcal{L} \times \mathcal{Q})^N$.

- State update process

a) Coverage score: From (11), the coverage score of PoI k is updated as

$$c_k^{t+1} = \begin{cases} c_k^t + \frac{1}{T}, & \text{if } D_{i,k}^t \le R \,\&\, u_t \ne i, i \in \{1, \cdots, N\}, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where $D_{i,k}^t$ is the distance between UAV i and PoI k at time slot t.

b) UAV location: Each UAV flies along the direction θ_i^t with distance d_i^t when providing coverage, and flies to the charging station directly at low energy level. The location of each UAV can be expressed as

$$\mathbf{v}_i^{t+1} = \begin{cases} \mathbf{v}_i^t + (d_i^t \cos\theta_i^t, d_i^t \sin\theta_i^t, 0), & \text{if } q_i^t = 0, \\ \mathbf{w}_0 + (0, 0, h_i), & \text{elif } u_t = 0 \,\&\, q_i^t = 1, \\ \mathbf{v}_i^t, & \text{otherwise.} \end{cases} \tag{14}$$

c) Residual energy: The residual energy of each UAV i is updated as

$$e_i^{t+1} = \min\{(e_i^t - E_i^t + G_i^t)^+, E_{max}\}, \tag{15}$$

where $(x)^+ = \max\{x, 0\}$, and G_i^t denotes the amount of energy recharged at time slot t. According to the system model in Section II, each UAV consumes a certain amount of energy on communicating, hovering and flying, and its energy consumption can be expressed as

$$E_i^{t+1} = \begin{cases} P_f(V_t)t_s, & \text{if } d_i^t > 0 \,\&\, u_t = 0, \\ (P_h + P_i^{tr})t_s & \text{elif } d_i^t = 0 \,\&\, u_t = 0, \\ P_h t_s, & \text{otherwise,} \end{cases} \tag{16}$$

where $V_t = \frac{d_i^t}{t_s}$ is the UAV's flight speed in time slot t. The amount of energy transferred can be expressed as

$$G_i^{t+1} = \begin{cases} \eta t_s, & \text{if } u_t = i \,\&\, d_{i,0}^t \le d_0, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

d) Recharging state: Any UAV i can get the recharging service only when the charging station located within a small distance d_0 is not occupied, i.e., $u_t = 0$. It still stays in the recharging state if its battery is not full. Otherwise, the charging station becomes idle. Accordingly, the recharging state is updated as

$$u_{t+1} = \begin{cases} i, & \text{if } u_t = 0 \,\&\, d_{i,0}^t \le d_0, \\ u_t, & \text{elif } u_t = i \,\&\, e_i^t < E_{max}, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

- Reward

The UAVs are encouraged to provide fair and sustainable coverage using less energy and to get recharging services when their batteries are at low levels. To this end, given the state-action pair $(\mathbf{s}_t, \mathbf{a}_t)$, the reward at time slot t is defined as:

$$r_t = \rho \frac{f_t \cdot \sum_{k=1}^K \psi_k^t}{\sum_{i=1}^N E_i^t} + \sum_{i=1}^N q_i^t \cdot g_i^t(e_i^t), \tag{19}$$

where ρ is the weighting factor, $g_i^t(e_i^t)$ is a function of the residual energy of UAV i, which is set positive to push UAV i to get recharged when its battery level is relatively low. In (19), the first term can be considered as the energy efficiency, i.e., how much fair coverage gain per unit of energy is obtained, and the second term means the UAVs' recharging reward at time slot t. In simulations, we adopt a quadratic polynomial as the recharging reward of UAV i, i.e., $g_i^t(e_i^t) = \frac{e_i^t}{E_{max}}(1 - \frac{e_i^t}{E_{max}})$.

In the sequel, we attempt to find the optimal policy π^* to maximize the long-term return as follows:

$$C_\pi = \mathbb{E}_\pi \left[\sum_{t=1}^T \gamma^t r_t(\mathbf{s}_t, \mathbf{a}_t) | \mathbf{s}_1 \right], \tag{20}$$

where $\mathbb{E}_\pi$ is the expectation following policy π, $\gamma \in [0, 1]$ is the discount factor, and $r_t(\mathbf{s}_t, \mathbf{a}_t)$ is the immediate reward when action $\mathbf{a}_t$ is taken at state $\mathbf{s}_t$.

4 ALGORITHM DESIGN

Notice that the proposed MDP problem has a relatively large state space and action space which contains discrete and continuous actions. To solve this problem, we select the proximal strategy optimization (PPO) method as the starting point of our algorithm design.

Like the other policy gradient algorithms, PPO estimates the policy using the Monte Carlo (MC) method. It collects samples of the policy loss $J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)}[r(\tau)]$ with $r(\tau) = \sum_t r_t(\mathbf{s}_t, \mathbf{a}_t)$ and its gradient

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta(\tau)} \left[\left(\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t | \mathbf{s}_t) \right) r(\tau) \right], \tag{21}$$

which are backpropagated through the neural network to update the hyperparameters θ. In the actor-critic architecture, the advantage function $A_t(\mathbf{s}_t, \mathbf{a}_t) = Q^\pi(\mathbf{s}_t, \mathbf{a}_t) - V^\pi(\mathbf{s}_t)$ is used to measure how good an action is compared to the other actions in the same state, where $Q^\pi(\mathbf{s}_t, \mathbf{a}_t)$ and $V^\pi(\mathbf{s}_t)$ are the state-action value function and value function, respectively.

PPO-UNC has one critic network, and two actor networks. The critic network is used to predict the value function A_t from the reward values in the collected samples. One actor network is to generate the strategy π_θ that we want to optimize. The other one uses importance sampling to obtain the expectation of samples collected under an old policy $\pi_{\theta_{old}}$ to refine the new policy π_θ.

Algorithm 1 PPO-UNC

Input: initialize policy parameters θ and θ_{old}, value function parameters ϕ, the learning rates for actor network α_A and critic network α_C;
for episode :=1, 2, $\cdots$ **do**
  Initialize the residual energy of each UAV $e_i^1 = E_{max}$ and its position $\mathbf{v}_i^t$, the coverage rate of each PoI c_k^t, the recharging state u_t, and the system state $\mathbf{s}_1$.
  for epoch t :=1, $\cdots$, T **do**
    Select action $\mathbf{a}_t = \pi(\mathbf{s}_t; \theta)$;
    Reshape inappropriate action $\mathbf{a}_t$, e.g., any UAV may fly beyond the area, or be out of battery, or collide with the others, and execute $\mathbf{a}_t$ in the environment;
    Calculate the reward r_t by (19), and obtain the updated system state $\mathbf{s}_{t+1}$ by (13)-(18);
    Compute the advantage estimate $\hat{A}_t$ based on the current value function $V_\phi(\mathbf{s}_t)$;
    Collect sample $\{\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}\}$ and add it to the fixed-length trajectory τ;
  end for
  Update the policy π_θ by maximizing the surrogate objective $L^{clip}(\theta)$ in (22) on the collected trajectories;
  Update the parameter ϕ by minimizing the objective $L(\phi) = \hat{\mathbb{E}}_t(V_\phi(\mathbf{s}_t) - \hat{r}_t)^2$;
  Update the parameter θ_{old} every l_2 steps;
end for
Output: the coverage score of each PoI c_k^T, the fairness index f_T, and the long-term return r_T.

The samples collected from $\pi_{\theta_{old}}$ can be trained multiple times to improve sample efficiency. In the training process, the two policies will diverge after a while, and therefore the old policy is updated every l_2 steps to match the new policy.

Let $p_t(\theta) = \frac{\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}{\pi_{\theta_{old}}(\mathbf{a}_t|\mathbf{s}_t)}$ be the probability ratio between the new policy $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ and the old policy $\pi_{\theta_{old}}(\mathbf{a}_t|\mathbf{s}_t)$. PPO-UNC aims to maximize the clipped surrogate objective function as follows:

$$L^{clip}(\theta) = \hat{\mathbb{E}}_{\tau\sim\pi_\theta(\tau)}\left[\sum_{t=1}^{l_1} \min(p_t(\theta)\hat{A}_t, \text{clip}(p_t(\theta), \epsilon)\hat{A}_t)\right], \quad (22)$$

where l_1 is the length of trajectory τ, $\text{clip}(p_t(\theta), \epsilon)$ clips the probability ratio into range $[1-\epsilon, 1+\epsilon]$, $\hat{A}_t$ and $\hat{\mathbb{E}}_t$ are the empirical estimates of the advantage function and expectation, respectively. Here, ϵ is a hyperparameter, like $\epsilon = 0.2$, and the second term in (22) constrains the policy update in a small range. The details of PPO-UNC are presented in Algorithm 1.

5 SIMULATION RESULTS

We evaluate the performance of the proposed algorithm by simulations. The target area is about $400m \times 400m$, and ground users are located in $K = 16$ PoIs. The UAVs fly and hover at an altitude 50m. Some important simulation parameters are listed in Table 1. The other parameters are as follows: $c_1 = 0.8554$, $c_2 = 0.3051$, $c_3 = 0.3177$, $c_4 = 0$, $c_5 = 0.00002$, $d_{max} = d_0 = 20$, $\rho = 1000$, $\gamma = 0.99$, $\alpha_A = 0.000001$, $\alpha_C = 0.00001$, $l_1 = 2000$, $l_2 = 20$, and batch size $B_s = 2000$. The algorithm is trained in 50,000 episodes, each of which has 2,000 epochs.

Table 1: Some simulation parameters.

Parameter	Value
The number of UAVs N	3
The number of PoIs K	16
The number of slots T	2000
Flight altitude h_i	50m
The battery capacity of each UAV E_{max}	2e5J
Recharging rate η	1e5J/s
The UAVs' maximum flight distance d_{max}	200m
The weight of each UAV mg	50N
The thrust of each UAV F	60N
The transmit power of UAV i P_i^{tr}	1.5w
The white Gaussian noise power σ^2	-100dBm

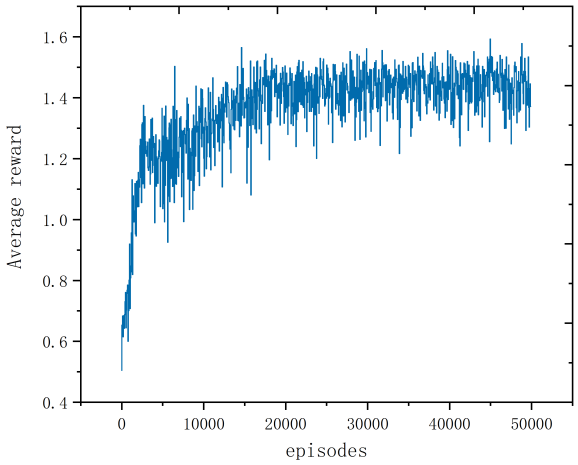

Fig. 2: The average reward accumulated during training

To highlight the advantage of the PPO-based method, we compare it to the two baseline policies: random control and greedy recharging. By random control, the UAVs take actions randomly from the action space at each time slot, including flight distance $d_i^t \in [0, 1]$, flight direction $\theta_i^t \in (0, 2\pi]$, and recharging demand $q_i^t \in \{0, 1\}$. With greedy recharging, each UAV flies to provide coverage using the PPO method and goes to the charging station for energy supplement when its energy level is below a fixed threshold 4KJ.

Fig. 2 shows the average reward accumulated over time during training. We can see that the average reward increases monotonously with the increasing number of episodes. As it reaches 17,500 episodes, the average reward growth slows down and becomes stable. The reason lies in the fact that the PoIs have not been fairly covered or the low-energy UAVs have no chance for energy recharging due to bad navigation policies at the beginning of the task. After a lot of training, the UAVs learn to take good actions to provide lasting and

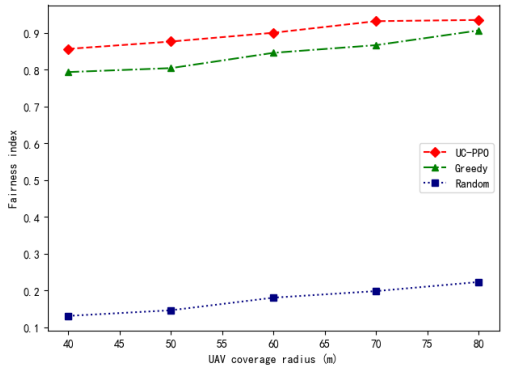

Fig. 3: Fairness index versus coverage radius

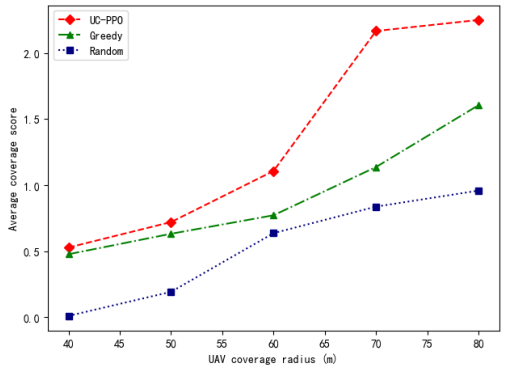

Fig. 4: Coverage score versus coverage radius

fair coverage for the PoIs while maintaining their energy levels by recharging.

In Fig. 3 and Fig. 4, we show the fair coverage performance in terms of fairness index and coverage score, respectively, for different coverage radii of the UAVs. When the coverage radius is small, each UAV can only cover one PoI at each time slot. With the increase of the coverage radius, the UAVs can cover multiple PoIs simultaneously. Hence, more PoIs can be covered within a reasonable period of time, and the system fairness index increases when each of the three algorithms is performed. Due to the same reason, the system coverage score also increases significantly when the coverage radius of each UAV is enlarged, and each PoI is covered for more times. By efficient UAV navigation using the PPO-based methods, both PPO-UNC and greedy recharging policies have much higher fair coverage performances than the random control policy, as shown in Fig. 3 and Fig. 4. Similarly, by joint UAV navigation and recharging, the proposed PPO-UNC algorithm outperforms the greedy recharging algorithm, since intelligent recharging enables the multiple UAVs to provide more efficient and fair coverage.

6 CONCLUSIONS

This paper studied the multi-UAV navigation and recharging problem for fair and sustainable coverage in wireless networks with a charging station. To solve this problem, we proposed an efficient PPO-based strategy, called PPO-UNC, to control multi-UAV navigation and recharging in the DRL framework. By learning based on the collected samples from the environment, PPO-UNC makes sequential decisions on the UAVs' flight, coverage and recharging actions such that each point-of-interest in the target area is covered fairly for a reasonable period of time. By simulations, we showed that PPO-UNC improves the coverage performance significantly compared with two commonly-used baseline policies: random control and greedy recharging. It achieves 0.6 times and 3.9 times higher fairness in terms of fairness index than the greedy and random policies, respectively.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under grant 61971249, the Natural Science Foundation of Zhejiang Province of China under grant LY19F010003.

REFERENCES

[1] B. Li, C. Chen, R. Zhang, H. Jiang, and X. Guo, "The energy-efficient UAV-based bs coverage in air-to-ground communications," in *Proc. IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Sheffield, UK, Jul. 2018, pp. 578–581.

[2] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, Dec. 2014.

[3] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage," *IEEE Communications Letters*, vol. 20, no. 8, pp. 1647–1650, Aug. 2016.

[4] ——, "Optimal transport theory for power-efficient deployment of unmanned aerial vehicles," in *Proc. IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[5] X. Li, H. Yao, J. Wang, X. Xu, C. Jiang, and L. Hanzo, "A near-optimal UAV-aided radio coverage strategy for dense urban areas," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9098–9109, Sep. 2019.

[6] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[7] S. Koulali, E. Sabir, T. Taleb, and M. Azizi, "A green strategic activity scheduling for uav networks: A sub-modular game perspective," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 58–64, May 2016.

[8] A. Trotta, M. D. Felice, F. Montori, K. R. Chowdhury, and L. Bononi, "Joint coverage, connectivity, and charging strategies for distributed UAV networks," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 883–900, Aug. 2018.

[9] W. Chen, S. Zhao, Q. Shi, and R. Zhang, "Resonant beam charging-powered UAV-assisted sensing data collection," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 1086–1090, Jan. 2020.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347 [cs]*, Jul. 2017.

[11] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, Sep. 2018.

[12] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. IEEE Global Communications Conference*, Austin, TX, USA, Dec. 2014, pp. 2898–2904.

[13] Z. Liu, R. Sengupta, and A. Kurzhanskiy, "A power consumption model for multi-rotor small unmanned aircraft systems," in *Proc. International Conference on Unmanned Aircraft Systems (ICUAS)*, Miami, FL, USA, Jun. 2017, pp. 310–315.