Reinforcement Learning Based Optimization on Energy Efficiency in UAV Networks for IoT

Dan Deng, *Senior Member, IEEE,* Junxia Li, Rutvij H. Jhaveri, Prayag Tiwari, Muhammad Fazal Ijaz, Jiangtao Ou and Chengyuan Fan

***Abstract*—The combination of Non-Orthogonal Multiplex Access and Unmanned Aerial Vehicles (UAV) can improve the energy efficiency (EE) for Internet-of-Things (IoT). On the condition of interference constraint and minimum achievable rate of the secondary users, we propose an iterative optimization algorithm on EE. Firstly, with given UAV trajectory, the Dinkelbach method based fractional programming is adopted to obtain the optimal transmission power factors. By using the previous power allocation scheme, the successive convex optimization algorithm is adopted in the second stage to update the system parameters. Finally, reinforcement learning based optimization is introduced to obtain the best UAV trajectory.**

***Index Terms*—Unmanned Aerial Vehicles; energy efficiency; Internet-of-Things (IoT); NOMA; power allocation optimization;**

I. Introduction

DUE to the rapid growth of wireless connections, both the total capacity and the spectrum efficiency of Internet-of-Things (IoT) are facing more and more challenges, including energy efficiency (EE) and QoS limitations. To tackle the bottleneck of the IoT, various new technologies are proposed and investigated in both academic and industrial areas [1]–[5].

Unmanned Aerial Vehicles (UAV) can be used in IoT to improve the system capacity [6], [7]. Considering energy harvesting at the user equipments, by using weight sum method, the authors in [8] introduced a solution to maximize the spectral efficiency. Focused on hybrid energy supplied networks, the authors in [9] proposed a novel wireless resource allocation scheme to improve the long-term mean EE. Considering the presence of an interference node, the suboptimal power splitting and energy allocation method is proposed in [10] to improve EE of cooperative relay networks. Furthermore, an EE and spectrum efficiency trade-off scheme is addressed in [11] for cooperative cognitive radio network, where mathematical analysis as well as the numerical results are present to show that about 19% gains can be obtained compared with the existing methods. Recently, the Dinkelbach method [12] based optimization on EE is introduced in [13] to solve the fractional programming, where the classical EE problem is transformed into q parameters iterative problem.

Some enabling technologies have been proposed and analyzed to improve the QoS in UAV networks [14], [15]. Considering both the coverage and the EE in UAV networks, with the help of potential games and Nash equilibrium, the authors in [16] developed a UAV deployment scheme to guarantee the energy-efficient coverage. Under the information causality constraint, the authors in [17] proposed the optimal energy-efficient UAV trajectory for relaying networks. By optimization on locations of backscattering UAV for future Internet-of-Things, the authors in [18] derived the system average outage probability by using Golden Section method. Considering the wireless powered UAV networks, by using of Q-learning method, ref. [19] proposed a learning based UAV route scheme for data collection.

Meanwhile, the deployment of Non-Orthogonal Multiplex Access (NOMA) networks is a promising solution for limited spectrum scenarios [20]–[23] in cognitive radio. Considering fairness of multiple users, energy-efficient power allocation is investigate in [24], where only statistical channel state information is used to obtain sub-optimal solution under the rate requirements as well as the power constraints. The authors in [25] formulated the EE optimization as a fractional programming problem, and proposed a solution method for the non-convex problem. Focusing on hybrid user pairing scheme, the authors in [26], [27] proposed a joint optimization algorithm on spectral efficiency and EE.

Furthermore, in the latest research literatures, machine learning has been adopted in [28] as an enabling algorithm in wireless networks. When the traffic load is high enough, the proposed scheme showed higher EE compared with classical method. Similarly, without prior knowledge, an online network selection algorithm is proposed in [29]–[31] for complex network environment. With imperfect successive interference cancellation (SIC) decoding receiver, both precoding and SIC decoding are jointly optimized in [32] based on deep neural networks. To maximizing the sum data rate as well as the EE, a learning based algorithm is proposed in [33] to tackle the challenge of massive connectivity. A deep reinforcement learning (RL) based optimization solution on power and channels for multi-carrier NOMA is proposed in [34].

Dan Deng is with Guangzhou Panyu Polytechnic, Guangzhou, 410630, China.(dengdan@ustc.edu)

Junxia Li is with Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, 454003, China. (lijunxia20020@163.com)

Rutvij H. Jhaveri is with Department of Computer Science and Engineering, Pandit Deendayal Energy University, India.

Prayag Tiwari is with Department of Computer Science, Aalto University, Finland. (prayag.tiwari@aalto.fi)

Muhammad Fazal Ijaz is with Department of Intelligent Mechatronics Engineering, Sejong University, South Korea.

Jiangtao Ou is with AI Sensing Technology, Foshan, China. (jiangtaoou@ieee.org)

Chengyuan Fan is with AI Sensing Technology, Chancheng District, Foshan, China, 528000. (chengyuanfan@ieee.org)

*The corresponding author is Junxia Li (lijunxia20020@163.com), Rutvij H. Jhaveri (rutvij.jhaveri@sot.pdpu.ac.in) and Muhammad Fazal Ijaz(fazal@sejong.ac.kr).

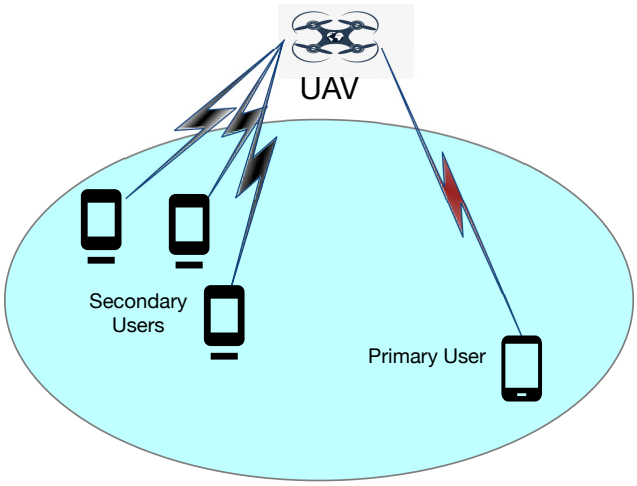

Fig. 1. System model of optimization on EE for UAV-NOMA.

The optimization method on power resource allocation and UAV flight trajectory for NOMA networks (UAV-NOMA) is investigated in this paper with interference constraint and minimum achievable rate constraint. In the considered wireless networks, sharing the spectrum of the primary user by using cognitive radio, the UAV node sends multiple data streams simultaneously to its users. By optimization on the UAV movement and the transmission power as well as the power resource allocation scheme, the optimal average EE can be obtained. To solve the difficult problem, an iterative reinforcement learning based algorithm on the resource allocation and the flying trajectory is proposed. In the first stage, with given UAV trajectory, the Dinkelbach method based fractional programming is adopted to get the optimal transmission power as well as the power allocation scheme. Afterwards, by using previous power allocation scheme, the successive convex optimization algorithm is used in the second stage to update the system parameters. Finally, reinforcement learning based optimization is introduced to obtain the optimal UAV movement directions in terms of mean EE. The contributions of this paper are summarized as follows:

1) Under the constraints of maximum interference power and and minimum achievable rate, we investigate the joint optimization algorithm on UAV flight trajectory and power allocation for UAV-NOMA networks in the presence of multiple secondary users.
2) Considering the interference constraint as well as the minimum achievable rate of the secondary users, Dinkelbach fractional programming and successive convex optimization algorithm are adopted to solve the fractional resource allocation problem, and RL-based update algorithm is introduced to solve the UAV trajectory optimization problem.

II. Network Model

The system model of UAV-NOMA is shown in Fig. 1, where a primary user, K secondary users and a single UAV node share the same wireless spectrum as a cognitive network. It is assumed that all wireless links follow line of sight propagation, and all nodes are configured with only one antenna. Because of the spectrum sharing protocol, the signals transmitted by the UAV node will interfere with the primary user. To guarantee the normal data transmission of the primary users, UAV node must strictly control its own transmit power. Using NOMA and SIC receiver, UAV node can serve multiple secondary users simultaneously. Specifically, all data streams are added together with different weights at the transmission node. Accordingly, data streams are decoded layer by layer at each secondary user.

To simplify the following analysis, the continuous trajectory of UAV is divided by equal time slots. The flight altitude of the UAV is fixed as H, and $\mathbf{W} = \{\mathbf{w}[n] = [x[n], y[n]]^T\}$ denotes the trajectory of the UAV node. The total time period of the trajectory is denoted as T, the number of locations is N, and the time interval is $\delta = T/N$.

Let L_p and $\mathbf{L}_i = [x_i, y_i]^T$ denote the the locations of primary users and the ith secondary user, respectively. Thus, the channel fading power of the destination users is

$$\begin{aligned} g_p[n] &= \rho d_p^{-\alpha}[n] \\ &= \rho(H^2 + ||\mathbf{w}[\mathbf{n}] - \mathbf{L}_p[n]||^2)^{-\alpha/2}, \end{aligned} \tag{1}$$

and

$$\begin{aligned} g_i[n] &= \rho d_i^{-\alpha}[n] \\ &= \rho(H^2 + ||\mathbf{w}[\mathbf{n}] - L_i||^2)^{-\alpha/2}, \end{aligned} \tag{2}$$

where ρ is the reference power gain with $d_i = 1m$, and $\alpha \geq 2$ is the path loss exponent.

Note that line-of-sight channel model without multi-path effect is adopted in this paper, which results in good agreement with the actual situation for UAV networks, and this assumption is widely used in existing literatures, such as [14], [35].

Consequently, the downlink NOMA signal transmitted from the UAV node is expressed as

$$x = \sum_{i=1}^{K} \sqrt{\xi_i P} x_i, \tag{3}$$

where P denotes the total signal power, x_i is the original information to the ith secondary user, and ξ_i denotes the power allocation coefficient for x_i. Moreover, $\xi[n]$ is used to represent the power allocation coefficients for nth time slot, i.e., $\xi[n] = [\xi_1[n], \xi_2[n], ..., \xi_K[n]]^T$. and $\mathbf{\Xi}$ means the total power allocation factors in all time slots as $\mathbf{\Xi} = \{\xi[n], n \in [1, N]\}$. In order to satisfy the power constraint, we obtain

$$\sum_{i \in \Omega} \xi_i[n] = 1. \tag{4}$$

Then, the received wireless signal at the destination users can be given by

$$y_i = \sqrt{g_i} x + n_i = \sqrt{g_i} \sum_{i \in \Omega} \sqrt{\xi_i P} x_i + n_i, \tag{5}$$

where $n_i \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise. Without losing generality, we assume that the channel gain

increases by index, that is $g_1 \leq g_1 \leq ... \leq g_K$. Since the classical NOMA protocol is adopted, the power allocation factors decrease by index, $\xi_1 \geq \xi_2 \geq ... \geq \xi_K$.

III. Problem Formulation

Since ideal interference cancellation receiver is used in all nodes, the receiver decodes all data stream one by one according to the order of the allocation power factors. In addition, the signal noise radio (SNR) of the *k*th data stream is calculated as

$$\begin{aligned}\gamma_{i,k}[n] &= \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{g_i[n]P[n]}} \\ &= \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{\rho_o P[n]}(H^2 + ||\mathbf{w}[n] - \mathbf{L}_i||^2)^{\alpha/2}},\end{aligned} \tag{6}$$

where $\eta_k[n]$ represent the interference items at the *k*th secondary user in the *n*th time slot, i.e.,

$$\eta_k[n] = \sum_{j=k+1}^{K} \xi_j[n], \forall n \in [1, N]. \tag{7}$$

From (6), we can see that $\gamma_{i,k}[n]$ is a monotone increasing function. Since $g_i[n]$'s are in an ascending order, we have $\gamma_{i,k}[n] \leq \gamma_{j,k}[n], \forall i \leq j$. According to the SIC receiver, the decoding order is the same with the channel power gains order. Then, the achievable rate of the *k*th data stream is

$$R_{k,k}[n] = \log_2(1 + \gamma_{k,k}[n]), \forall k \in [1, K], n \in [1, N]. \tag{8}$$

Meanwhile, due to the QoS requirement for each secondary user, we have the following SNR constraint

$$R_{k,k}[n] \geq C_{th}, \forall k \in [1, K], n \in [1, N]. \tag{9}$$

Since $R_{k,k}[n]$ is a monotone increasing function of $\gamma_{k,k}[n]$, (9) is equivalent as

$$\gamma_{k,k}[n] \geq \Gamma, \forall k \in [1, K], \forall n \in [1, N], \tag{10}$$

with $\Gamma = 2^{C_{th}} - 1$.

The interference power received by the primary user is

$$I_p[n] = P[n]g_p[n] = \frac{P[n]\rho}{(H^2 + ||\mathbf{w}[\mathbf{n}] - \mathbf{L}_p[n]||^2)^{\alpha/2}}. \tag{11}$$

In each time slot, the sum of achievable transmission rates of all secondary users are

$$R[n] = \sum_{k=1}^{K} R_{k,k}[n] = \sum_{k=1}^{K} \log_2(1 + \gamma_{k,k}[n]). \tag{12}$$

Thus, the EE of nth time slot is

$$EE[n] = \frac{R[n]}{P[n]}. \tag{13}$$

In the following section, our purpose is to maximize the average EE of NOMA assisted UAV networks with interference constraints. Thus, the optimization problem on EE can be expressed as

$$\begin{aligned}
&\text{(P0)}: \max_{\mathbf{w},\xi,P} \frac{1}{N}\sum_{n=1}^{N} EE[n] && (14)\\
&s.t. \quad 0 \leq \xi_i[n] \leq 1, && (14a)\\
&\sum_{i\in\Omega} \xi_i[n] = 1, && (14b)\\
&||\mathbf{w}[n] - \mathbf{w}[n+1]||^2 \leq (v_m\delta)^2, && (14c)\\
&\gamma_{k,k}[n] \geq \Gamma, && (14d)\\
&I_p[n] \leq I_{th}, && (14e)
\end{aligned}$$

where v_m denotes the maximum flying speed.

IV. Joint Optimization Algorithm

In this section, we will introduce the RL-based solution for (P0). Firstly, with given UAV trajectory, the Dinkelbach method based fractional programming is adopted to give the solution of the transmission power as well as the power allocation scheme. Afterwards, using the previous power allocation scheme, the successive convex optimization algorithm is adopted in the second stage to update the system parameters. Finally, reinforcement learning based optimization is introduced to obtain the best UAV trajectory in terms of mean EE.

A. Subproblem with given UAV trajectory

Firstly, with given UAV trajectory, we focus on the EE in each time slot, i.e. $EE[n]$ as defined in (13). With definition that

$$R[n] = \sum_{k=1}^{K} \log_2(1 + \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{g_k[n]P[n]}}), \tag{15}$$

and introducing variables $\eta[n]$, the optimal subproblem can be rewritten as

$$\begin{aligned}
&\text{(P1)}: \max_{\xi[n],\eta[n],P[n]} EE[n] = \frac{R[n]}{P[n]} && (16)\\
&s.t. \quad 0 \leq \xi_i[n] \leq 1, && (16a)\\
&\sum_{i\in\Omega} \xi_i[n] = 1, && (16b)\\
&\eta_k[n] = \sum_{j=k+1}^{K} \xi_j[n], && (16c)\\
&\gamma_{k,k}[n] \geq \Gamma, && (16d)\\
&P[n] \leq I_{th}/g_p[n]. && (16e)
\end{aligned}$$

Since the objective function in (16) is a nonlinear fractional optimization problem, it is difficult to be solved. The Dinkelbach algorithm [12] can be adopted to solve the problem with some mathematical transforms.

Proposition 1. The numerator of $EE[n]$, i.e., $R[n]$, is concave w.r.t. $\xi_k[n]$ and $P[n]$, and convex w.r.t. $\eta_k[n]$, respectively.

Proof: Consider a continuous function defined as follows

$$w = \ln(1 + \frac{x}{y}), x > 0, y > 0. \tag{17}$$

Applying first-order partial derivative on w, we obtain

$$\frac{\partial w}{\partial x} = \frac{1}{(1 + \frac{x}{y})}, \tag{18}$$

and

$$\frac{\partial w}{\partial y} = \frac{-x}{y^2 + xy}. \tag{19}$$

Furthermore, applying second-order partial derivative on w, yields

$$\frac{\partial^2 w}{\partial x^2} = \frac{-1}{(1 + \frac{x}{y})^2} < 0, \tag{20}$$

and

$$\frac{\partial^2 w}{\partial y^2} = \frac{x(x + 2y)}{(y^2 + xy)^2} > 0. \tag{21}$$

Consider another continuous function

$$w = \ln(1 + \frac{a}{b + c/z}), z > 0. \tag{22}$$

Similarly, w is concave w.r.t. z. According to the definition of $R[n]$, we have

$$\begin{aligned} R[n] &= \sum_{k=1}^{K} \log_2(1 + \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{g_i[n]P[n]}}) \\ &= \frac{1}{\ln 2} \sum_{k=1}^{K} \ln(1 + \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{g_i[n]P[n]}}) \end{aligned} \tag{23}$$

Similar with (21) and (22), with given $g_i[n]$, it is concluded that $R[n]$ is a concave function w.r.t. $\xi_k[n]$ and $P[n]$, and is convex w.r.t. $\eta_k[n]$.

This is the end of the proof.

Applying the first-order partial derivative on $R_{k,k}[n]$ in (8), yields

$$\begin{aligned} &\frac{\partial R_{k,k}[n]}{\partial \eta_k[n]} \\ &= \frac{1}{\ln 2} \frac{-\xi_k[n]}{[(\eta_k[n] + \frac{\sigma^2}{g_k[n]P[n]})^2 + \xi_k[n](\eta_k[n] + \frac{\sigma^2}{g_k[n]P[n]})]}. \end{aligned} \tag{24}$$

According to *Proposition 1*, $R_{k,k}[n]$ is a convex functions of $\eta_k[n]$. By using first-order Taylor expansion on $R_{k,k}[n]$, fields

$$\begin{aligned} R_{k,k}[n] &= \frac{1}{\ln 2} \ln(1 + \frac{\xi_k[n]}{\eta_k[n] + \frac{\sigma^2}{g_k[n]P[n]}}) \\ &\geq \frac{1}{\ln 2} \ln(1 + \frac{\xi_k[n]}{\eta_k^r[n] + \frac{\sigma^2}{g_k[n]P[n]}}) + \nabla_k^r(\eta_k[n] - \eta_k^r[n]) \\ &= R_k^L[n], \end{aligned} \tag{25}$$

where $R_k^L[n]$ denotes the lower bound of $R_{k,k}[n]$, and the definition of ∇^r is given as follows

$$\nabla_k^r = \frac{1}{ln2} \frac{-\xi_k^r[n]}{[(\eta_k^r[n] + \frac{\sigma^2}{g_k[n]P^r[n]})^2 + \xi_k^r[n](\eta_k^r[n] + \frac{\sigma^2}{g_k[n]P^r[n]})]}, \tag{26}$$

where $\xi_k^r[n], \eta_k^r[n], P^r[n]$ are the previous rth iteration output results. Note that $R_k^L[n]$ is concave of $\xi_k[n]$ and $P[n]$, respectively, and is linear w.r.t. $\eta_k[n]$, respectively. By substituting (11) into (16d), the maximum value of $P[n]$ is

$$\eta_k[n] + \frac{\sigma^2}{g_k[n]P[n]} \leq \frac{\xi_k[n]}{\Gamma}. \tag{27}$$

Note that (27) is a convex function of $P[n]$. Thus, given rth iteration output results $\xi_k^r[n], \eta_k^r[n], P^r[n]$, the optimal problem of $(r+1)$th iteration is formulated as

$$\text{(P2)}: \max_{\xi[n], \eta[n], P[n]} EE^L[n] = \frac{\sum_{k=1}^{K} R_k^L[n]}{P[n]} \tag{28}$$

$$s.t. \quad (16a), (16b), (16c), (16e), (27). \tag{28a}$$

Since it is difficult to solve (P1), we present another approximated solution as in (P2). Note that the cost of the transform from (P1) to (P2) is that first-order Taylor expansion is adopted on the achievable rate as in (25), and the approximated version $R_k^L[n]$ is used instead of the accurate achievable rate $R_{k,k}(n)$. The similar method is widely used in existing works, such as [13], [36], and simulation results show that the approximation error is negligible.

B. Fractional Programming Using Dinkelbach Method

We can see that all of the constraints in (28) are linear, the numerator is concave w.r.t. $\xi[n], \eta[n], P[n]$, and the denominator of the objective function is linear. Thus, the Dinkelbach optimization algorithm [12], [13] can be used to solve the problem in (28).

Proposition 2. Let S denoting a compact subset of E^n, where n is the dimension of Euclidean space. Let $N(x)$ and $D(x)$ are continuous real-value functions, with assumption as follows

$$D(x) > 0, \forall x \in S. \tag{29}$$

Given the following fractional equation

$$\max_{x \in S} \frac{N(x)}{D(x)}, \tag{30}$$

and

$$F(q) = \max_{x \in S}[N(x) - qD(x)], \text{for } q \in E^1. \tag{31}$$

we have

1) $F(q)$ is continuous and convex over $q \in E^1$.
2) $F(q)$ is strictly monotone decreasing w.r.t. q.
3) Let

$$q_0 = \frac{N(x_0)}{D(x_0)} = \max_{x \in S} \frac{N(x)}{D(x)}, \tag{32}$$

if and only if

$$\begin{aligned} F(q_0) &= F(q_0, x_0) \\ &= \max_{x \in S}[N(x) - q_0 D(x)] = 0. \end{aligned} \tag{33}$$

Proof. The detailed proof is present in [12].

Furthermore, if $N(x)$ is concave and $D(x)$ is convex, the nonlinear programming in (30) can be solved by iterative Dinkelbach algorithm as follows.

1) Initialization with iteration index $t = 0$, $q_0 = 0$, maximum iteration number T and convergence threshold δ, definition $F(q, x) = N(x) - qD(x)$.
2) Given q_t, find the solution of $x_t = \arg\max_x F(q_t, x)$ by using classical convex optimization tools.
3) If $[(F(q_t, x_t) > \delta)$ and $(t < T)]$ then update parameters with $t = t + 1$ and $q_t = \frac{N(x_t)}{D(x_t)}$, goto step (b).
4) Output the solution x_t and calculate the programming result with $q(x_t) = \frac{N(x_t)}{D(x_t)}$.

With definition that

$$F(q, \xi[n], \eta[n], P[n]) = \sum_{k=1}^{L} R_k^L[n] - q \cdot P[n] \tag{34}$$

the programming problem in (28) can be rewritten as

$$\text{(P3)}: \max_{\xi[n], \eta[n], P[n]} F(q, \xi[n], \eta[n], P[n]) \tag{35}$$

$$s.t. \quad (16a), (16b), (16c), (16e), (27). \tag{35a}$$

Since $R_k^L[n]$ is a concave function of $\xi_k[n]$ and $P[n]$, respectively, and is linear w.r.t. $\eta_k[n]$, respectively, it can be concluded that the programming problem in (P3) is convex. Moreover, the standard convex optimization tools, such as CVX [37], can be deployed to solve the problem.

By applying Dinkelbach algorithm on (P2) in (28), we can give the iterative optimization method as described in **Algorithm 1**.

Algorithm 1 Dinkelbach method based optimization algorithm for (P2)

1: Initialize the parameters with $t = 0$, $q_0 = 0$, maximum iterative number T_1 and convergence threshold τ_1.
2: **repeat**
3: With given q_t, use CVX to solve (P3) and obtain the optimal solution $\xi_{t+1}[n], \eta_{t+1}[n], P_{t+1}[n]$.
4: Update iteration index: $t = t+1$ and update parameters with $q_{t+1} = \frac{R_{t+1}^L[n]}{P_{t+1}[n]}$.
5: **until** $t \geq T_1$ or $F(q_{t+1}, \xi_{t+1}[n], \eta_{t+1}[n], P_{t+1}[n]) \leq \tau_1$.
6: Output the results $q_{t+1}, \xi_{t+1}[n], \eta_{t+1}[n], P_{t+1}[n]$.

According to [12], Algorithm 1 is convergent. Followed by the complexity analysis in [38], the computational complexity of Algorithm 1 is

$$C_1 = \mathcal{O}\{T_1 \left(N(3K+2)\right)^4 \left(N(K+1)\right)^{1/2} \log(1/\epsilon)\}, \tag{36}$$

where ϵ denotes the solution accuracy of CVX tools.

C. Iterative Optimization Algorithm for (P1)

Since the solution of (P2) can be obtained by Algorithm 1, we can summary the detailed solution of (P1) by using successive convex optimization algorithm as **Algorithm 2**.

Moreover, in each iterative process of Algorithm 2, we can see that $EE_{r+1}[n] \geq EE_r[n], \forall r, \forall n$. In another hand, $EE_{r+1}[n]$ is obviously upper bounded. Thus, we conclude that Algorithm 2 is convergent.

The computational complexity of Algorithm 2 can be given by

$$\begin{aligned} C_2 &= T_2 C_1 \\ &= \mathcal{O}\{T_2 T_1 \left(N(3K+2)\right)^4 \left(N(K+1)\right)^{1/2} \log(1/\epsilon)\}. \end{aligned} \tag{37}$$

Algorithm 2 Successive convex optimization algorithm for (P1)

1: Initialize the parameters with $r = 0$, maximum iterative number T_2, convergence threshold τ_2 and initial solution $\xi_0[n], \eta_0[n], P_0[n]$, calculate the objective function value $EE_r[n]$ as in (16).
2: **repeat**
3: With given $\xi_r[n], \eta_r[n], P_r[n]$, use Algorithm 1 to solve problem (P2) and obtain the optimal solution $\xi_{r+1}[n], \eta_{r+1}[n], P_{r+1}[n]$.
4: Update iteration index $r = r + 1$ and calculate the increase of the objective function $\chi_r = EE_{r+1}[n] - EE_r[n]$.
5: **until** $t \geq T_2$ or $\chi_r \leq \tau_2$.
6: Output the solution $\xi_{r+1}[n], \eta_{r+1}[n], P_{r+1}[n]$ and optimal objective function with $EE_{r+1}[n]$.

V. Reinforcement Learning Based Optimization on UAV Trajectory

Because the optimization on UAV trajectory is hard to solve directly, we turn to propose a Q-learning based method to search the action space and select the optimal action according to the reward function. Firstly, the action space is defined as follows: $A = \{0, 1, 2, 3, 4\}$; $\lambda = v_m \delta$ denotes the distance of each movement; the action zero, i.e., $a = 0$, indicates that the UAV keeps current position; and other actions, such as $a = 1, 2, 3, 4$, denote that the UAV moves to four different directions. Moreover, all users are located within a 2D space with size $M\lambda \times M\lambda$.

$$S = \{s_n = [x[n], y[n]]^T\}, \tag{38}$$

with

$$x[n] \in [0, M-1], y[n] \in [0, M-1]. \tag{39}$$

According to the optimization problem in (14), the reward function of each time slot R_a is

$$R_a = EE[n], \tag{40}$$

where the definition of $EE[n]$ is given in (16).

$$\begin{aligned} Q_{n+1}(s_n, a_n) = &(1-\theta) Q_n(s_n, a_n) \\ &+ \theta \Big[R_n + \beta \max_{a \in A} Q_n(s_{n+1}, a) \Big], \end{aligned} \tag{41}$$

where $\theta \in (0, 1]$ is the learning rate, and $\beta \in [0, 1]$ is the discount factor. In each step, the UAV action is updated

according to the ϵ-greedy policy, i.e., the optimal action is selected with probability ϵ, while the other actions are selected equally with total probability $(1-\epsilon)$. The detailed solution procedure for (P0) is given in **Algorithm 3**.

The computational complexity of Algorithm 3 is calculated by

$$\begin{aligned} C_3 &= 5M^2C_2 \\ &= \mathcal{O}\{5M^2T_2T_1\left(N(3K+2)\right)^4\left(N(K+1)\right)^{1/2}\log(1/\epsilon)\}. \end{aligned} \tag{42}$$

Furthermore, in each iterative process of **Algorithm 3**, we can see that $Q_{n+1}(s_n,a_n) \geq Q_n(s_n,a_n), \forall n$. On the other hand, $Q_{n+1}(s_n,a_n)$ is obviously upper bounded. Thus, we conclude that **Algorithm 3** is convergent.

Algorithm 3 Deep reinforcement learning optimization algorithm for (P0)

1: Initialize ξ, s, initialize $Q(s,a)$ with arbitrary value, and set iteration index $n=1$.
2: **repeat**
3: For each step of iterations
4: Employ ϵ-greedy policy to select action a_n, updated state s_{n+1}.
5: Given UAV location s_{n+1}, calculate the optimal power allocation factors $\xi[n+1]$ and transmission power $P[n+1]$ as in **Algorithm 2** .
6: Observe the reward R_a according to (40).
7: Update Q-table as in (41).
8: Update iteration index: $n=n+1$.
9: **until** n reaches the maximum iteration number T_3.

VI. Simulation Results

The proposed optimization method is verified by numerical results.

Without special comments, some parameters in reference configuration are listed as follows: the target area is a square with 100x100m; the number of the secondary users is $K=3$; the locations of the secondary users are $[20,20]m, [40,40]m$ and $[60,60]m$; the initial location is $w_0 = [80,0]m$; the location of the primary user is $[0,50]m$; the reference power gain is $\rho = 10^{-3}$; the UAV trajectory altitude is $H = 20m$; the path loss exponent is $\alpha = 2$; the iteration numbers are set as $T_1 = 8, T_2 = 3$; and the convergence thresholds are $\tau_1 = 10^{-3}, \tau_2 = 10^{-1}$, the threshold of the epsilon-greedy algorithm is $\epsilon = 0.9$; the discount factor is $\theta = 0.95$; the learning rate is $\beta = 0.4$; and the total number of time slots is set as $N = 20000$.

The convergence of algorithm 1 for (P2) is depicted in Fig. 2. We can see from this figure that Algorithm 1 converges rapidly w.r.t. the iteration index. When the iteration number is larger than 3, the EE becomes stationary, which can verify the convergence of the Dinkelbach method based iteration optimization algorithm in **Algorithm 1**. Also, the convergence performance of the RL-based optimization algorithm given in **Algorithm 3** is presented in Fig. 3. where the reference fading gain ρ changes from -30dB to -39dB. We can see from this

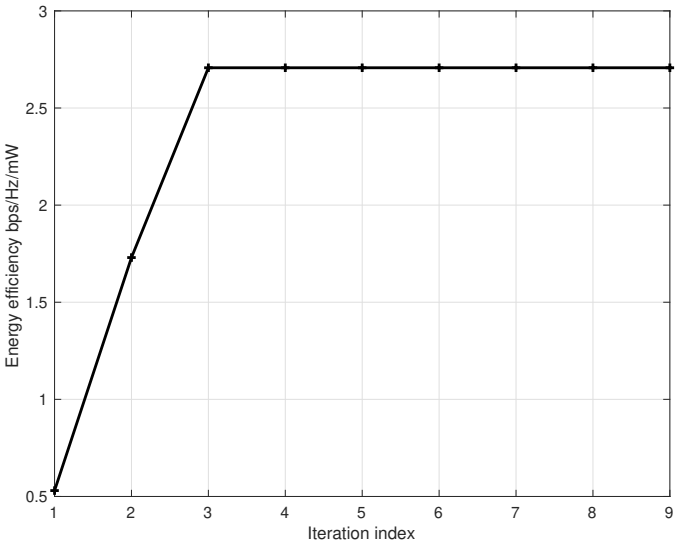

Fig. 2. Convergence performance of Algorithm 1 for (P2).

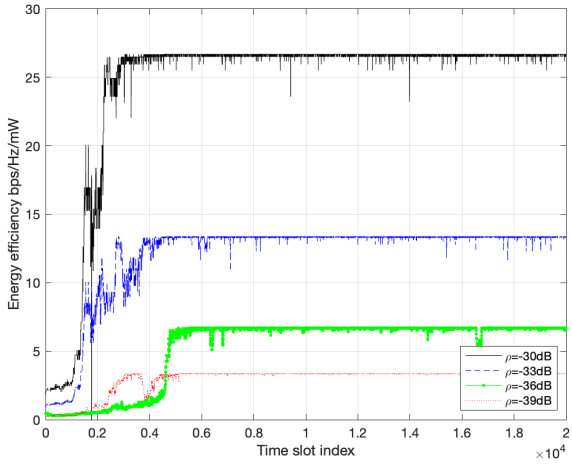

Fig. 3. Convergence performance of Algorithm 3 with different ρ.

figure that the energy efficiency increases rapidly for the first 5000 time slots and then remains stationary, which verifies the convergence performance of the proposed method.

The impacts of UAV altitude H on mean EE is present in Fig. 4. The UAV altitude H changes from $20m$ to $50m$. We can see that with $H = 20$m, when the time slot number is large enough, the mean EE reaches about 23 bps/Hz/mW. While with $H = 50m$, the mean EE drops quickly down to 7 bps/Hz/mW. A fixed trajectory scheme, which surrounds the boundary of the target area and uses the same power allocation algorithm as given in **Algorithm 1**, is adopted as a performance benchmark. The proposed algorithm shows remarkable performance gain compared with the baseline in all setup cases. Furthermore, the UAV trajectory with $H = 20$m is given in Fig. 5. One can see that by using of about 3000 explores, the UAV trajectory becomes stationary within an area of about $10m \times 10m$.

Fig. 6 shows the effects of capacity threshold C_{th} on

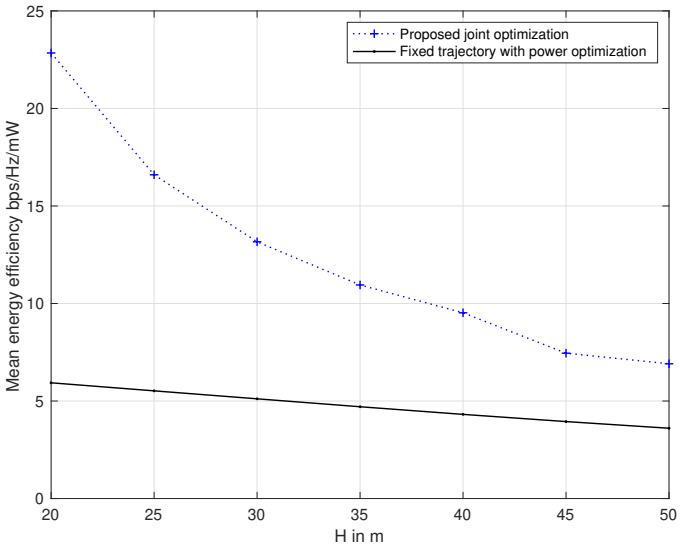

Fig. 4. Performance comparison on mean EE.

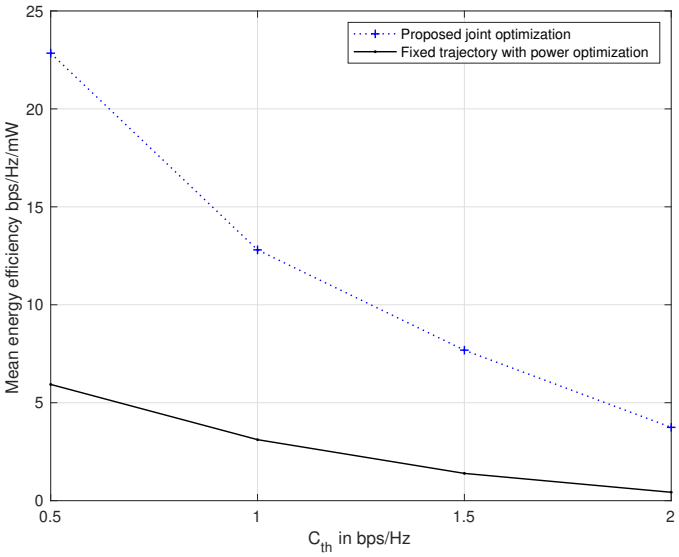

Fig. 6. Effects of capacity threshold C_{th} on mean EE.

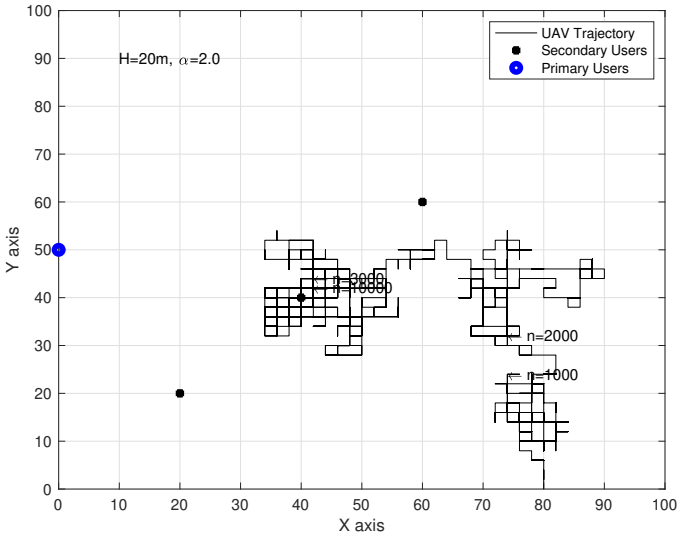

Fig. 5. UAV trajectory with $H = 20m$

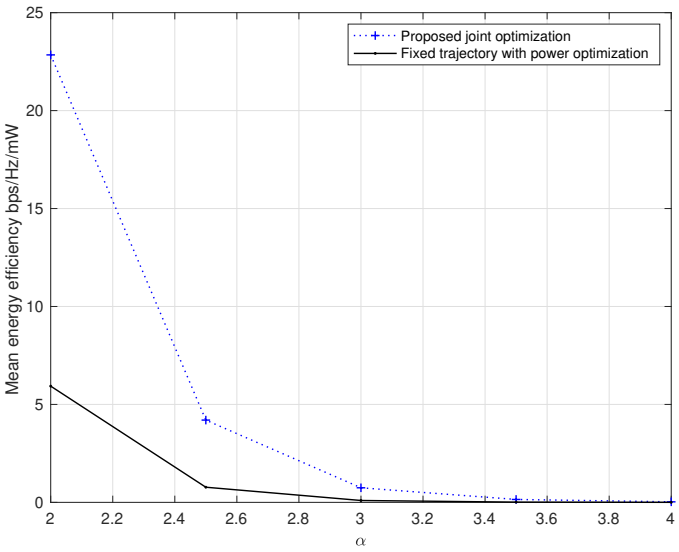

Fig. 7. Effects of path fading exponent α on mean EE.

mean EE, where the capacity threshold C_{th} changes from 0.5 to 2.0 bps/Hz. It can be concluded that the larger the capacity threshold C_{th} is, the smaller the stationary EE is. Specifically, with $C_{th} = 0.5$ bps/Hz, the mean EE reaches 23 bps/Hz/mW when the slot number is 20000. While the corresponding result is only 4 bps/Hz/mW with $C_{th} = 2.0$ bps/Hz, i.e., 82.6% reduction in terms of mean EE. The reason is that when capacity threshold C_{th} becomes larger, more transmission power is used to support the minimum capacity of each secondary user. Afterward, the feasible solution is compressed and the value of objective function is decreased.

The effects of path fading exponent α on mean EE are given in Fig. 7, where the path fading exponent α changes from 2 to 4. We can observe that when α changes larger, the EE reduces from 23 to smaller than 1 bps/Hz/mW. The reason is that when the path fading exponent becomes larger, more energy is consumed to achieve the same transmission rate. Thus, the mean EE becomes smaller rapidly. Also, significant performance gain can be obtained by the proposed iterative optimization algorithm compare with the fixed trajectory scheme.

VII. Conclusion

The combination of NOMA and UAV can be adopted in Industrial IoT to improve the wireless connections and spectrum efficiency. On the condition of interference constraint and minimum rate of each secondary user, we propose an iterative reinforcement learning based algorithm on joint optimization of UAV trajectory and power allocation scheme for NOMA protocol. Firstly, with given UAV trajectory, the Dinkelbach method based fractional programming is adopted to obtain the optimal transmission power and the power allocation scheme. With the help of previous power allocation scheme, the successive convex optimization algorithm is used in the second

stage to update the system parameters. Finally, reinforcement learning based optimization is introduced to obtain the best UAV trajectory in terms of mean EE.

ACKNOWLEDGMENT

This work was partly supported by Natural Science Foundation of Guangdong, China (2022A1515010999), Science and Technology Program of Guangzhou, China (202201011850), Scientific Research Project of Colleges in Guangdong, China (2021KCXTD061), Scientific Research Project of Guangzhou Education Bureau (202032761).

REFERENCES

[1] R. Jhaveri, R. Sagar, G. Srivastava, T. R. Gadekallu, and V. Aggarwal, "Fault-resilience for bandwidth management in industrial software-defined networks," *IEEE Transactions on Network Science and Engineering*, vol. 2, no. 6, pp. 1–11, 2021.

[2] X. Li, Z. Xie, Z. Chu, V. G. Menon, S. Mumtaz, and J. Zhang, "Exploiting benefits of irs in wireless powered NOMA networks," *IEEE Transactions on Green Communications and Networking*, vol. 6, no. 1, pp. 175–186, 2022.

[3] W. Wang, C. Qiu, Z. Yin, G. Srivastava, T. R. Gadekallu, F. Alsolami, and C. Su, "Blockchain and puf-based lightweight authentication protocol for wireless medical sensor networks," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[4] X. Li, Y. Zheng, W. U. Khan, M. Zeng, D. Li, G. K. Ragesh, and L. Li, "Physical layer security of cognitive ambient backscatter communications for green internet-of-things," *IEEE Transactions on Green Communications and Networking*, vol. 5, no. 3, pp. 1066–1076, 2021.

[5] H. Jin, X. Dai, J. Xiao, B. Li, H. Li, and Y. Zhang, "Cross-cluster federated learning and blockchain for internet of medical things," *IEEE Internet of Things Journal*, vol. 8, no. 21, pp. 15 776–15 784, 2021.

[6] X. Huang, X. Yang, Q. Chen, and J. Zhang, "Task offloading optimization for uav-assisted fog-enabled internet of things networks," *IEEE Internet of Things Journal*, vol. pp, no. 99, pp. 1–1, 2021.

[7] B. Jiang, J. Yang, H. Xu, H. Song, and G. Zheng, "Multimedia data throughput maximization in internet-of-things system based on optimization of cache-enabled uav," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3525–3532, 2019.

[8] Z. Kuang, L. Zhang, and L. Zhao, "Energy- and spectral-efficiency tradeoff with α-fairness in energy harvesting D2D communication," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9972–9983, 2020.

[9] B. Xu, P. Zhu, J. Li, D. Wang, and X. You, "Joint long-term energy efficiency optimization in C-RAN with hybrid energy supply," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 10, pp. 11 128–11 138, 2020.

[10] F. K. Ojo and M. F. Mohd Salleh, "Energy efficiency optimization for SWIPT-enabled cooperative relay networks in the presence of interfering transmitter," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1806–1810, 2019.

[11] S. Chatterjee, S. P. Maity, and T. Acharya, "Energy-spectrum efficiency trade-off in energy harvesting cooperative cognitive radio networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 2, pp. 295–303, 2019.

[12] W. Dinkelbach, "On nonlinear fractional programming," *Management Science*, vol. 13, no. 7, pp. 492–498, 1967.

[13] X. Wang, Z. Na, K. Lam, X. Liu, Z. Gao, F. Li, and L. Wang, "Energy efficiency optimization for noma-based cognitive radio with energy harvesting," *IEEE Access*, vol. 7, pp. 139 172–139 180, 2019.

[14] D. Deng, X. Li, V. Menon, M. J. Piran, H. Chen, and M. A. Janf, "Learning based joint UAV trajectory and power allocation optimization for secure IoT networks," *Digital Communications and Networks*, vol. 8, no. 1, pp. 411–418, 2022.

[15] L. FXingwang, G. Xuesong, L. Yingting, H. Gaojian, Z. Ming, and Q. Dawei, "Energy-efficient uav-enabled data collection via wireless charging: A reinforcement learning approach," *Chinese Journal of Electronics*, vol. pp, no. 99, pp. 1–12, accepted in 2022.

[16] L. Ruan, J. Wang, J. Chen, Y. Xu, Y. Yang, H. Jiang, Y. Zhang, and Y. Xu, "Energy-efficient multi-UAV coverage deployment in UAV networks: A game-theoretic framework," *China Communications*, vol. 15, no. 10, pp. 194–209, 2018.

[17] 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.

[18] S. Yang, Y. Deng, X. Tang, Y. Ding, and J. Zhou, "Energy efficiency optimization for UAV-assisted backscatter communications," *IEEE Communications Letters*, vol. 23, no. 11, pp. 2041–2045, 2019.

[19] S. Fu, Y. Tang, Y. Wu, N. Zhang, H. Gu, C. Chen, and M. Liu, "Energy-efficient uav-enabled data collection via wireless charging: A reinforcement learning approach," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10 209–10 219, 2021.

[20] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "Noma-enabled mobile edge computing for internet of things via joint communication and computation resource allocations," *IEEE Internet of Things Journal*, vol. 7, no. 1, pp. 718–733, 2020.

[21] L. Han, R. Liu, Z. Wang, X. Yue, and J. S. Thompson, "Millimeter-wave mimo-noma-based positioning system for internet-of-things applications," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 068–11 077, 2020.

[22] D. Deng and M. Zhu, "Joint UAV trajectory and power allocation optimization for NOMA in cognitive radio network," *Physical Communication*, vol. 46, no. 6, pp. 101 328–101 340, 2021.

[23] L. Geng, L. Huiling, H. Gaojian, L. Xingwang, R. Bichu, and K. Ferdi, "Eeffective capacity analysis of reconfigurable intelligent surfaces aided NOMA network," *EURASIP Journal on Wireless Communications and Networking*, vol. 198, no. 12, pp. 1–16, 2021.

[24] Y. Lin, Z. Yang, and H. Guo, "Proportional fairness-based energy-efficient power allocation in downlink MIMO-NOMA systems with statistical CSI," *China Communications*, vol. 16, no. 12, pp. 47–55, 2019.

[25] M. Song and M. Zheng, "Energy efficiency optimization for wireless powered sensor networks with nonorthogonal multiple access," *IEEE Sensors Letters*, vol. 2, no. 1, pp. 1–4, 2018.

[26] T. V. Nguyen, V. D. Nguyen, D. B. da Costa, and B. An, "Hybrid user pairing for spectral and energy efficiencies in multiuser MISO-NOMA networks with SWIPT," *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4874–4890, 2020.

[27] D. Wang and S. Men, "Secure energy efficiency for NOMA based cognitive radio networks with nonlinear energy harvesting," *IEEE Access*, vol. 6, pp. 62 707–62 716, 2018.

[28] M. C. Hlophe and B. T. Maharaj, "QoS provisioning and energy saving scheme for distributed cognitive radio networks using deep learning," *Journal of Communications and Networks*, vol. 22, no. 3, pp. 185–204, 2020.

[29] Y. Yang, Y. Wang, K. Liu, N. Zhang, S. Gu, and Q. Zhang, "Deep reinforcement learning based online network selection in crns with multiple primary networks," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7691–7699, 2020.

[30] Z. Zhang, Y. Lu, Y. Huang, and P. Zhang, "Neural network-based relay selection in two-way SWIPT-enabled cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6264–6274, 2020.

[31] G. Baggio, R. Bassoli, and F. Granelli, "Cognitive software-defined networking using fuzzy cognitive maps," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 3, pp. 517–539, 2019.

[32] J. M. Kang, I. M. Kim, and C. J. Chun, "Deep learning-based MIMO-NOMA with imperfect sic decoding," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3414–3417, 2020.

[33] H. Huang, Y. Yang, Z. Ding, H. Wang, H. Sari, and F. Adachi, "Deep learning-based sum data rate and energy efficiency optimization for MIMO-NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 8, pp. 5373–5388, 2020.

[34] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2200–2210, 2019.

[35] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for uav-enabled multiple access," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2018.

[36] N. Zhao, X. Pang, Z. Li, Y. Chen, F. Li, Z. Ding, and M. Alouini, "Joint trajectory and precoding optimization for uav-assisted noma networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3723–3735, 2019.

[37] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[38] Z. Q. Luo, W. K. Ma, M. C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *Signal Processing Magazine, IEEE*, vol. 27, no. 3, pp. 20–34, 2010.

Dan Deng (Senior Member, IEEE) received his Bachelor and Ph.D. degrees from University of Science and Technology of China, in 2003 and 2008, respectively. From 2008 to 2014, he was with Comba Telecom Ltd. in Guangzhou China, as a Director. Since 2014, he has joined Guangzhou Panyu Polytechnic as a full professor. His research interests include wireless communication and machine learning for signal processing in next generation wireless communication systems. He has published 73 papers in international journals and conferences. Also, he holds 25 patents, and has served as a member of Technical Program Committees for several conferences.