Actor-Critic Learning Based Energy Optimization for UAV Access-and-Backhaul Networks

Yaxiong Yuan (yaxiong.yuan@uni.lu)
Universite du Luxembourg https://orcid.org/0000-0002-4118-039X
Lei Lei
Universite du Luxembourg
Thang X. Vu
Universite du Luxembourg
Symeon Chatzinotas
Universite du Luxembourg
Sumei Sun
Institute for Infocomm Research, Agency for Science, Technology and Research
Björn Ottersten
Universite du Luxembourg

Research

Keywords: UAV, deep reinforcement learning, user scheduling, backhaul power allocation, energy optimization, actor-critic

Posted Date: October 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-95073/v1

Version of Record: A version of this preprint was published at EURASIP Journal on Wireless Communications and Networking on April 7th, 2021. See the published version at https://doi.org/10.1186/s13638-021-01960-0.

RESEARCH

Actor-Critic Learning based Energy Optimization for UAV Access-and-Backhaul Networks

Yaxiong Yuan[1*], Lei Lei[1], Thang X. Vu[1], Symeon Chatzinotas[1], Sumei Sun[2] and Björn Ottersten[1]

*Correspondence:
yaxiong.yuan@uni.lu
[1]Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, Kirchberg, 1855, Luxembourg, Luxembourg
Full list of author information is available at the end of the article
[†]An earlier version of the paper [1] was presented in the 29th edition of EuCNC - European Conference on Networks and Communications. This is an extended version.

Abstract

In unmanned aerial vehicle (UAV)-assisted networks, UAV acts as an aerial base station to serve ground users (GUs) through an access network, and meanwhile, the requested data traffic is acquired through the backhaul link. In this paper, we investigate an energy minimization problem with limited power supply for both backhaul and access links. The difficulties for solving such a non-convex and combinatorial problem lie at the high computational complexity/time. In solution development, we consider the approaches from both actor-critic deep reinforcement learning (AC-DRL) and optimization perspectives. Firstly, two offline non-learning algorithms, i.e., an optimal and a heuristic algorithms, based on piecewise linear approximation and relaxation are developed as benchmarks. Secondly, towards real-time decision making, we improve the conventional AC-DRL and propose two learning schemes: AC-based user group scheduling and backhaul power allocation (ACGP), and joint AC-based user group scheduling and optimization-based backhaul power allocation (ACGOP). Numerical results show that the computation time of both ACGP and ACGOP are reduced tenfold to hundredfold compared to the offline approaches, where ACGOP is better than ACGP in energy savings. The results also verify the superiority of proposed learning solutions in terms of guaranteeing the feasibility and minimizing the system energy compared to the conventional AC-DRL.

Keywords: UAV; deep reinforcement learning; user scheduling; backhaul power allocation; energy optimization; actor-critic

1 Introduction

Unmanned aerial vehicle (UAV)-assisted communication has been widely applied to various domains, e.g., aerial inspection, precision agriculture, traffic control, and after-disaster rescue [2]. Compared to terrestrial cellular systems, UAV-assisted systems 1) provide on-the-fly communication, which expands the coverage of ground wireless devices, and 2) have more probability to experience Line-of-Sight (LoS) transmission, which improves channel quality. In addition, the advances of UAVs' manufacturing technologies reduce the deployment cost of UAV networks and popularize their commercial and civilian usages [3].

However, one of the most critical issues of UAV-assisted networks is the limited on-board energy, which may shorten the UAVs' endurance and lead to service failure. Therefore, minimizing the UAV's energy consumption is of great importance. In [4], the authors proposed a joint power allocation and trajectory design algorithm to maximize UAV's propulsion energy efficiency. With the consideration of both communication energy and propulsion energy of UAV, the authors in [5] and [6] proposed energy-efficient communication schemes via user scheduling and

sub-channel allocation, respectively. We note that the works in [4–6] focused on the access link in UAV-assisted networks, where the UAV serves as an aerial base station (BS) that carries all the ground users' (GUs') requested data. In practice, due to limited storage capacity, the GU's requested data may be not available at the UAV's cache. In such case, the UAV needs to acquire the data from an auxiliary base station (ABS) through a backhaul link [7]. In [8], an energy efficiency maximization problem was investigated via power allocation and trajectory design, where the UAV performs as a relay between ABS and GUs. The authors in [9] proposed a joint trajectory design and spectrum allocation algorithm to minimize UAV's propulsion energy while satisfying the backhaul constraint, meaning that the transmitted data of the access link must be less than that of the backhaul link.

The user scheduling schemes in [8,9] are based on pure time division multiple access (TDMA) or frequency division multiple access (FDMA) with a single-antenna UAV. However, spatial division multiple access (SDMA) mode with multiple-antenna techniques and precoding design is able to improve network capacity, thereby reducing the tasks' completion time and total energy consumption. In [10], a non-orthogonal multiple access-based user scheduling and power allocation algorithm was proposed to minimize UAV's transmission energy with the backhaul constraint. In [11], the authors designed a game theory-based precoding scheme for multi-antenna UAV-assisted cluster networks. To maximize the UAV's propulsion energy efficiency, the authors in [12] proposed a power allocation scheme for multi-antenna UAV-enabled relay systems. However, the energy consumption of the backhaul link is studied to a limited extent in the above works [10–12], which is a large proportion of the total energy consumption and could be optimized by backhaul power control [13]. This motivates us to investigate an energy minimization problem, including both backhaul and access energy, in multiple-antenna UAV-assisted networks.

Optimization-based solutions, e.g., successive convex approximation [5] or Lagrangian dual method [6], might not be able to make time-efficient decisions. Firstly, the SDMA-based transmission mode enables the UAV to serve more than one GU simultaneously, resulting in an exponential growth of decision variables as well as the complexity [1]. Moreover, diversified energy models in UAV systems may lead to non-convexity in problem formulation, which makes the problem difficult to be solved optimally.

Deep reinforcement learning (DRL) learns the optimal policy from the interaction between environment and actions, instead of directly solving the optimization problem. DRL combines artificial neural networks with a reinforcement learning architecture to improve the learning efficiency and solution quality. Different from deep neural networks (DNNs), DRL is not necessary to prepare a large amount of data in advance for offline training. To maximize the energy efficiency, the authors in [14] and [15] applied deep Q network (DQN) to make decisions for resource block allocation and flight path planning, respectively. DQN needs to establish a Q-table containing all the possible actions before executing the algorithm so that it is usually for the decision tasks with discrete action space and a small number of decision variables [16].

Actor-critic-based DRL (AC-DRL) can tackle both discrete and continuous action space. For the problem with continuous variables, e.g., power control, AC-DRL

adopts a stochastic policy to select an action by probability. In [17], an energy-efficient UAV's direction control policy was proposed based on AC-DRL. To minimize UAV's energy consumption, in [18], the authors applied an AC-based deep deterministic policy gradient algorithm for UAV's velocity and direction control. In [17, 18], multiple decision variables in the problem modelings may lead to huge action space and slow convergence (more than 1000 learning episodes). It is noted that the solution proposed in [17, 18] can be applied to only unconstrained problems. However, for general UAV-assisted networks, the optimization problems have constraints [4–9, 11–13]. Therefore, directly applying AC-DRL may not lead to a high-quality and feasible solution.

In this paper, we propose two tailored AC-DRL-based schemes: AC-based user group scheduling and backhaul power allocation (ACGP), and joint AC-based user group scheduling and optimization-based backhaul power allocation (ACGOP). The main contributions are summarized as follows:

- We formulate a non-convex mixed-integer programming (NCMIP) problem to minimize both backhaul energy and access energy in UAV-assisted networks.
- To approach the optimum, we first transform the non-linear terms to linear by piece-wise linear approximation and McCormic envelopes, leading to a mixed-integer linear programming (MILP) problem, which can be solved optimally by branch and bound (B&B).
- We provide a near-optimal algorithm with lower computation time than the optimal method. Firstly, the original NCMIP problem is relaxed to a continuous optimization problem. Secondly, the relaxed problem is converted to a linear programming (LP) problem by piecewise linear approximation. Then, the heuristic solutions can be obtained after taking a rounding-up operation.
- Being aware of the high-complexity optimization methods, we propose ACGP and ACGOP learning schemes. To enable the learning algorithms to adapt to the considered NCMIP, in ACGP and ACGOP, we improve the conventional AC-DRL by a set of approaches, i.e., action filtering and reward re-design, to improve learning performance and avoid infeasible solutions.
- From the numerical results, we conclude that, compared with non-learning algorithms, ACGP and ACGOP have superiority in computational time efficiency, while compared with conventional AC-DRL, ACGP and ACGOP achieve better performance in delivering feasible solutions. Experiments also show that the combined learning-optimization scheme, i.e., ACGOP, achieves better energy-saving performance than ACGP.

The rest of the paper is organized as follows. Section 2 provides the system model. In Section 3, we formulate the considered optimization problem and solve it by proposing an optimal algorithm and a heuristic algorithm. In Section 4, we resolve the problem by DRL and develop an AC-DRL-based algorithm. Numerical results are presented and analyzed in Section 5. Finally, we draw the conclusions in Section 6.

Notations: Some mathematical operators are defined as follows. For a vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|$ and $\boldsymbol{a}^{\mathrm{H}}$ represent its Euclidean norm and conjugate transpose, respectively. For a matrix $\boldsymbol{A}$, $\boldsymbol{A}^{\mathrm{H}}$ refers to its conjugate transpose, and $\boldsymbol{A}^{\dagger}$ denotes its generalized inverse matrix. For scalars x and y, $\lceil x \rceil$ and $\lfloor x \rfloor$ means rounding-up and rounding-down operations, respectively. $[x]^{+}$ is equivalent to $\max\{0, x\}$. $\mathcal{N}(x, y)$ means a

Gaussian distribution with a mean x and a variance y. For a random variable X, $\mathbb{E}[X]$ is the statistical expectation of X.

2 System Model

We consider the UAV-assisted system, in which, a UAV acts as the aerial BS to serve multiple GUs in remote areas or when the terrestrial BS in the current service area is not available, e.g., destroyed in a disaster, as shown in Fig. 1. As the UAV operates at high altitude, it can overcome the influence of obstacles on the ground, e.g., buildings or mountains, and has more probability to experience LoS transmission. Specifically, when the UAV receives GUs' data requests, it first downloads these data from a remote ABS through a backhaul link and then distributes data to GUs through access links. The GUs in the service area are divided into several clusters due to the limited communication coverage of the UAV. We assume that the UAV serves all the clusters in a predetermined flying path at a fixed altitude. In a cluster, there exist K single-antenna GUs and each has q_k (bits) demands. The user set is denoted as $\mathcal{K} = \{1, ..., k, ..., K\}$. The total demands is denoted by $D = \sum_{k=1}^{K} q_k$. In each transmission task, all the GUs' demands need to be served within the time limitation T_{max} (seconds), including the time used for acquiring data from ABS and delivering data to GUs [1]. As shown in Fig. 2, the system spectrum is reused in a TDMA fashion so that the time domain of a transmission task is divided into sequence of timeslots $\mathcal{I} = \{1, ..., i, ..., I\}$, where I is the maximum number of timeslots, given by $\lfloor \frac{T_{max}}{\Phi} \rfloor$, and Φ (seconds) refers to the duration of each timeslot. In the access network, a timeslot accommodates multiple GUs with the SDMA transmission mode to further improve network capacity.

2.1 Backhaul Transmission

The ABS and UAV are equipped with L_t and L_r antennas, respectively, so that the backhaul link can be modeled as a MIMO channel. We assume that signals propagate through LoS transmission from ABS to UAV. Let $\mathbf{G} \in \mathbb{C}^{L_t \times L_r}$ be the channel matrix of the wireless backhaul link. The received signal at the UAV from the ABS can be described by:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{1}$$

where $\mathbf{x}$ and $\mathbf{n}$ denote the transmitted signal and white Gaussian noise of the UAV, respectively. In order to maximize the backhaul capacity, we employ the water-filling based power allocation [19]. The matrix $\mathbf{G}$ has a singular value decomposition (SVD):

$$\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\dagger}, \tag{2}$$

where $\mathbf{U} \in \mathbb{C}^{L_t \times L_t}$ and $\mathbf{V} \in \mathbb{C}^{L_r \times L_r}$ are unitary matrices, and $\mathbf{\Lambda} \in \mathbb{C}^{L_t \times L_r}$ is a diagonal matrix whose elements are non-negative real numbers. The diagonal

[1] The time and energy consumed on sending requests from GUs to UAV are not considered in this paper, since they are negligible compared to those on content delivery.

elements $\lambda_1, ..., \lambda_L$ in $\mathbf{\Lambda}$ are the ordered singular values (from large to small) for $\mathbf{G}$. Under the assumption that $\mathbf{G}$ is a full-rank matrix, let $L = \min\{L_t, L_r\}$. We process the UAV's received signal by:

$$\tilde{\mathbf{y}} = \mathbf{U}^{\dagger}\mathbf{y} = \sqrt{\mathbf{P}}\mathbf{\Lambda}\mathbf{V}^{\dagger}\mathbf{x} + \mathbf{U}^{\dagger}\mathbf{n}, \tag{3}$$

where $\sqrt{\mathbf{P}} = diag(\sqrt{p_1}, ..., \sqrt{p_L})$ referring to a diagonal matrix, and p_l means the power for each sub-channel. Thus, the capacity of the MIMO channel can be calculated by:

$$r^{bh} = B^{bh} \sum_{l=1}^{L} \log_2 \left(1 + \frac{p_l \lambda_l^2}{\sigma^2}\right), \tag{4}$$

where B^{bh} is the bandwidth of the backhaul link, and σ^2 is the receiver noise power of the UAV. Based on the water-filling power allocation, $p_l^* = \left[\mu - \frac{\sigma^2}{\lambda_l^2}\right]^+$, where μ is the water-filling level [19]. Thus, the total transmit power on the backhaul is:

$$p^{bh}(\mu) = \sum_{l=1}^{L} p_l^* = \sum_{l=1}^{L} \left[\mu - \frac{\sigma^2}{\lambda_l^2}\right]^+. \tag{5}$$

The achievable rate of the backhaul can be rewritten as:

$$r^{bh}(\mu) = B^{bh} \sum_{l=1}^{L} \left[\log_2 \left(\frac{\mu \lambda_l^2}{\sigma^2}\right)\right]^+. \tag{6}$$

At a timeslot, the backhaul transmission energy and the achievable transmitted data volume are:

$$e^{bh}(\mu) = \Phi p^{bh}(\mu), \tag{7}$$
$$d^{bh}(\mu) = \Phi r^{bh}(\mu). \tag{8}$$

2.2 Access Transmission

From Fig. 2, in the access transmission, the shaded block indicates that the user is scheduled. We define the scheduled users as a user group. Therefore, the maximum number of candidate groups can be calculated by $G = \sum_{l=1}^{L_r} \frac{K!}{l!(K-l)!}$, which increases exponentially with K. The group combination is $\mathcal{G} = \{1, ..., g, ..., G\}$. Towards eliminating multi-user interference within a group, minimum mean square error (MMSE) precoding is applied [20]. We denote K_g and $\mathcal{K}_g$ as the number and set of users in group g, and $\mathbf{h}_{k,g} \in \mathbb{C}^{L_r \times 1}$ as the channel vector for user $k \in \mathcal{K}_g$. We form $\mathbf{H}_g = \left[\mathbf{h}_{1,g}, ..., \mathbf{h}_{K_g,g}\right]$ as the channel matrix of group g. Based on the MMSE, the precoding vector $\mathbf{w}_{k,g} \in \mathbb{C}^{L_r \times 1}$ can be calculated by:

$$\mathbf{w}_{k,g} = \frac{\tilde{\mathbf{h}}_{k,g}}{\|\tilde{\mathbf{h}}_{k,g}\|}, \tag{9}$$

where $\tilde{\mathbf{h}}_{k,g}$ is to the k-th column of the MMSE precoding matrix $\mathbf{H}_g^{\mathrm{H}}(\sigma_{k,g}^2\mathbf{I} + \mathbf{H}_g\mathbf{H}_g^{\mathrm{H}})^{-1}$, $\sigma_{k,g}^2$ is the noise power for user $k \in \mathcal{K}_g$ and $\mathbf{I}$ is an identity matrix.

Since the UAV's transmit power is a constant selected from 0.1 W to 10 W in practical UAV application [21], we assume the transmit power for user k in group g is fixed, denoted as $p_{k,g}$. The received signal at GU $k \in \mathcal{K}_g$ is given by:

$$\begin{aligned} y_{k,g} =& \sqrt{p_{k,g}}\mathbf{h}_{k,g}^{\mathrm{H}}\mathbf{w}_{k,g}x_{k,g} + \sum_{j\in\mathcal{K}_g\setminus\{k\}} \sqrt{p_{j,g}}\mathbf{h}_{k,g}^{\mathrm{H}}\mathbf{w}_{j,g}x_{j,g} \\ &+ n_{k,g}, \quad k \in \mathcal{K}_g,\ g \in \mathcal{G}. \end{aligned} \tag{10}$$

where $x_{k,g}$ and $n_{k,g}$ denote the transmitted signal and white Gaussian noise of GU $k \in \mathcal{K}_g$. According to (10), we obtain the SINR of GUs $k \in \mathcal{K}_g$ as:

$$SINR_{k,g} = \frac{p_{k,g}|\mathbf{h}_{k,g}^{\mathrm{H}}\mathbf{w}_{k,g}|^2}{\sum_{j\in\mathcal{K}_g\setminus\{k\}} p_{j,g}|\mathbf{h}_{k,g}^{\mathrm{H}}\mathbf{w}_{j,g}|^2 + \sigma_{k,g}^2}, \tag{11}$$

Thus, the transmitted data volume for GU $k \in \mathcal{K}_g$ and the transmission energy for group g can be expressed as:

$$d_{k,g} = \Phi r_{k,g} = \Phi B^{ac} \log_2\left(1 + SINR_{k,g}\right), \tag{12}$$

$$e_g = \Phi p_g = \Phi \sum_{k\in\mathcal{K}_g} p_{k,g}, \tag{13}$$

where B^{ac} is the bandwidth of the access link.

2.3 UAV energy model

The propulsion power can be modeled as a function with regards to the flying velocity U [22], which is given by:

$$\begin{aligned} \mathcal{P}(U) =& P_0\left(1 + \frac{3U^2}{U_{tip}^2}\right) + P_1\left(\sqrt{1 + \frac{U^4}{4U_{ind}^4}} - \frac{U^2}{2U_{ind}^2}\right)^{\frac{1}{2}} \\ &+ \frac{1}{2}\varrho_1\varrho_2 U^3, \end{aligned} \tag{14}$$

where P_0 and P_1 are the blade profile power and induced power in hovering status, respectively. U_{tip} and U_{ind} refer to the tip speed of the rotor blade and mean rotor induced velocity, respectively. ϱ_1 is the parameter related to the fuselage drag ratio, rotor solidity, and the rotor disc area. ϱ_2 is denoted as the air density.

In the hovering phase, the UAV flies circularly around a hovering point with a small radius. To minimize the hovering power, the hovering velocity is given by:

$$U^{hov} = \underset{U\geq 0}{\operatorname{argmin}}\, \mathcal{P}(U). \tag{15}$$

Therefore, the hovering energy is only related to the hovering time. In the flying phase, the energy consumption with flying distance S is expressed as $\frac{S\mathcal{P}(U)}{U}$. When the flying path is predetermined, S is a constant parameter such that the flying velocity that minimize the flying energy is:

$$U^{fly} = \underset{U\geq 0}{\operatorname{argmin}}\, \frac{\mathcal{P}(U)}{U}. \tag{16}$$

Both U^{hov} and U^{fly} can be obtained by graph-based numerical methods [23]. Therefore, the hovering power p^{hov} and flying power p^{fly} are $\mathcal{P}(U^{hov})$ and $\mathcal{P}(U^{fly})$. Because the UAV suspends data transmission when flying between the clusters in the fly-hover-communicate protocol [5], the minimum flying energy is $\frac{S\mathcal{P}(U^{fly})}{U^{fly}}$.

3 Problem Formulation and Heuristic Approach

3.1 Problem Formulation

Our goal is to minimize the total system energy consumption via a joint design for user-timeslot scheduling and backhaul power allocation subject to the users' quality of service requirements. The total energy consumption consists of four parts: 1) the flying energy, 2) the hovering energy, 3) the backhaul transmission energy, and 4) the access transmission energy. As analyzed in the previous section, the flying energy is independent from the scheduling and power transmission decisions, hence can be skipped in the joint design. On the other hand, the hovering energy is determined by the transmission time, hence needs to be optimized.

We denote a set of binary variables indicating the channel occupation at each timeslot as follows:

$$\alpha_{g,i}^{ac} = \begin{cases} 1, & \text{group } g \in \mathcal{G} \text{ is scheduled at timeslot } i, \\ 0, & \text{otherwise.} \end{cases}$$

$$\alpha_{i}^{bh} = \begin{cases} 1, & \text{backhaul link is scheduled at timeslot } i, \\ 0, & \text{otherwise.} \end{cases}$$

Then joint design of user-timeslot scheduling (via $\alpha_{g,i}^{ac}, \alpha_i^{bh}$) and power allocation (via μ) for energy minimization can be formulated as follows:

$$\mathcal{P}_1:$$

$$\min_{\alpha_{g,i}^{ac}, \alpha_i^{bh}, \mu} \quad \sum_{i=1}^{I} \alpha_i^{bh} \left(e^{bh}(\mu) + e^{hov}\right) + \sum_{i=1}^{I}\sum_{g=1}^{G} \alpha_{g,i}^{ac} \left(e_g + e^{hov}\right) \tag{17a}$$

$$s.t. \quad \sum_{i=1}^{I}\sum_{g=1}^{G} \alpha_{g,i}^{ac} d_{k,g} \geq q_k, \ \forall k \in \mathcal{K}, \tag{17b}$$

$$d^{bh}(\mu) \sum_{i=1}^{I} \alpha_i^{bh} \geq D, \tag{17c}$$

$$\alpha_i^{bh} + \sum_{g=1}^{G} \alpha_{g,i}^{ac} \leq 1, \tag{17d}$$

$$\mu \leq u_{max}, \tag{17e}$$

$$\alpha_{g,i}^{ac}, \in \{0,1\}, \ \forall g \in \mathcal{G}, \ i \in \mathcal{I}, \tag{17f}$$

$$\alpha_i^{bh} \in \{0,1\}, \ \forall i \in \mathcal{I}, \tag{17g}$$

where $e^{bh}(\mu)$ and e_g are given in (7) and (13), respectively, and $e^{hov} = \Phi \cdot p^{hov}$ is the hovering energy at each timeslot.

In (17a), the first summation represents the transmission and hovering energy spent on the backhaul, and the second summation is the energy consumed on the

access links. Note that we optimize water-filling level μ instead of directly optimizing backhaul power p^{bh} since p^{bh} depends on μ based on Eq. (5). Constraints (17b) guarantee that each GU's request is satisfied in the access network. Constraint (17c) states that contents delivered through the backhaul should accommodate the total demands from the GUs. Constraint (17d) is to avoid concurrent transmission of the backhaul and access links. Constraint (17e) upper bounds the water-filling level to u_{max}, which is the maximal water-filling level under the backhaul's limitted transmit power. Constraints (17f) and (17g) confine variables $\alpha_{g,i}^{ac}$ and α_i^{bh} to binary.

Due to the non-convex items $e^{bh}(\mu)\alpha_i^{bh}$ and $d^{bh}(\mu)\alpha_i^{bh}$, $\mathcal{P}_1$ is a NCMIP problem which is difficult to obtain the optimal solution. One method to solve this problem is to apply a piece-wise linear approximation to linearize non-linear functions, i.e., $e^{bh}(\mu)$ and $d^{bh}(\mu)$ [24]. Thus, the approximations of $e^{bh}(\mu)\alpha_i^{bh}$ and $d^{bh}(\mu)\alpha_i^{bh}$ have a form of bilinear function, which can be transfomred to linear problems by using the McCormick envelopes [25]. The resulting problem is an integer linear programming (ILP) problem, which can be solved optimally by the B&B method. [26]. When the number of linear pieces is sufficient in fitted functions and the bounds of the McCormick envelopes are sufficiently tight, the solutions can approach to the global optimum. However, the operations of relaxation and approximation bring about high computation time (minutes level) which is unaffordable in practice.

3.2 Heuristic Approach

To reduce the computation time of the problem $\mathcal{P}_1$, we propose a heuristic algorithm. Firstly, we consider an extreme condition $\Phi \to 0$, such that $\mathcal{P}_1$ can be relaxed to a continuous optimization problem $\mathcal{P}_2$ in (20a). After relaxation, the allocated time for group g and the backhaul link are continuous values:

$$\tau_g = \lim_{\Phi \to 0} \Phi \sum_{i=1}^{T_{max}/\Phi} \alpha_{g,i}^{ac}, \tag{18}$$

$$\tau^{bh}(\mu) = D/r^{bh}(\mu). \tag{19}$$

$\mathcal{P}_2$ can be formulated as follows:

$$\min_{\tau_g,\ \mu} \quad \mathcal{F}(\mu) + \sum_{g=1}^{G} \tau_g \left(p_g + p^{hov}\right) \tag{20a}$$

$$s.t. \ \sum_{g=1}^{G} \tau_g d_{k,g} = q_k, \ \forall k \in \mathcal{K}, \tag{20b}$$

$$\tau^{bh}(\mu) + \sum_{g=1}^{G} \tau_g \leq T_{max}, \tag{20c}$$

$$\tau_g > 0, \ \forall g \in \mathcal{G}, \tag{20d}$$

$$\sigma^2/\lambda_1^2 < \mu \leq \mu_{max}, \tag{20e}$$

where

$$\begin{aligned}\mathcal{F}(\mu) =& \tau^{bh}(\mu)\cdot\left(p^{bh}(\mu)+p^{hov}\right)\\ \overset{\text{Eq.(5)}}{\underset{\text{Eq.(6)}}{=}}& \frac{D}{B^{bh}}\cdot\frac{\sum_{l=1}^{L}\left(\mu-\frac{\sigma^2}{\lambda_l^2}\right)^+ + p^{hov}}{\sum_{l=1}^{L}\left[\log_2\left(\frac{\mu\lambda_l^2}{\sigma^2}\right)\right]^+}.\end{aligned} \tag{21}$$

By fitting $\mathcal{F}(\mu)$ and $\tau^{bh}(\mu)$ with piece-wise linear approximations, $\mathcal{P}_2$ can be approximated as a linear programming (LP) problem, which can be solved by classical algorithms such as simplex method [27]. In practice, when $\Phi > 0$, $\mathcal{P}_2$ provides a lower bound of $\mathcal{P}_1$ and variables $\tau_1, ..., \tau_g$ are integer multiples of Φ. Thus, we take a rounding-up operation for post-processing, which introduces errors but makes the solutions of $\mathcal{P}_2$ feasible. We summarize the proposed heuristic algorithm in Alg. 1.

Algorithm 1 Proposed Heuristic algorithm

Inputs:
GUs' demands: $q_1, ..., q_K$;
Channel state: $\mathbf{G}, \mathbf{H}_1, ..., \mathbf{H}_G$;
GUs' transmit power: $p_{k,g}$; Bandwidth: B^{ac}, B^{bh};
Hovering power: p^{hov}; Duration of timesolt: Φ;
Maximum time: T_{max}; Maximum water-filling level μ_{max}.
Outputs:
Timeslots allocation for groups: $\bar{\tau}_1, ..., \bar{\tau}_1, ..., \bar{\tau}_G$;
Timeslots allocation for backhaul network: $\bar{\tau}^{bh}$;
Water-filling level : μ.
1: Reformulate the original problem $\mathcal{P}_1$ to the relaxed problem $\mathcal{P}_2$.
2: Transform $\mathcal{P}_2$ to a LP problem by fitting $\mathcal{F}(\mu)$ and $r^{bh}(\mu)$ as a piece-wise linear function.
3: Solve the LP problem and obtain the solutions: μ and $\{\tau_1, ..., \tau_g\}$.
4: Calculate $r^{bh}(\mu)$ by Eq. (6).
5: Calculate $\tau^{bh}(\mu)$ by Eq. (19).
6: Take ceiling operations by: $\bar{\tau}_g = \lceil \frac{\tau_g}{\Phi} \rceil$ and $\bar{\tau}^{bh} = \lceil \frac{\tau^{bh}(\mu)}{\Phi} \rceil$.

When Φ is sufficiently small, i.e., the solution of $\mathcal{P}_2$ approaches the optimal solution of $\mathcal{P}_1$ and the proposed heuristic method provides near-optimal solutions. The heuristic algorithm is more efficient than the optimal algorithm as solving the relaxed continuous problem is easier than solving its original integer programming problem. However, $\mathcal{P}_2$ is still suffered from high computation complexity as the number of variable is $G+1$, which exponentially increases with the number of GUs. This limits its application practice when the number of users is large or the latency requirement is stringent.

4 AC Overview and The Proposed Solutions

Being aware of the high computation complexity of the iterative optimal and suboptimal algorithms, We develop ACGP and ACGOP towards real-time applications.

4.1 AC-DRL Framework

To make the paper self-contained, we provide a brief overview for the adopted AC-DRL framework first. AC is one of the DRL frameworks, which integrates the strengths of both value-based and policy-based methods [28]. AC-DRL split the

learning agent into two components, where the actor is responsible for updating policies and making decisions while the critic is used for evaluating the decisions by value functions.

For the actor, the stochastic policy is applied, which is denoted as $\pi(a|s_t)$ representing the probability of taking action a under state s_t. Usually, we model $\pi(a|s_t)$ as Gaussian distribution with a mean $\psi(s_t)$ and a variance $\chi(s_t)$ [29]. At each learning step t, an action a_t is taken by following the policy $\pi(a|s_t)$. After that, the agent receives a reward r_t as the feedback. The objective of AC-DRL is to maximize the cumulative reward so that the loss function of the actor can be defined as:

$$J = \mathbb{E}[-Q^{\pi}(s_t, a_t)], \tag{22}$$

where $Q^{\pi}(s_t, a_t) = \mathbb{E}_{\pi}[\sum_{t'=t}^{\infty} \gamma^{t'-t} r'_t | s_t, a_t]$, representing a Q-value function with a discount factor γ. The critic is to evaluate the quality of the action by estimating the current Q-value. Temporal difference (TD) learning can be applied for Q-value estimation with high learning efficiency [28]. In TD learning, the TD error is the difference between the TD target $r_t + Q^{\pi}(s_{t+1}, a_{t+1})$ and the estimated Q-value $Q^{\pi}(s_t, a_t)$. The loss function of the critic is the square of TD error:

$$L = \mathbb{E}\left[(r_t + \gamma Q^{\pi}(s_{t+1}, a_{t+1})) - Q^{\pi}(s_t, a_t)\right]^2. \tag{23}$$

To update the policy and Q-value, we use parameterized functions, i.e., $\psi_{\boldsymbol{\theta}_t}(s_t)$, $\chi_{\boldsymbol{\theta}_t}(s_t)$ and $Q_{\boldsymbol{\omega}_t}(s_t, a_t)$, to approximate $\pi(a|s_t)$ and $Q^{\pi}(s_t, a_t)$:

$$\pi(a|s_t) \approx \pi_{\boldsymbol{\theta}_t}(a|s_t) \sim \mathcal{N}(\psi_{\boldsymbol{\theta}_t}(s_t), \chi_{\boldsymbol{\theta}_t}(s_t)), \tag{24}$$

$$Q^{\pi}(s_t, a_t) \approx Q_{\boldsymbol{\omega}_t}(s_t, a_t), \tag{25}$$

where $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_t$ are the parameters of the approximators. Based on the fundamental results of the policy gradient theorem [16], the gradient of $J(\boldsymbol{\theta}_t)$ and $L(\boldsymbol{\omega}_t)$ are given by:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t) = \mathbb{E}\left[-\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) Q_{\boldsymbol{\omega}_t}(s_t, a_t)\right], \tag{26}$$

$$\nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega}_t) = \mathbb{E}\left[2L(\boldsymbol{\omega}_t) \nabla_{\boldsymbol{\omega}} \left(Q_{\boldsymbol{\omega}_t}(s_{t+1}, a_{t+1}) - Q_{\boldsymbol{\omega}_t}(s_t, a_t)\right)\right]. \tag{27}$$

The update rules for $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_t$ can be derived based on gradient descend:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \rho \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t), \tag{28}$$

$$\boldsymbol{\omega}_{t+1} = \boldsymbol{\omega}_t - \rho \nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega}_t), \tag{29}$$

where ρ refers to the learning rate.

However, approximating $Q^{\pi}(s_t, a_t)$ directly brings about a large variance on gradient $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t)$, resulting in poor convergence [30]. To reduce the variance, we estimate a V-value function $V^{\pi}(\boldsymbol{s}_t) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{\infty} \gamma^{t'-t} r'_t | s_t\right]$ instead of Q-value. Based on TD learning and parameterized V-value $V_{\boldsymbol{\omega}_t}(s_t)$, the loss function of the critic can be expressed as:

$$L(\boldsymbol{\omega}_t) = \mathbb{E}[\delta_V(\boldsymbol{\omega}_t)]^2 = \mathbb{E}[r_t + \gamma V_{\boldsymbol{\omega}_t}(s_{t+1}) - V_{\boldsymbol{\omega}_t}(s_t)]^2. \tag{30}$$

In addition, the TD error $\delta_{_V}(\boldsymbol{\omega}_t)$ provides an unbiased estimation of Q-value [30]. Thus, we can rewrite Eq. (26) and Eq. (27) by:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_t) =& \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log(\pi(a_t|s_t;\boldsymbol{\theta}_t)) Q^{\pi}(s_t,a_t)\right] \\ =& \mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \log(\pi(a_t|s_t;\boldsymbol{\theta}_t)) \delta_{_V}(\boldsymbol{\omega}_t)\right], \qquad (31)\\ \nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega}_t) =& \mathbb{E}\left[2L(\boldsymbol{\omega}_t) \cdot \nabla_{\boldsymbol{\omega}} \left(V_{\boldsymbol{\omega}_t}(s_{t+1}) - V_{\boldsymbol{\omega}_t}(s_t)\right)\right] \qquad (32)\end{aligned}$$

In this paper, we apply DNNs as the approximators. The tuple $\{s_t, s_{t+1}, r_t, \delta_{_V}(\boldsymbol{\omega}_t)\}$ is stored in a repository over the learning process. At each learning step, a batch of tuples will be extracted as the training data for parameter updating.

4.2 The Proposed ACGP and ACGOP

We firstly reformulate P1 by defining states, actions, and rewards, such that an RL framework can apply. Next, we propose two AC-based solutions with highlighting the differences from conventional AC-DRL and tailored design for solving $\mathcal{P}_1$. In a learning episode, we denote the learning steps range from $t = 1$ to $t = t_e$, where t_e represents the last step when any termination condition reaches. We set the termination conditions by:

- The GUs' requests have been completed.
- The service runs out of time.

Based on the AC-DRL framework, we consider two schemes: (1) A straightforward learning approach ACGP, i.e., the agent makes decisions for all the variables. (2) A combined AC learning and simple optimization approach, i.e., ACGOP.

For ACGP, the system states s_t are jointly determined by the undelivered demands $b_{k,t}$ and the remaining timeslots η_t:

$$s_t = \{b_{1,t}, ..., b_{K,t}, \eta_t\}. \qquad (33)$$

The undelivered demands $b_{k,t}$ is the residual data to be transmitted to GU k at timeslots t. The actions a_t in ACGP are corresponding to the decision variables in $\mathcal{P}_1$. When $t = 1$, the agent predicts the water-filling level, i.e., $a_t = \mu$. The backhaul power $p^{bh}(a_t)$ and backhaul transmission rate $r^{bh}(a_t)$ can be calculated by Eq.(5), Eq.(6). Then, the backhaul energy is expressed as:

$$e^{bh}(a_t) = \bar{\tau}^{bh}(a_t)(p^{bh}(a_t) + p^{hov}), \qquad (34)$$

where $\bar{\tau}^{bh}(a_t) = \lceil D/r^{bh}(a_t) \rceil$. When $t = 2, ..., t_e$, the agent makes the decisions for user scheduling in the access network. The action $a_t = g$, representing the index of the selected user group. The expressions of the state transition are given by:

$$b_{k,t+1} = \begin{cases} q_k, & t = 1, \\ b_{k,t} - d_{k,a_t}, & t = 2, ..., t_e. \end{cases} \qquad (35)$$

$$\eta_{t+1} = \begin{cases} I - \bar{\tau}^{bh}(a_t), & t = 1, \\ \eta_t - 1, & t = 2, ..., t_e. \end{cases} \qquad (36)$$

The reward function r_t is commonly related to the objective of the original problem. For example, $r_t = -e_t$ is widely adopted for min-energy problems [31], where e_t is the energy consumed at step t, given by:

$$e_t = \begin{cases} e^{bh}(a_t), & t = 1, \\ e_{a_t} + e^{hov}, & t = 2, ..., t_e. \end{cases} \tag{37}$$

Note that, In the simulation $-e_t$ will be treated as a benchmark. A tailored reward function for ACGP and ACGOP can be found in (46).

In ACGOP, we observe that when user scheduling is fixed, the remaining backhaul power allocation becomes a single-variable optimization problem which is computationally light. Thus, the agent in ACGOP only takes actions for user scheduling while the backhaul power is determined by an efficient golden-section search approach. Specifically, the state s_t keeps same as in ACGP. When $t = 1, ..., t_e$, the learning agent makes decision for user scheduling, i.e., $a_t = g$. The expressions of state transition can be rewritten as:

$$b_{k,t+1} = \begin{cases} q_k - d_{k,a_t}, & t = 1, \\ b_{k,t} - d_{k,a_t}, & t = 2, ..., t_e. \end{cases} \tag{38}$$

$$\eta_{t+1} = \begin{cases} I - 1, & t = 1, \\ \eta_t - 1, & t = 2, ..., t_e. \end{cases} \tag{39}$$

When a termination condition is reached, i.e., $t = t_e$, if $\eta_t \leq 0$, then the solutions are not feasible, otherwise η_t can be regarded as the available number of timeslots for backhaul transmission. Since the user scheduling is obtained by the learning agent, the original problem can be reduced to a single-variable power control problem $\mathcal{P}_3$:

$$\min_{\frac{\sigma^2}{\lambda_1^2} < \mu \leq \mu_{max}} \quad \mathcal{F}(\mu) \tag{40a}$$

$$s.t. \quad \tau^{bh}(\mu) = \Phi \eta_{t_e}, \tag{40b}$$

Lemma 1 *Assume* $\frac{\sigma^2}{\lambda_1^2} < 1$, $\mathcal{F}(\mu)$ *is a unique function with a unique minimum point in* $\left[\frac{\sigma^2}{\lambda_1^2}, +\infty\right]$.

Proof See the appendix 7.1. □

Fig. 3 illustrates the function graph of $\mathcal{F}(\mu)$. Based on Lemma 1, the optimal value μ^* can be quickly found by golden section search [32]. After that, the backhaul energy $e^{bh}(\mu^*)$ can be calculated by Eq. (34). The energy consumption at each time step is rewritten as:

$$e_t = \begin{cases} e_{a_t} + e^{hov}, & t = 1, ..., t_e - 1, \\ e_{a_t} + e^{hov} + e^{bh}(\mu^*), & t = t_e, \end{cases} \tag{41}$$

We observe that conventional AC-DRL may have limitations on dealing with $\mathcal{P}_1$. Firstly, the decision variables in $\mathcal{P}_1$ are both continuous and discrete. Thus, we need to map the stochastic policy in AC-DRL to the corresponding action space. Secondly, the action spaces is huge due to the combinatorial nature of $\mathcal{P}_1$. Searching in such a huge space may reduce learning efficiency and solution quality. Thirdly, conventional AC-DRL may converge to an infeasible solution without tailored reward design. In this paper, we propose a set of approaches to address the above issues.

4.2.1 Action Mapping

Denote $\hat{a}_t$ as the original action selected by the stochastic policy $\pi(a|s_t)$. Since $\pi(a|s_t)$ follows Gaussian distribution, $\hat{a}_t$ is a continuous value on $[-\infty, +\infty]$. We introduce two mapping functions:

$$\mathcal{M}_1(x) = \min\{\max\{-\kappa, 0\}, \kappa\}, \tag{42}$$

$$\mathcal{M}_2(x) = \lceil \frac{\kappa + \mathcal{M}_1(x)}{2\kappa/G} \rceil. \tag{43}$$

$\mathcal{M}_1(x)$ maps x to a continuous space $[-\kappa, \kappa]$, where κ is a positive parameter, while $\mathcal{M}_2(x)$ maps x to a discrete space $\mathcal{G} = \{1, 2, ..., G\}$. In order to map $\hat{a}_t$ to the corresponding action space, we define the after-mapped action a_t as:

- ACGP : $$a_t = \begin{cases} \frac{\mu_{max}(\mathcal{M}_1(\hat{a}_t)+\kappa)}{2\kappa}, & t = 1, \\ \mathcal{M}_2(\hat{a}_t), & t = 2, ..., t_e. \end{cases} \tag{44}$$
- ACGOP : $$a_t = \mathcal{M}_2(\hat{a}_t),\ t = 1, ..., t_e. \tag{45}$$

4.2.2 Action Filtering

The size of discrete space $\mathcal{G}$ increases exponentially with the number of users. To confine the action space, we eliminate a considerable amount of redundant actions which bring no benefit to rewards. Specifically, the redundant actions refer to scheduling the groups that contain demand-satisfied GUs. Therefore, at the beginning of each step, we take an action filtering operation to find which GUs' demands have been satisfied and remove the corresponding groups. As a result, the action space decreases gradually over the learning steps, thereby improving the search efficiency and the solution quality.

4.2.3 Reward Design

All the constraints in $\mathcal{P}_1$ except (17b) can be met by properly defining actions and states. The constraints (17b) cannot be guaranteed as the commonly-used reward function, i.e., $r_t = -e_t$, purely minimizes energy, and the GU's demand is not taken into account. We re-design a tailored reward function. Firstly, if the after-learned policy is infeasible at the end of each episode, the agent will get a penalty $-\zeta$ which is negative [33]. Secondly, an extra reward $\epsilon \sum_{k=1}^{K} d_{k,a_t}$ will be added to r_t. That is, the reward enforces the actor to deliver more data to meet GUs' demands. However, transmitting more data results in more energy consumption. In this case, we can decrease the weight factor ϵ to control energy growth. The re-designed reward is

expressed as:

$$r_t = \begin{cases} -\zeta, & \text{if } t = t_e \text{ and } \sum_{k=1}^{K} b_{k,t} > 0 \\ -e_t + \epsilon \sum_{k=1}^{K} d_{k,a_t}, & \text{otherwise} \end{cases} \tag{46}$$

In Alg. 2, we summarize the pseudo-code of ACGOP. Analogous to ACGOP, Alg. 2 can apply to ACGP by replacing Eq.(45), Eq. (41), Eq. (38) and Eq. (39) with Eq. (44), Eq. (37), Eq. (35) and Eq. (36), respectively.

Algorithm 2 ACGOP algorithm

Inputs:
The current state s_t.
Outputs: The selected action a_t.
Initialize s_1, $\boldsymbol{\theta}_1$ and $\boldsymbol{\omega}_1$.
for each learning step t **do**
  Remove the groups containing the demand-satisfied GUs.
  Predict $\psi_{\boldsymbol{\theta}}(s_t)$ and $\chi_{\boldsymbol{\theta}}(s_t)$ by the actor DNN.
  Obtain $\pi_{\boldsymbol{\theta}_t}(a_t|s_t)$ based on Eq. (24).
  Randomly choose $\hat{a}_t$ following $\pi_{\boldsymbol{\theta}_t}(a_t|s_t)$.
  Map $\hat{a}_t$ to a_t by Eq. (45).
  Take the action and calculate e_t by Eq. (41)
  Obtain r_t by Eq. (46).
  Observe s_{t+1} by Eq. (38) and Eq. (39).
  Predict $V_{\boldsymbol{\omega}_t}(s_t)$ and $V_{\boldsymbol{\omega}_t}(s_{t+1})$ by the critic DNN.
  Calculate TD error $\delta_V(\boldsymbol{\omega}_t)$ by Eq. (30).
  Store a tuple $\{s_t, s_{t+1}, r_t, \delta_V(\boldsymbol{\omega}_t)\}$ in the memory.
  Obtain $\boldsymbol{\theta}_{t+1}$ and $\boldsymbol{\omega}_{t+1}$ by gradient descent.
  if $\eta_t < 0$ or $\sum_{k=1}^{K} b_{k,t} = 0$ **then**
    Break.
  end if
  $s_t = s_{t+1}$; $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t+1}$; $\boldsymbol{\omega}_t = \boldsymbol{\omega}_{t+1}$.
end for

5 Numerical Results

In this section, we evaluate the performance of the proposed solutions and other three non-learning benchmarks:

- Optimal approach (OPT): McCormick envelopes + B&B (refer to Section 3).
- Prop-HEU: Near-optimal algorithm in Alg. 1.
- Semi-orthogonal user scheduling-based heuristic algorithm (SUS-HEU) [34]: Applying SUS for user scheduling and solving $\mathcal{P}_3$ backhaul for power allocation.

In addition, we simulate two conventional AC-DRL schemes based on [31] for performance comparison.

5.1 Parameter Settings

The parameter setting is similar to that in [12]. We consider both the ABS and UAV are equipped with $L_t = L_r = 3$ antennas. In the backhaul link, the communication channel is modeled as a LoS-MIMO channel. The channel matrix $\mathbf{G}$ is determined

with the spherical wave model [35] which is given by:

$$\mathbf{G} = \begin{bmatrix} o_{1,1}^{-\beta} e^{\mathrm{j}2\pi f_c o_{1,1}} & \cdots & o_{1,L_r}^{-\beta} e^{\mathrm{j}2\pi f_c o_{1,L_r}} \\ \vdots & \ddots & \vdots \\ o_{L_t,1}^{-\beta} e^{\mathrm{j}2\pi f_c o_{L_t,1}} & \cdots & o_{L_t,L_r}^{-\beta} e^{\mathrm{j}2\pi f_c o_{L_t,L_r}} \end{bmatrix},$$

where o_{l_t,l_r} corresponds to the path length between the l_t-th transmitting antenna and the l_r-th receiving antenna, $f_c = 2.4$ (GHz) refers to the carrier frequency, and $\beta = 2.6$ is the path loss exponent. In the access link, the GUs are randomly scattered and separated into $N = 3$ clusters. In each cluster, the number of GUs is up to $K = 10$. The GUs' demands are randomly selected from the set {3, 3.5, 4, 4.5, 5} (Gbits). We model the access channel as the multi-user LoS-MISO channel expressed as:

$$\mathbf{h}_{k,g} = \left[\iota_{k,g,1}^{-\beta} e^{\mathrm{j}2\pi f_c \iota_{k,g,1}}, \cdots, \iota_{k,g,L_r}^{-\beta} e^{\mathrm{j}2\pi f_c \iota_{k,g,L_r}} \right],$$

where ι_{k,g,l_r} means the distance between the UAV's l_r-th antenna and the k-th GU of the g-th group. We assume the bandwidth for the ABS and UAV are $B^{bh} = 1$ (GHz) and $B^{ac} = 0.05$ (GHz). The maximum water-filling level μ_{max} is set to 10 units. The UAV's hovering power p^{hov} and GUs' transmit power $p_{k,g}$ is 5 (Watt) and 2 (Watt), respectively. The noise power in UAV σ^2 and GUs $\sigma_{k,g}^2$ are -87.49 (dB) and -116.98 (dB). The duration of timeslot Φ is set as 0.1 (s).

Two fully-connected DNNs are employed as the actor and the critic. The adopted parameters in ACGOP and ACGP are summarized in Table 1.

5.2 Results and Analysis

We compare the performance of the algorithms in terms of energy minimization and computation time. Fig. 4 shows the objective energy with the number of users K. We can observe that ACGOP has 3.97% gap to the optimum, while for ACGP, the gap increases to 10.27%. Prop-HEU obtains a near-optimal solution with 1.61% average gap but requires much more computation time, e.g., see Fig. 5. SUS-HEU results in the highest energy consumption among all the schemes due to its inappropriate grouping strategy in energy savings.

Fig. 5 compares the computation time with respect to K. The computation time refers to the time from giving inputs to algorithms until receiving the results. From Fig. 5, the computation time of OPT and Prop-HEU grows exponentially with K. When K=10, the computation time reaches at 11 (s) and 90 (s), respectively. ACGOP, ACGP, and SUS-HEU can provide online solutions by applying the after-learned DRL policy or low-complexity SUS strategy to avoid directly solving complex optimization problems, thereby saving tenfold to hundredfold computation time compared with OPT and Prop-HEU. The average computation time of the three algorithms is relatively close. However, by recalling the energy-saving performance, ACGOP saves 8.21% and 15.28% energy compared to ACGP and SUS-HEU, respectively.

Fig. 6 demonstrates the total energy consumption with respect to T_{max}. When T_{max} increases from 14 (s) to 17 (s), the energy consumption reduces by 10.43%,

12.34%, and 15.31% for Prop-HEU, ACGP, and ACGOP, respectively. This is because, in the access network, a small T_{max} may enforce more GUs to share the same timeslot, which increases inter-user interference as well as the precoding energy. On the other hand, in the backhaul network, the system needs to allocate more backhaul power to satisfy backhaul constraint within a very limited time. When the transmission time is sufficient, e.g., T_{max} >17 (s), the min-energy points in all the schemes are achieved.

Fig. 7 and Fig. 8 illustrates the impacts of different learning rates ρ for ACGOP on the performance of convergence and feasibility. From Fig. 7, we can obverse that the objective energy converges over the learning episodes. The convergence speed in the case of $\rho = 10^{-3}$ is faster than that of $\rho = 10^{-4}$. Whereas, when ρ increases to 10^{-2}, the curve has large fluctuations and the energy at the convergence is higher than that of $\rho = 10^{-3}$ and $\rho = 10^{-4}$. Fig. 8 depicts the total transmitted data over learning episodes. When $\rho = 10^{-3}$ and $\rho = 10^{-4}$, the two curves are overlapped and the after-converged solutions for both are feasible, i.e., the transmitted data are equal to the demands. But for $\rho = 10^{-2}$, the feasibility can not be guaranteed. Therefore, to achieve a fast learning speed while ensuring the feasibility of the solutions, the learning rates need to be appropriately selected. Taking ACGOP as an example, ACGP has the same tendency.

Fig. 9 and Fig. 10 compares the proposed solutions with conventional AC-DRL. From Fig. 9, ACGOP, ACGP and conventional AC-DRL with the reward in [31] and action filtering have similar performance in energy minimization. Conventional AC-DRL with the reward in Eq. (46) and without action filtering performs badly, which has slow convergence speed and high after-converged energy. Moreover, Fig. 10 demonstrates that neither the Conventional AC-DRL schemes can guarantee feasibility. The reason is that the reward in [31] is only related to the objective function but fails to consider the constraints of the problem. For AC-DRL without action filtering, a huge space may lead to low exploration efficiency and degraded performance.

To demonstrate the superiority of water-filling-based sub-channel power allocation in the backhaul network, we compare it with the uniform allocation scheme in Fig. 11. It can be found that backhaul energy of the water-filling-based scheme is 40.35% lower than that of the uniform allocation scheme on average. This is because, the water-filling method is able to maximize the capacity for the MIMO system. With a given total power p^{bh}, the water-filling-based scheme has a higher transmission rate and less transmission time τ^{bh} than other schemes. Thus, the backhaul transmission energy $p^{bh}\tau^{bh}$ is reduced.

6 Conclusion

In this paper, we studied a joint user-timeslot scheduling and backhaul power allocation problem to minimize the energy consumed in the access and backhaul UAV-assisted systems. We developed an optimal method and a heuristic algorithm as the non-learning benchmarks. However, due to the high computation time, the above methods cannot provide real-time solutions. To this end, we addressed the problem from the perspective of AC-DRL and proposed two learning schemes, i.e., ACGP and ACGOP. Different from conventional AC-DRL, the proposed ACGOP

combines AC and optimization to accelerate learning performance. In addition, we design a set of approaches, such as action filtering and reward re-design, to reduce huge action space and guarantee the feasibility. Numerical results demonstrated that ACGOP and ACGP improve computational efficiency and guarantee solution feasibility. Simulations also showed that ACGOP achieves better energy-saving performance than ACGP.

7 Appendix

7.1 Proof of Lemma 1

$\mathcal{F}(\mu)$ can be expressed as a piece-wise function:

$$\mathcal{F}(\mu) = f_l(\mu), \quad \mu \in \left[\frac{\sigma^2}{\lambda_l^2}, \frac{\sigma^2}{\lambda_{l+1}^2}\right), \ l = 1, ..., L, \tag{47}$$

where $f_l(\mu) = \frac{D}{B^{bh}} \cdot \frac{\mu - a_l + c_l}{\log_2(\mu) + b_l}$, $a_l = \frac{1}{l}\sum_{l'=1}^{l} \frac{\sigma^2}{\lambda_{l'}^2}$, $b_l = \frac{1}{l}\sum_{l'=1}^{l} \log_2\left(\frac{\sigma^2}{\lambda_{l'}^2}\right)$, $c_l = \frac{p^{hov}}{l}$ and $\lambda_{L+1} = 0$. The function $f(\mu)$ can prove to be continuous but not differentiable at the breakpoints between adjacent intervals. We define $\phi_l(\mu) = \mu(\log_2(\mu) + b_l)$ and $\varphi_l(\mu) = \mu - a_l + c_l$. The first derivative and second derivative of $f_l(\mu)$ are given by:

$$f_l'(\mu) = \frac{D}{B^{bh}} \cdot \frac{\ln 2 \cdot \phi_l(\mu) - \varphi_l(\mu)}{\ln 2 \cdot \phi_l^2(\mu)/\mu}, \tag{48}$$

Based on Eq. (48), we can derive:

$$(1) \text{ if } \ \mu_l^* < \frac{\sigma^2}{\lambda_l^2}, \quad f_l'(\mu) > 0, \mu \in \left[\frac{\sigma^2}{\lambda_l^2}, \frac{\sigma^2}{\lambda_{l+1}^2}\right), \tag{49}$$

$$(2) \text{ if } \ \mu_l^* \geq \frac{\sigma^2}{\lambda_{l+1}^2}, \quad f_l'(\mu) < 0, \mu \in \left[\frac{\sigma^2}{\lambda_l^2}, \frac{\sigma^2}{\lambda_{l+1}^2}\right), \tag{50}$$

$$(3) \text{ if } \ \frac{\sigma^2}{\lambda_l^2} \leq \mu_l^* < \frac{\sigma^2}{\lambda_{l+1}^2}, \quad f_l'(\mu) \begin{cases} < 0, & \mu \in \left[\frac{\sigma^2}{\lambda_l^2}, \mu_l^*\right), \\ = 0, & \mu = \mu_l^*, \\ > 0, & \mu \in \left(\mu_l^*, \frac{\sigma^2}{\lambda_{l+1}^2}\right), \end{cases} \tag{51}$$

where μ_l^* is the point that satisfies $\ln 2 \cdot \phi_l(\mu_l^*) = \varphi_l(\mu_l^*)$. Since $\lambda_l > \lambda_{l+1}$, we can derive that $a_l < a_{l+1}$, $b_l < b_{l+1}$, $c_l > c_{l+1}$ and $\mu_l^* > \mu_{l+1}^*$ by graphical method, as shown in Fig. 12.

Recalling the precondition that $\frac{\sigma^2}{\lambda_1^2} < 1$, it is not difficult to prove that $\mu_1^* > \frac{\sigma^2}{\lambda_1^2}$. Then, $\mathcal{F}'(\frac{\sigma^2}{\lambda_1^2}) = f_1'(\frac{\sigma^2}{\lambda_1^2}) < 0$ based on Eq.(50) and Eq.(51). Moreover, $\mathcal{F}'(\mu) = f_L'(\mu) > 0$ when $\mu > \max\{\mu_L, \frac{\sigma^2}{\lambda_L^2}\}$. Thus, there must exist minimum points in $\left[\frac{\sigma^2}{\lambda_1^2}, +\infty\right)$ on $\mathcal{F}(\mu)$. We assume a minimum point μ^{**} is located in $\left[\frac{\sigma^2}{\lambda_l^2}, \frac{\sigma^2}{\lambda_{l+1}^2}\right)$. There are two situations:

- $\mu^{**} \in \left(\frac{\sigma^2}{\lambda_l^2}, \frac{\sigma^2}{\lambda_{l+1}^2}\right)$. In this case, $\mu^{**} = \mu_l^*$. On one hand, we can derive that $\mu_L^* < \cdots < \mu_{l+1}^* < \mu_l^* < \frac{\sigma^2}{\lambda_{l+1}^2} < \cdots < \frac{\sigma^2}{\lambda_L^2}$, i.e., $\mu_m^* < \frac{\sigma^2}{\lambda_m^2}$, $m = l+1, ..., L$. Based on Eq. (49) and Eq.(51), it can be concluded:

$$\mathcal{F}'(\mu) > 0, \mu \in (\mu_l^*, +\infty). \tag{52}$$

On the other hand, we can also obtain that $\mu_1^* > \cdots > \mu_{l-1}^* > \mu_l^* > \frac{\sigma^2}{\lambda_{l+1}^2} > \frac{\sigma^2}{\lambda_l^2} > \cdots > \frac{\sigma^2}{\lambda_2^2}$, i.e., $\mu_m^* > \frac{\sigma^2}{\lambda_{m+1}^2},\ m = 1, ..., l-1$. Based on Eq. (50) and Eq. (51), it can be concluded:

$$\mathcal{F}'(\mu) < 0, \mu \in \left[\frac{\sigma^2}{\lambda_1^2}, \mu_l^*\right). \tag{53}$$

Therefore, $\mu^{**} = \mu_l^*$ is the only minimum point on $\mathcal{F}(\mu)$.

- $\mu^{**} = \frac{\sigma^2}{\lambda_l^2}$. In this case, we can derive that $\mu_L^* < \cdots < \mu_{l+1}^* < \mu_l^* < \frac{\sigma^2}{\lambda_l^2} < \frac{\sigma^2}{\lambda_{l+1}^2} < \cdots < \frac{\sigma^2}{\lambda_L^2}$, i.e., $\mu_m^* < \frac{\sigma^2}{\lambda_m^2},\ m = l, ..., L$, and $\mu_1^* > \cdots > \mu_{l-1}^* > \mu_l^* > \mu_{l-1}^* > \frac{\sigma^2}{\lambda_l^2} > \frac{\sigma^2}{\lambda_{l-1}^2} > \cdots > \frac{\sigma^2}{\lambda_2^2}$, i.e., $\mu_m^* > \frac{\sigma^2}{\lambda_{m+1}^2},\ m = 1, ..., l-1$. Based on Eq. (49) and Eq. (50), we can conclude:

$$\mathcal{F}'(\mu) > 0, \mu \in \left(\frac{\sigma^2}{\lambda_l^2}, +\infty\right), \tag{54}$$

$$\mathcal{F}'(\mu) < 0, \mu \in \left[\frac{\sigma^2}{\lambda_1^2}, \frac{\sigma^2}{\lambda_l^2}\right). \tag{55}$$

Therefore, $\mu^{**} = \frac{\sigma^2}{\lambda_l^2}$ is the only minimum point.

Thus the conclusion.

Acknowledgements
The work has been supported by the ERC project AGNOSTIC (742648), by the FNR CORE projects RO-SETTA (11632107), ProCAST (C17/IS/11691338) and 5G-Sky (C19/IS/13713801), and by the FNR bilateral project LARGOS (12173206).

Funding
The work has been supported by the ERC project AGNOSTIC (742648), by the FNR CORE projects RO-SETTA (11632107), ProCAST (C17/IS/11691338) and 5G-Sky (C19/IS/13713801), and by the FNR bilateral project LARGOS (12173206).

Abbreviation
UAV: Unmanned aerial vehicle; LoS: Line-of-Sight; GU: Ground user; TDMA: Time division multiple access; FDMA: Frequency division multiple access; SDMA: Spatial division multiple access; DRL: Deep reinforcement learning; DNN: Deep neural network; AC-DRL: Actor-critic-based DRL; NCMIP: Non-convex mixed-integer programming; MILP: Mixed-integer linear programming; ILP: Integer programming LP: Linear programming; B&B: Branch and bound; MMSE: minimum mean square error;

Availability of data and materials
The codes for generating the results are online available at the link: https://github.com/ArthuretYuan.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YY is the main author of the current paper, contributing the ideas, modeling, solutions, and writing. LL and TV contributed to the conception and design of the study as well as paper revision. SC, SS and BO commented the work. All authors read and approved the final manuscript.

Author details
[1]Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, Kirchberg, 1855, Luxembourg, Luxembourg. [2]Institute for Infocomm Research, Agency for Science, Technology, and Research, 138632, Singapore, Singapore.

References

1. Y. Yuan, L. Lei, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Actor-critic deep reinforcement learning forenergy minimization in uav-aided networks," in *2020 European Conference on Networks and Communications (EuCNC)*, 2020.
2. M. Mozaffari, W. Saad, M. Bennis, Y. Nam, and M. Debbah, "A tutorial on uavs for wireless networks: Applications, challenges, and open problems," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

3. M. M. U. Chowdhury, S. J. Maeng, E. Bulut, and I. Güvenç, "3d trajectory optimization in uav-assisted cellular networks considering antenna radiation pattern and backhaul constraint," *IEEE Transactions on Aerospace and Electronic Systems*, 2020.
4. S. Ahmed, M. Z. Chowdhury, and Y. M. Jang, "Energy-efficient uav-to-user scheduling to maximize throughput in wireless networks," *IEEE Access*, vol. 8, pp. 21 215–21 225, 2020.
5. Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing uav," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2329–2345, 2019.
6. H. Yang and X. Xie, "Energy-efficient joint scheduling and resource management for uav-enabled multicell networks," *IEEE Systems Journal*, vol. 14, no. 1, pp. 363–374, 2020.
7. H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in uav-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Communications Magazine*, vol. 56, no. 6, pp. 28–34, 2018.
8. S. Ahmed, M. Z. Chowdhury, and Y. M. Jang, "Energy-efficient uav relaying communications to serve ground nodes," *IEEE Communications Letters*, vol. 24, no. 4, pp. 849–852, 2020.
9. C. Qiu, Z. Wei, Z. Feng, and P. Zhang, "Backhaul-aware trajectory optimization of fixed-wing uav-mounted base station for continuous available wireless service," *IEEE Access*, vol. 8, pp. 60 940–60 950, 2020.
10. M. Youssef, J. Farah, C. Abdel Nour, and C. Douillard, "Full-duplex and backhaul-constrained uav-enabled networks using noma," *IEEE Transactions on Vehicular Technology*, 2020.
11. Z. Xu, L. Li, H. Xu, A. Gao, X. Li, W. Chen, and Z. Han, "Precoding design for drone small cells cluster network with massive mimo: A game theoretical approach," in *2018 14th International Wireless Communications Mobile Computing Conference (IWCMC)*, pp. 1477–1482, 2018.
12. Q. Song and F. Zheng, "Energy efficient multi-antenna uav-enabled mobile relay," *China Communications*, vol. 15, no. 5, pp. 41–50, 2018.
13. T. X. Vu, S. Chatzinotas, B. Ottersten, and T. Q. Duong, "Energy minimization for cache-assisted content delivery networks with wireless backhaul," *IEEE Wireless Communications Letters*, vol. 7, no. 3, pp. 332–335, 2018.
14. F. Ghavimi and R. Jantti, "Energy-efficient uav communications with interference management: Deep learning framework," in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2020.
15. W. Liu, P. Si, E. Sun, M. Li, C. Fang, and Y. Zhang, "Green mobility management in uav-assisted iot based on dueling dqn," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019.
16. R. S. Sutton and A. Barto, "Reinforcement learning: An introduction," *Robotica*, vol. 17, no. 2, pp. 229–235, 1999.
17. 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, 2018.
18. H. Qi, Z. Hu, H. Huang, X. Wen, and Z. Lu, "Energy efficient 3-d uav control for persistent communication service and fairness: A deep reinforcement learning approach," *IEEE Access*, vol. 8, pp. 53 172–53 184, 2020.
19. D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
20. C. Zhang, W. Xu, and M. Chen, "Robust mmse beamforming for multiuser miso systems with limited feedback," *IEEE Signal Processing Letters*, vol. 16, no. 7, pp. 588–591, 2009.
21. C. Yan, L. Fu, J. Zhang, and J. Wang, "A comprehensive survey on uav communication channel modeling," *IEEE Access*, vol. 7, pp. 107 769–107 792, 2019.
22. H. D. Tran, T. X. Vu, S. Chatzinotas, S. Shahbazpanahi, and B. OTTERSTEN, "Coarse trajectory design for energy minimization in uav-enabled wireless communications with latency constraints," *IEEE Transactions on Vehicular Technology*, 2020.
23. A. Filippone, *Flight performance of fixed and rotary wing aircraft*, Elsevier, 2006.
24. H. Ahmadi, J. R. Martí, and A. Moshref, "Piecewise linear approximation of generators cost functions using max-affine functions," in *2013 IEEE Power Energy Society General Meeting*, 2013.
25. G. P. McCormick, "Computability of global solutions to factorable nonconvex programs: Part i: Convex underestimating problems," *Mathematical programming*, vol. 10, no. 1, pp. 147–175, 1976.
26. W. Zhang, "Branch-and-bound search algorithms and their computational complexity." University of southern california marina del rey information sciences institution, Tech. Rep., 1996.
27. F. A. Ficken, *The simplex method of linear programming*, Courier Dover Publications, 2015.
28. V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000.
29. Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in hetnets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680–692, 2018.
30. J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.
31. Y. Lu, H. Lu, L. Cao, F. Wu, and D. Zhu, "Learning deterministic policy with target for power control in wireless networks," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018.
32. T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT press, 2009.
33. T. Yang, Y. Hu, M. C. Gursoy, A. Schmeink, and R. Mathar, "Deep reinforcement learning based resource allocation in low latency edge computing networks," in *2018 15th International Symposium on Wireless Communication Systems (ISWCS)*, 2018.
34. T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 528–541, 2006.
35. M. H. Castañeda Garcia, M. Iwanow, and R. A. Stirling-Gallacher, "Los mimo design based on multiple optimum antenna separations," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018.
36. Z. Li, Y. Wang, M. Liu, R. Sun, Y. Chen, J. Yuan, and J. Li, "Energy efficient resource allocation for uav-assisted space-air-ground internet of remote things networks," *IEEE Access*, vol. 7, 2019.

Figures

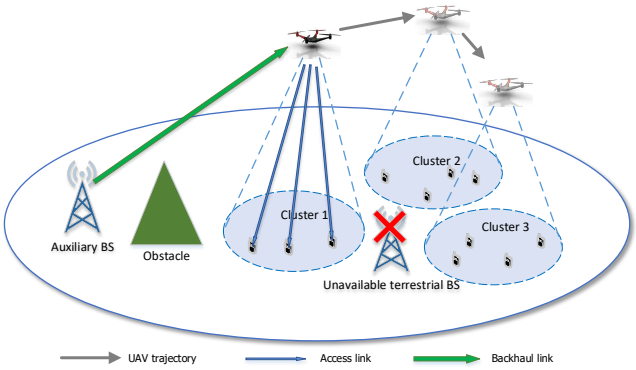

Figure 1: A UAV network with $N = 3$ clusters.

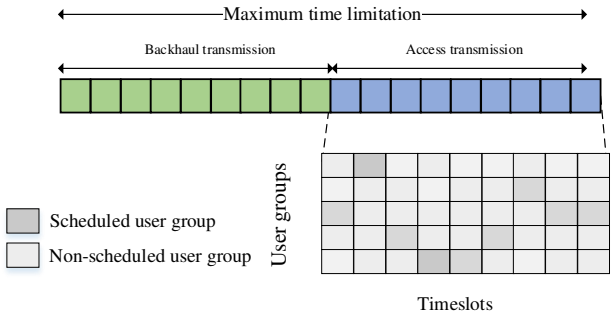

Figure 2: An illustration of the structure of the transmission process.

Tables

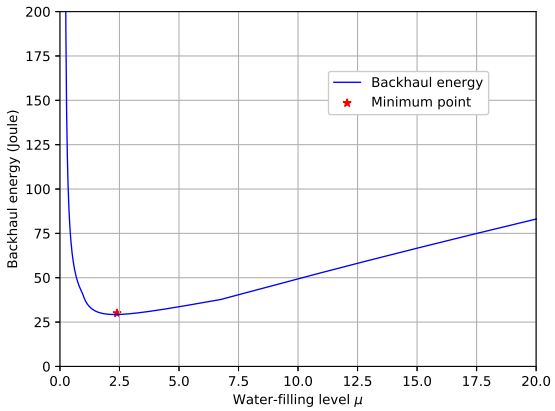

Figure 3: Function graph of $\mathcal{F}(\mu)$.

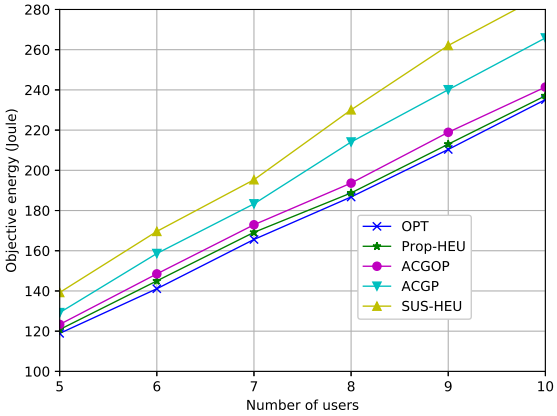

Figure 4: Objective energy vs. K ($T_{max} = 16s$).

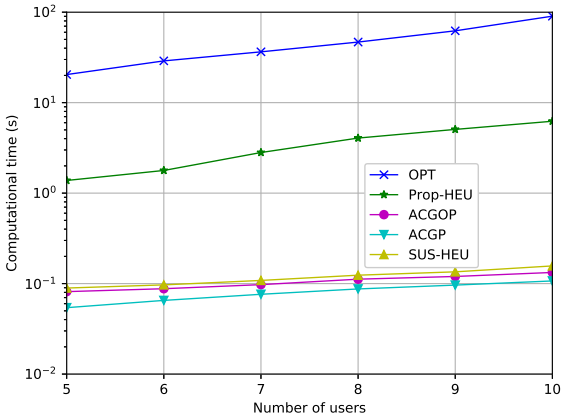

Figure 5: Computational time vs. K ($T_{max} = 16s$).

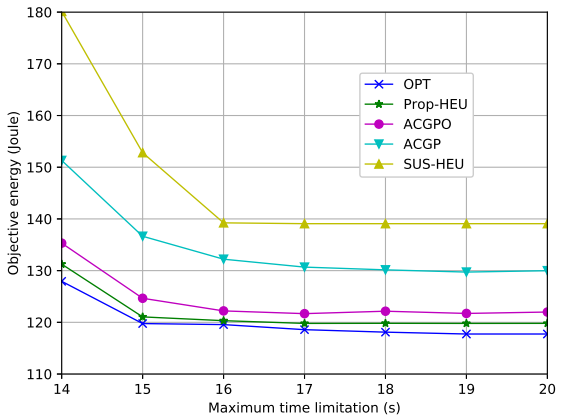

Figure 6: Objective energy vs. T_{max} ($K = 7$).

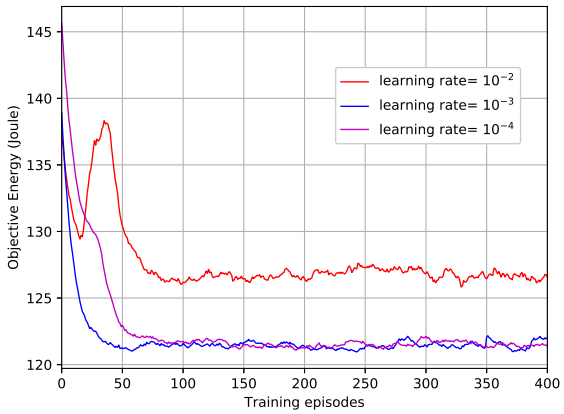

Figure 7: Objective energy with different learning rate.

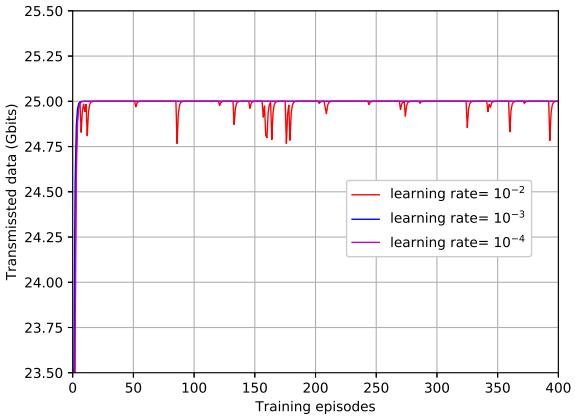

Figure 8: Transmitted data with different learning rate.

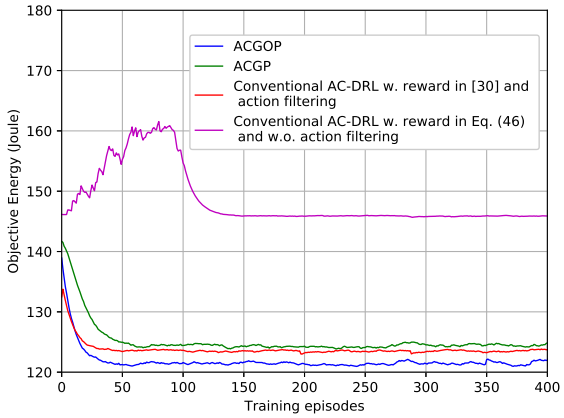

Figure 9: Objective energy with different AC-DRL methods.

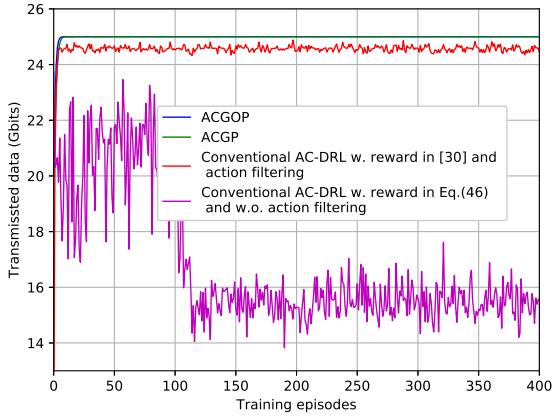

Figure 10: Transmitted data with different AC-DRL methods.

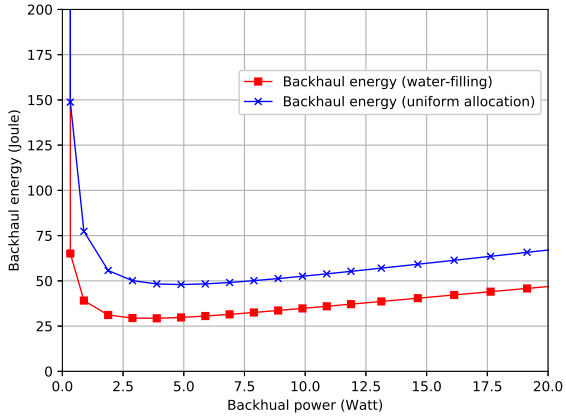

Figure 11: Backhaul energy vs. backhaul power.

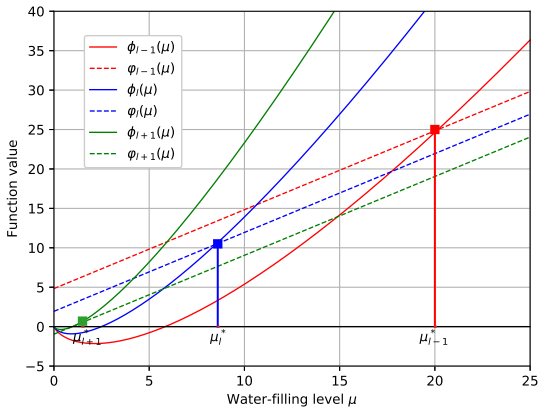

Figure 12: Function graphs of $\phi_l(\mu)$ and $\varphi_l(\mu)$.

Table 1: Parameters in ACGOP and ACGP

Parameters	Actor	Critic
Number of hidden layers	3	3
Number of nodes/layer	300	300
Activation function (hidden layers)	ReLU	ReLU
Activation function (output layer)	Sigmoid	None
Learning rate ρ	0.001	0.001
Loss function	Eq. (22)	Eq. (23)
Optimizer	Adam	Adam
Batch size	64	64
Discount factor γ	0.9	
Size of repository	10,000 tuples	
Number of learning episodes	400	
Software platform	Python 3.6 with TensorFlow 0.12.1	

Figures

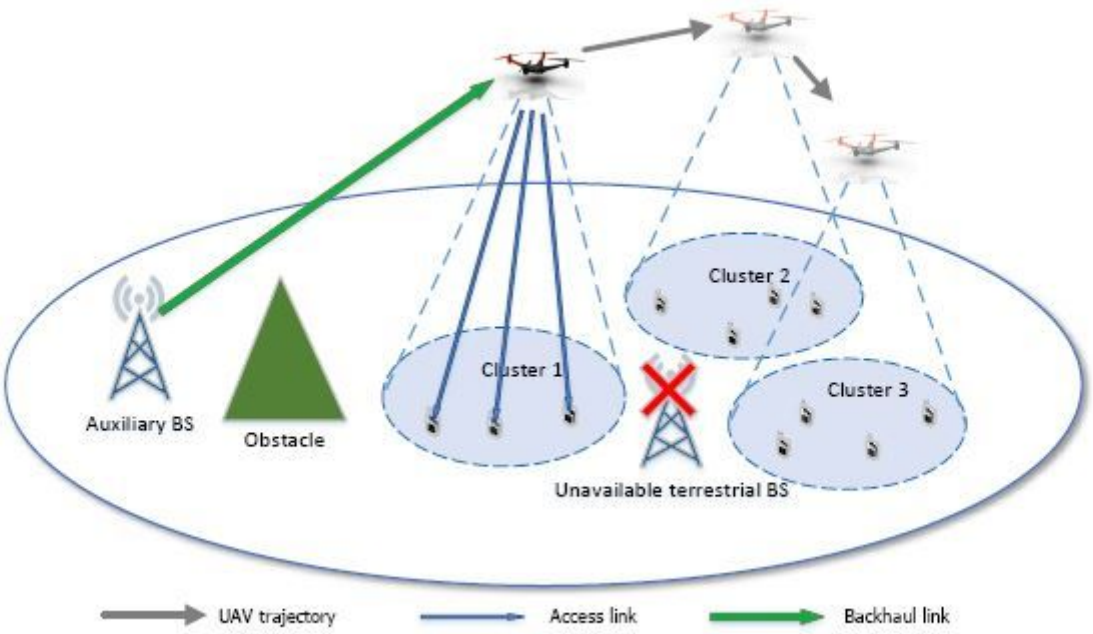

Figure 1

A UAV network with N = 3 clusters.

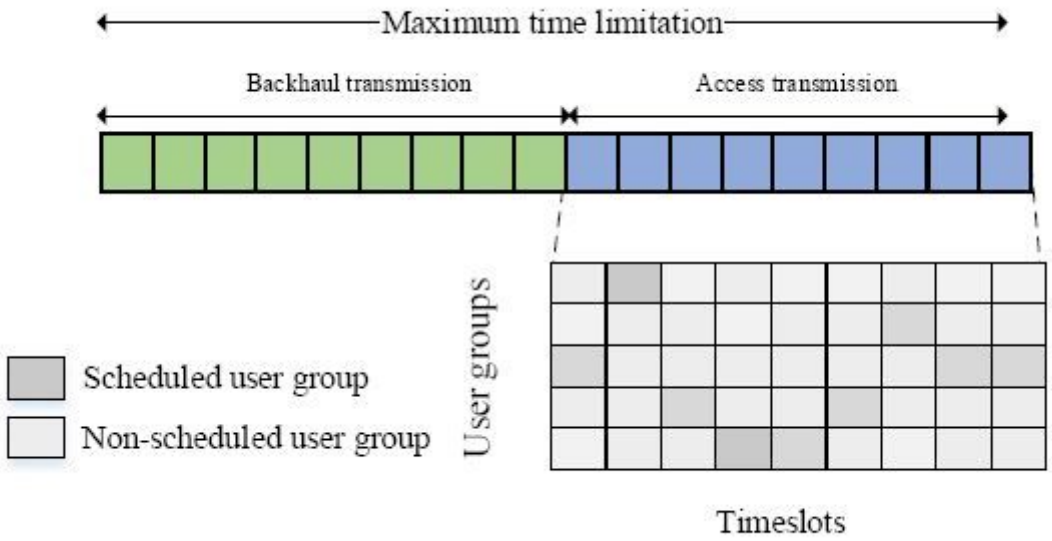

Figure 2

An illustration of the structure of the transmission process.

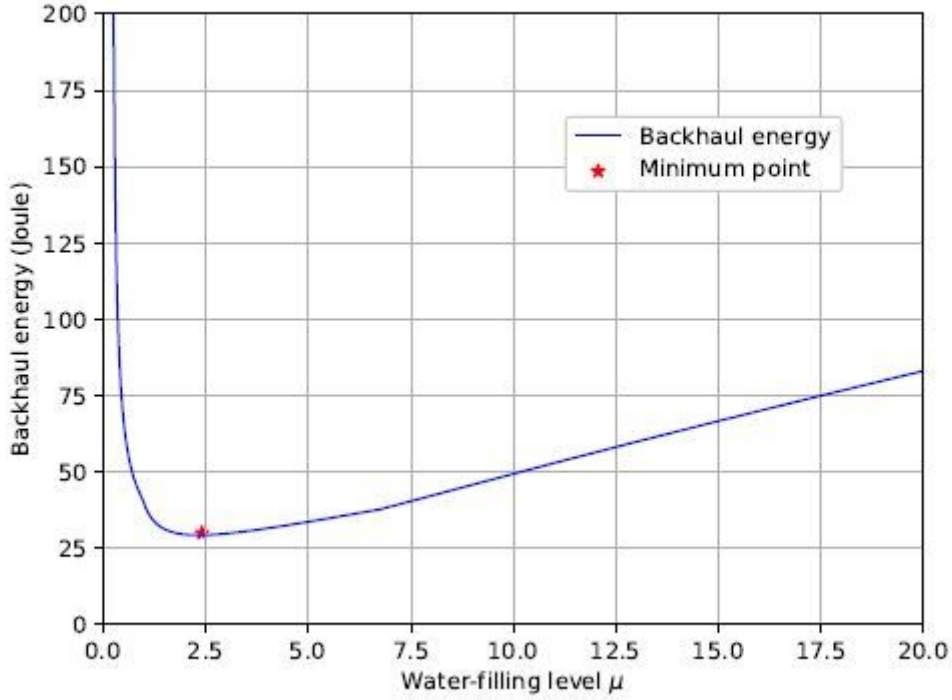

Figure 3

Function graph of F(μ).

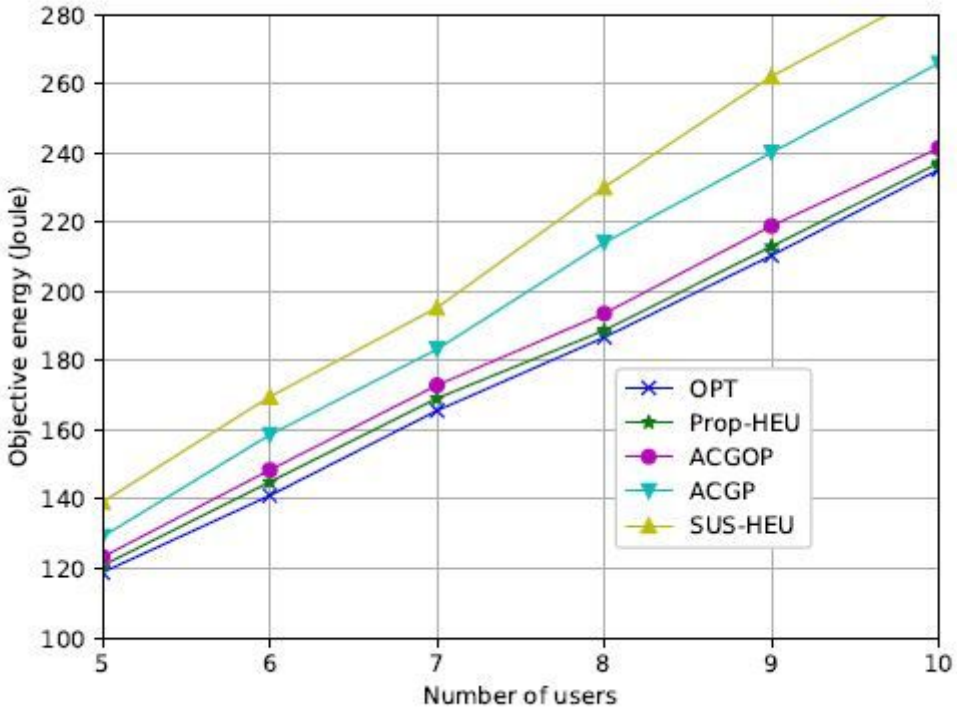

Figure 4

Objective energy vs. K (Tmax = 16s).

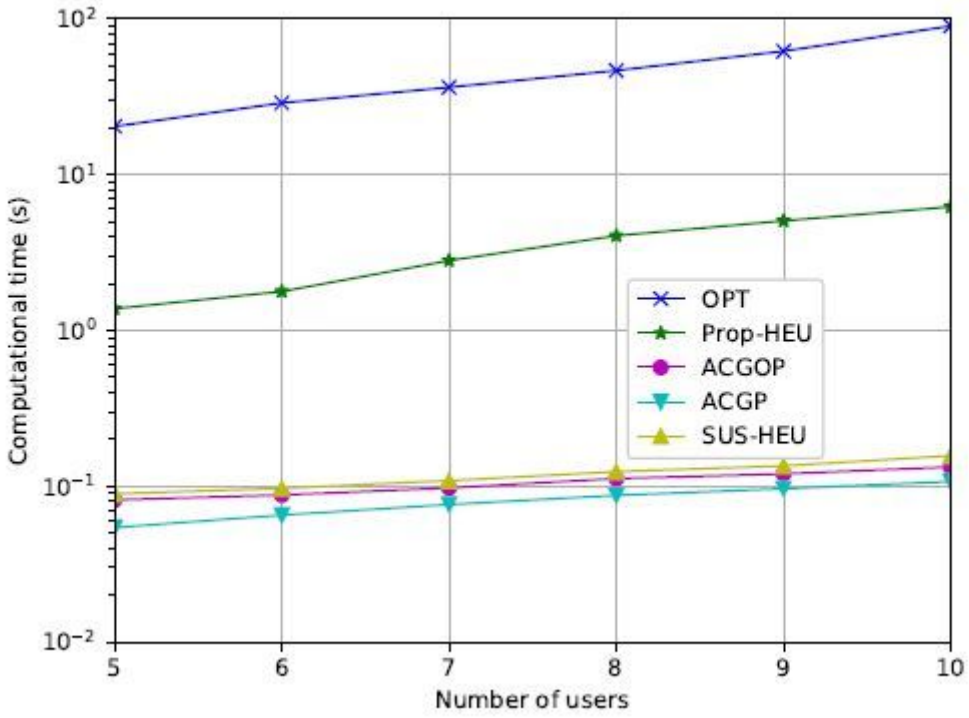

Figure 5

Computational time vs. K (Tmax = 16s).

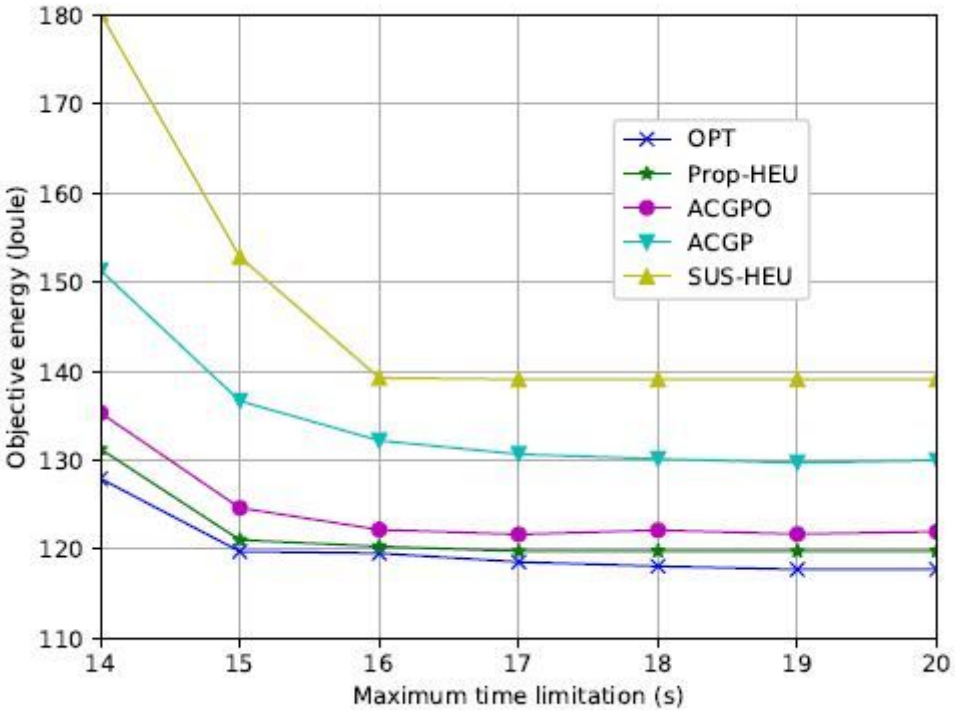

Figure 6

Objective energy vs. Tmax (K = 7).

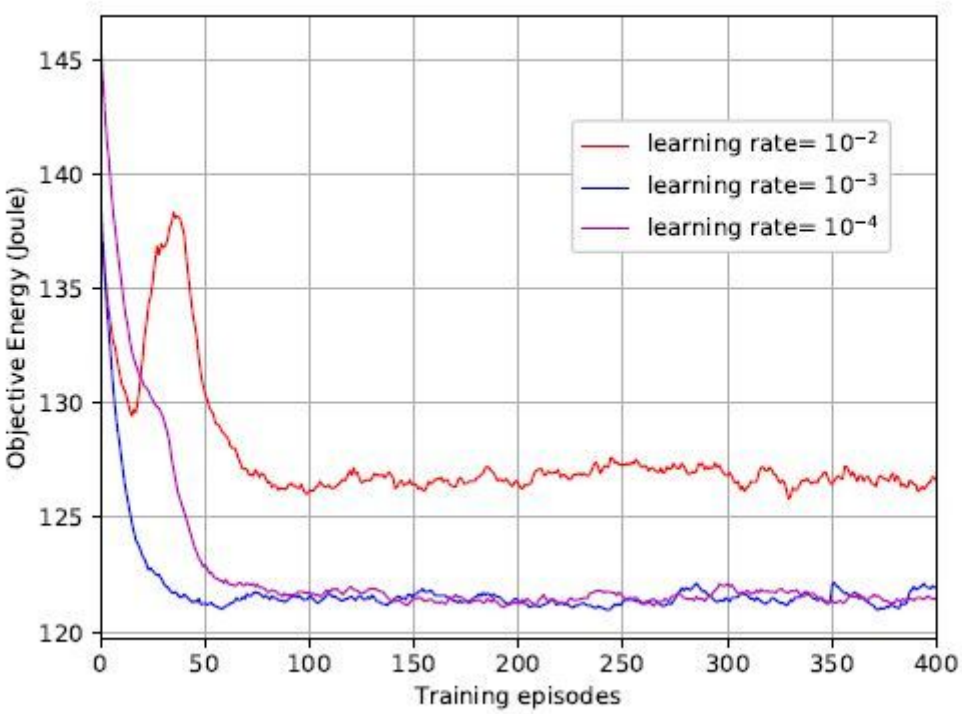

Figure 7

Objective energy with different learning rate.

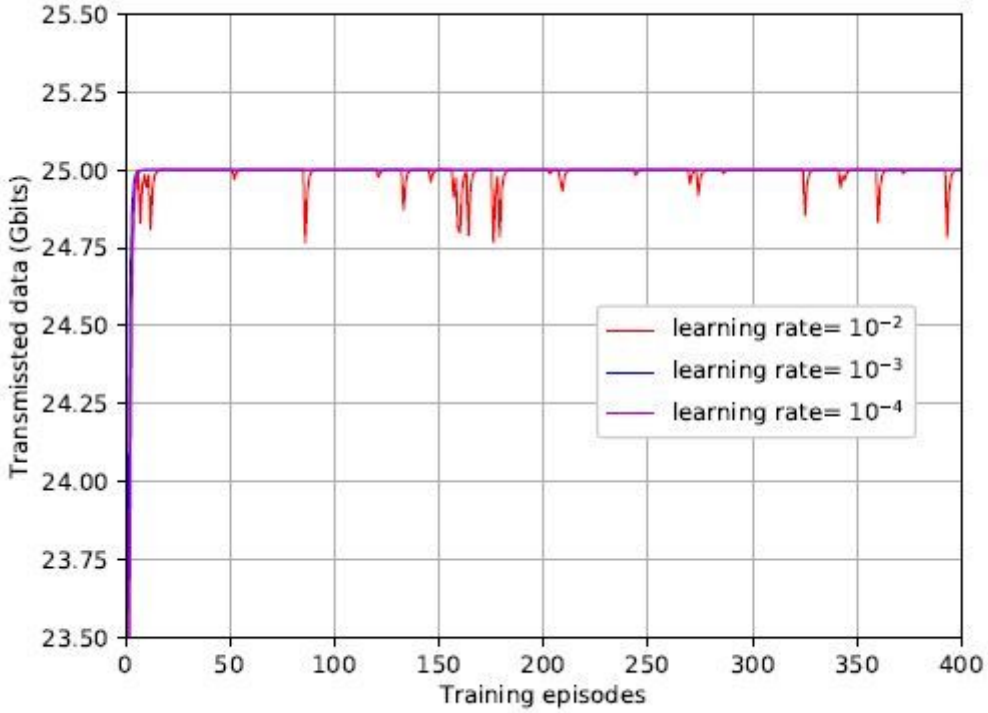

Figure 8

Transmitted data with different learning rate.

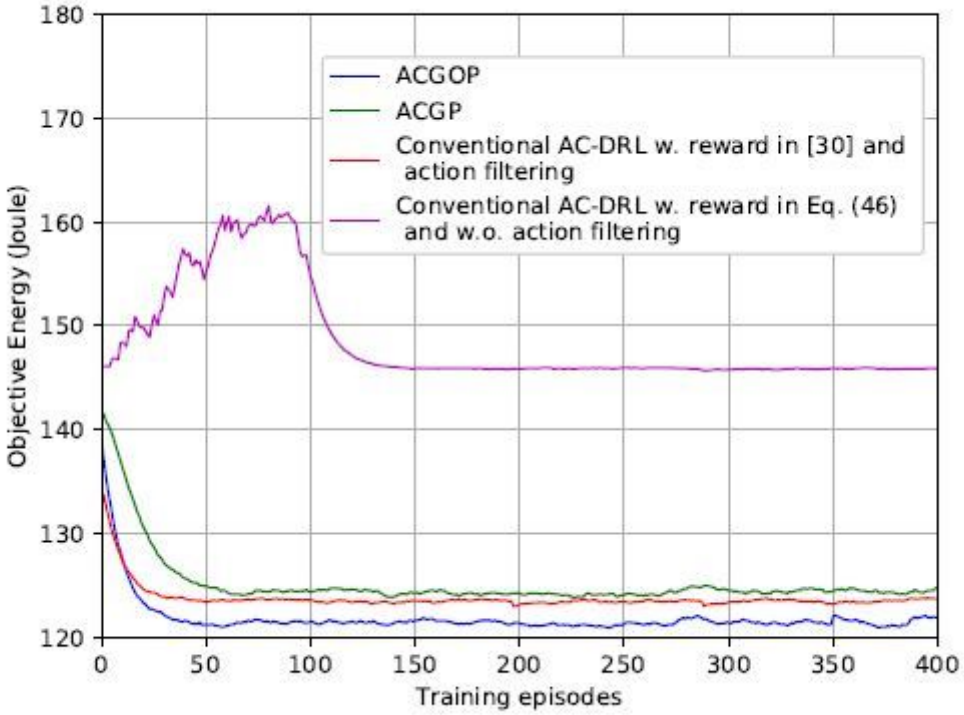

Figure 9

Objective energy with different AC-DRL methods.

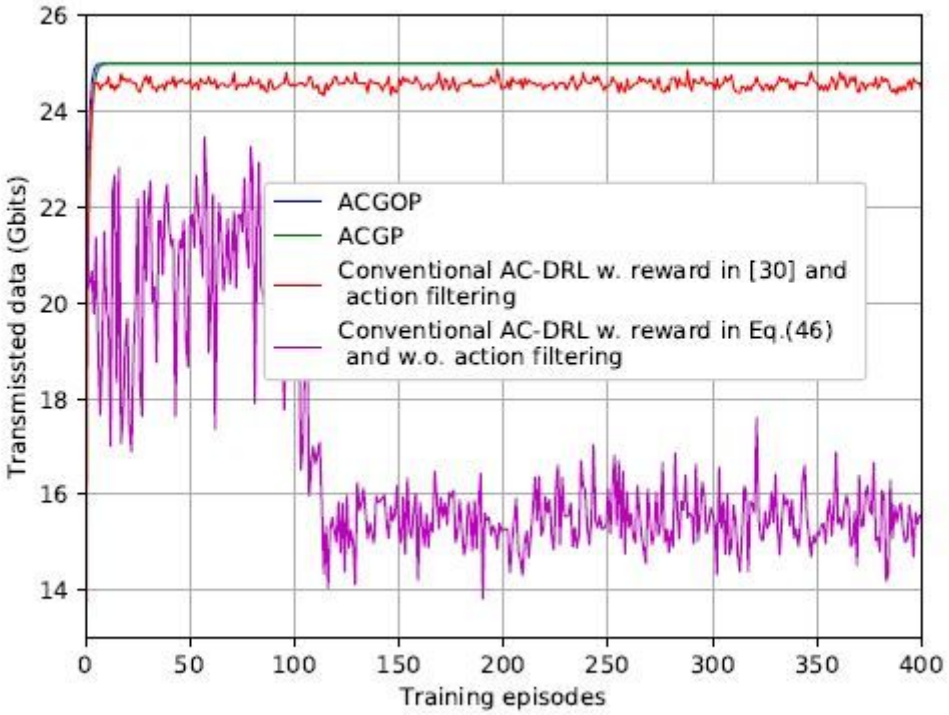

Figure 10

Transmitted data with different AC-DRL methods.

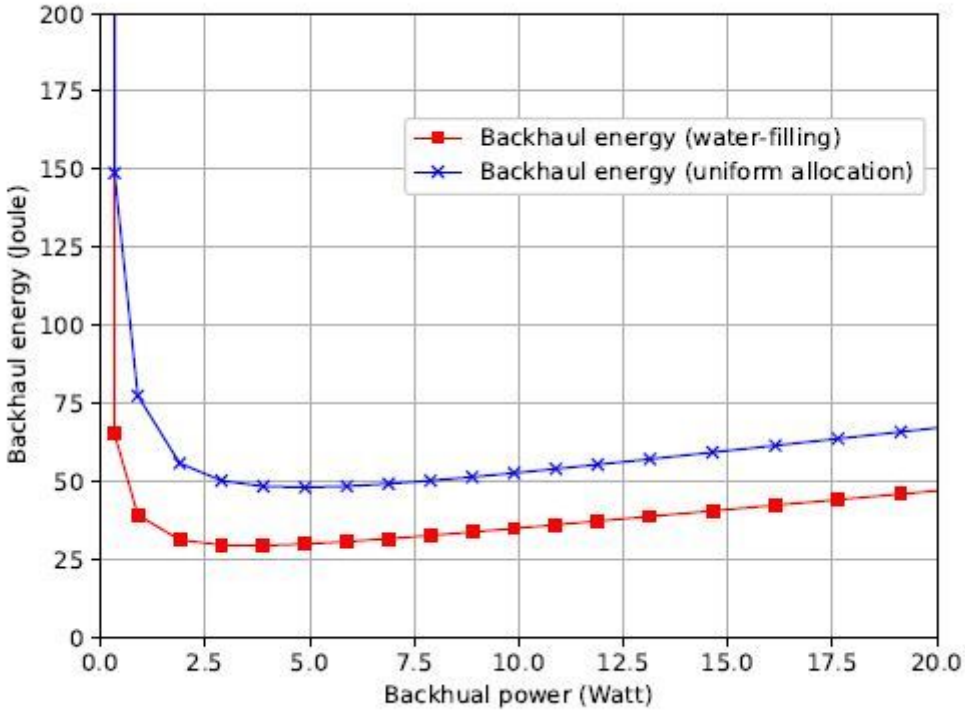

Figure 11

Backhaul energy vs. backhaul power.

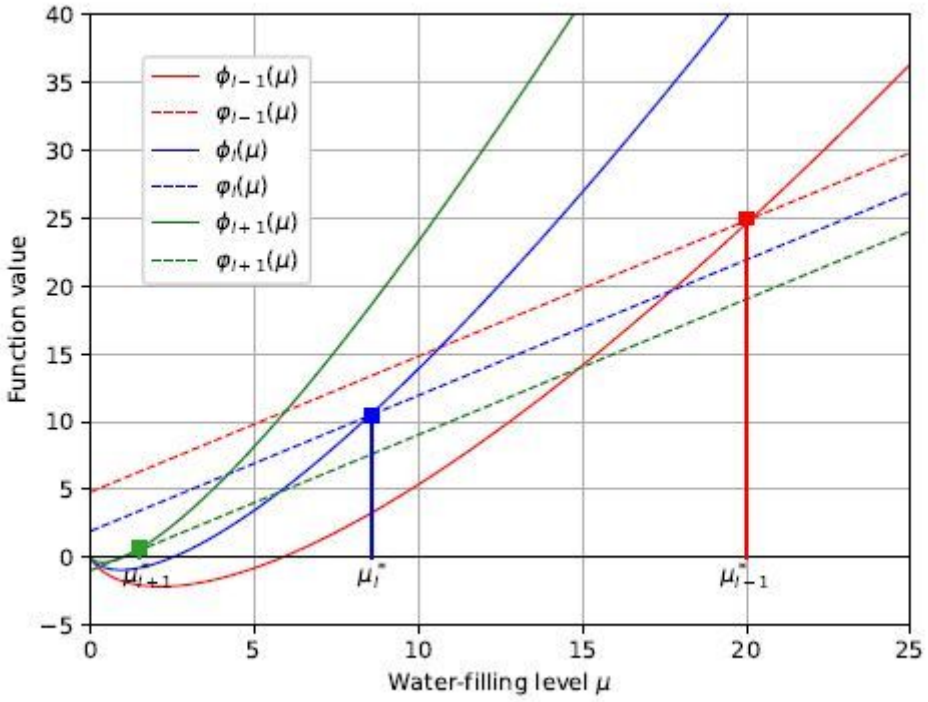

Figure 12

Function graphs of l(μ) and φl(μ).