Learning-Aided UAV 3D Placement and Power Allocation for Sum-Capacity Enhancement Under Varying Altitudes

Zeeshan Kaleem, *Senior Member, IEEE*, Waqas Khalid, Ali Muqaibel, *Senior Member, IEEE*, Ali Arshad Nasir, Chau Yuen, *Fellow, IEEE*, and George K. Karagiannidis, *Fellow, IEEE*

***Abstract*— Unmanned air vehicle (UAV) as an aerial base station (ABS) has attracted the attention of cellular service providers to enable emergency communications. However, the unplanned multiple ABS deployment poses severe interference challenges that degrade the user's performance. To maximize the system sum capacity, we propose the use of K-means and Q-learning assisted 3D ABS Placement and Power allocation algorithm (KQPP). Specifically, we combine the benefits of K-means and Q-learning algorithms to achieve this goal. As a result, we successfully improve the sum capacity by satisfying all the users' minimum data rate requirements. The proposed approach achieves 6bps/Hz and 16bps/Hz higher sum-capacity gain compared to equal power allocation and particle swarm optimization (PSO)-based power allocation schemes, respectively.**

***Index Terms*— ABS placement, power allocation, reinforcement learning, sum-capacity maximization.**

I. Introduction

UNMANNED air vehicle (UAV) deployment as an aerial base station (ABS) has attracted service providers' attention to enable emergency communications. However, this deployment results in co-channel interference because of reusing the same frequency band, which degrades the sum-capacity.

Scanning the literature, several existing schemes have addressed the 3D ABS placement problem. For instance, the authors in [1] optimized the UAV placement and power allocation for uplink and downlink cellular operations. They employed block coordinate decent algorithm and particle swarm optimization (PSO) methods to achieve the optimal solution. Results proved that the algorithm converges quickly and the optimal altitude was higher in uplink than that in the downlink scenario.

Manuscript received 25 March 2022; revised 26 April 2022; accepted 29 April 2022. Date of publication 3 May 2022; date of current version 12 July 2022. The author(s) would like to acknowledge the support provided by the Deanship of Research Oversight and Coordination (DROC) at King Fahd University of Petroleum & Minerals (KFUPM) for funding under the Interdisciplinary Research Center for Communication Systems and Sensing through project No. INCS2110. The associate editor coordinating the review of this letter and approving it for publication was H. Ghazzai. *(Corresponding author: Zeeshan Kaleem.)*

Zeeshan Kaleem is with the Electrical and Computer Engineering Department, COMSATS University Islamabad, Wah Campus, Wah Cantt 47040, Pakistan (e-mail: zeeshankaleem@gmail.com).

Waqas Khalid is with the Institute of Industrial Technology, Korea University, Sejong 30019, South Korea.

Ali Muqaibel and Ali Arshad Nasir are with the Electrical Engineering Department and Center for Communication Systems and Sensing, King Fahd University for Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia.

Chau Yuen is with the Engineering Product Development Department, Singapore University of Technology and Design, Singapore 138682.

George K. Karagiannidis is with the Computer Engineering Department, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece.

Digital Object Identifier 10.1109/LCOMM.2022.3172171

In [2], the authors presented two solutions to maximize the minimum achievable throughput by optimizing 3D ABS deployment. Firstly, the authors optimized the ABS horizontal $(x_m^{\mathrm{B}}, y_m^{\mathrm{B}})$ positions by adopting the mean shift technique. Secondly, to jointly optimize the ABS transmit power and altitude, they implemented the block coordinate descent technique. Their results proved that substantial throughput gain was achieved as compared to the conventional schemes.

The authors in [3] summarized the existing literature for ABS trajectory and placement optimization. Most of the existing solutions in the literature either focused on non-intelligent solutions, ABS 2D placement, or considered either K-means or Q-learning models, but none of them considered both models simultaneously. To improve users' quality-of-experience for multi-UAV communication framework, in [4], the authors adopted both genetic algorithm-based K-means algorithm for clustering and Q-learning for placement to optimize UAV 3D deployment and trajectory. The main advantage of using the reinforcement learning schemes was the provisioning of real-time optimization of the problem and the capability to adjust ABS positions based on users' feedback.

The existing solutions are focused on improving users' power allocation [1] and maximizing the minimum achievable system sum-rate [2] by adopting the conventional optimization and clustering algorithms such as PSO and mean shift techniques, respectively [5]. Moreover, those solutions did not account for the impact of varying ABS altitudes and the increase in the number of ABSs on the system sum capacity. However, in [4], [6] the authors proposed reinforcement learning solutions for optimizing the ABS trajectory and deployment by considering quality-of-experience constraints. Similar to [1]–[3], [7], they have shortcomings; ignored the impact of varying ABS altitudes, independently utilized either K-means or Q-learning algorithms, the reward function was not intelligent enough to model necessary quality-of-service (QoS) constraints, and the ABS placement search space was huge. Although the authors in [4] implemented both K-means and Q-learning algorithms, it is not comparable with the proposed scheme as they ignored QoS constraints and have high complexity due to the genetic algorithm-assisted K-means algorithm, hence the fair comparison not possible.

Contributions: Motivated by this, we propose **K**-means and **Q**-learning assisted 3D ABS **P**lacement and **P**ower allocation algorithm (**KQPP**) for sum-capacity maximization, where we adopt both K-means and Q-learning algorithms. Our major contributions are; 1) K-means and discrete search algorithms are adopted to optimize the ABS horizontal positions, 2) model-free Q-learning reinforcement learning algorithm is adopted to enhance sum-capacity by optimizing ABS power allocation under varying ABS altitudes, 3) a new reward

function is developed to maximize the system sum-capacity by giving high reward to the users with more capacity, and also to meet the users' required QoS by reducing the reward of users that deviates from their required QoS.

We discuss the system model, proposed KQPP algorithm, simulation results, and conclusions in Section II, III, IV, and V of the manuscript, respectively.

II. System Model

We deploy a multi-ABS downlink communication system to serve the ground users. We assume there are $M > 1$ ABSs sharing the same bandwidth, where each ABS at least serves one user, which may generate a co-channel interference to the neighbor ABS. Here, we represented the ABS and user sets as $\mathcal{M}$ and $\mathcal{U}$, respectively. Therefore, we can write $|\mathcal{M}| = M$ and $|\mathcal{U}| = U$. Moreover, we consider a 3D Cartesian coordinate system, where the position of each user u is represented as $\Phi_u^{\mathrm{U}} = [x_u^{\mathrm{U}}, y_u^{\mathrm{U}}]^T \in \mathbb{R}^{2\times 1}, u \in \mathcal{U}$, whereas ABS $m \in \mathcal{M}$ also considers an altitude H for its deployment and thus can be represented as $\Psi_m^{\mathrm{B}} = [x_m^{\mathrm{B}}, y_m^{\mathrm{B}}, H_m]^T \in \mathbb{R}^{3\times 1}, m \in \mathcal{M}$. We consider that each ABS height varies in range $[H_{\min}, H_{\max}]$. Moreover, we implement the model-free reinforcement learning (i.e., Q-learning) algorithm for ABS power allocation.[1] To implement Q-learning, we model ABS as an agent, 3D position (x, y, H) as a state (S), and power allocation as an action (A). We record these state-action pairs in the knowledge matrix, Q, where each element in the Q-matrix represents one of the state-action pairs.

We utilize the most adopted probabilistic model to calculate the air-to-ground (A2G) channel between the ABS and the users. This A2G model assumes that the received signals might have experienced a non-line of sight (NLoS) or line of sight (LoS) communication that depends on the probability. These LoS and NLoS probabilities are $P(\mathrm{LoS}) = \frac{1}{1+\kappa \exp(-\lambda[\frac{180}{\pi}\theta - \kappa])}$ and $P(\mathrm{NLoS}) = 1 - P(\mathrm{LoS})$, respectively [8]. The LoS and NLoS probabilities depend on the elevation angle ($\theta = \tan^{-1}(H_m/\rho_u^m)$) and environment dependent Sigmoid function (S-curve) fitting parameters (κ, λ), that vary for rural, urban, and suburban environments [8]. The horizontal distance between ABS m and user u is represented as $\rho_u^m = \sqrt{(x_m^{\mathrm{B}} - x_u^{\mathrm{U}})^2 + (y_m^{\mathrm{B}} - y_u^{\mathrm{U}})^2}$.

The average probabilistic pathloss of user u located at a distance d from the ABS m deployed at an altitude H_m is $\overline{L_u^m} = L_{\mathrm{LoS}} \times P(\mathrm{LoS}) + L_{\mathrm{NLoS}} \times P(\mathrm{NLoS})[\mathrm{dB}]$. The average probabilistic pathloss can be expressed in linear scale as $\overline{l_u^m} = 10^{\overline{L_u^m}/10}$.

Here, the mean LoS and NLoS losses for communication links are expressed as $L_{\mathrm{LoS}} = 20\log\left(\frac{4\pi f_c d_u^m}{\mathrm{c}}\right) + \xi_{LoS}$ and $L_{\mathrm{NLoS}} = 20\log\left(\frac{4\pi f_c d_u^m}{\mathrm{c}}\right) + \xi_{NLoS}$, respectively, where d_u^m is the 3D distance between the ABS m and the user u, calculated as $d_u^m = \sqrt{\rho_u^2 + H_m^2}$, f_c, c denote the carrier frequency, speed of light, ξ_{LoS} and ξ_{NLoS} are environment-related constants.

[1] In our future work, we will allocate continuous power, which results in a huge state-action space. For huge power allocation space, we will plan to implement a deep-deterministic policy gradient (DDPG) based algorithm. However, in this work, discrete transmission power levels selection is reasonable because practical communication devices transmit in discrete power levels.

The downlink signal received at the user from the associated ABS includes interference from the neighboring ABS and thermal noise. Hence, the SINR (Γ) at the user u from ABS m is $\Gamma_u^m = \frac{P_k \overline{l_u^m} g_u^m}{\sum_{s=1, s\neq k}^{M} P_s \overline{l_u^s} g_u^s + \sigma^2}$, where g_u^m represents the short-term channel gain between the m-th ABS and the u-th user, that follows exponential distribution and σ^2 is the variance of the additive white Gaussian noise (AWGN). Therefore, the normalized capacity of the user u associated with ABS m is $C_u^m = \log_2(1 + \Gamma_u^m)$ [9].

III. 3D ABS Placement for Sum Capacity Maximization Using Proposed KQPP Algorithm

Since, 3D ABS position and transmit power both affects the users' sum capacity. In this letter, our aim is to maximize the users' sum capacity under the constraints of ABS transmit power, minimum data rate requirement, and ABS 3D positions. $\boldsymbol{\mathcal{P}} = \{P_1, P_2, \ldots, P_M\}$ is a vector containing the ABS transmit powers and $\boldsymbol{w_l}$, $1 \le l \le \mathrm{v_{pos}}$ is a vector containing the ABS 2D horizontal positions ($\mathrm{v_{pos}}$), with initial position $\boldsymbol{w_1} = \boldsymbol{w}_{\min} = (x_1, y_1)$ and final position $\boldsymbol{w}_{\mathrm{v_{pos}}} = \boldsymbol{w}_{\max} = (x_{\mathrm{v_{pos}}}, y_{\mathrm{v_{pos}}})$. We formulate the sum capacity maximization problem as

$$\max_{\boldsymbol{\mathcal{P}}, \boldsymbol{w_l}, H_m} \sum_{m=1}^{M}\sum_{u=1}^{U} C_u^m \tag{1a}$$

$$\text{s.t. } C_u^m \ge \tilde{q}, \quad \forall m \in \mathcal{M} \tag{1b}$$

$$(x_m^{\mathrm{B}} - x_u^{\mathrm{U}})^2 + (y_m^{\mathrm{B}} - y_u^{\mathrm{U}})^2 \le \rho^2, \quad \forall u \in \mathcal{U},\ m \in \mathcal{M} \tag{1c}$$

$$(x_m^{\mathrm{B}} - x_u^{\mathrm{U}})^2 + (y_m^{\mathrm{B}} - y_u^{\mathrm{U}})^2 \ge 4\rho^2, \quad \forall m \in \mathcal{M} \tag{1d}$$

$$H_{\min} \le H_m \le H_{\max} \quad \forall m \in \mathcal{M} \tag{1e}$$

$$P_{\min} \le P_m \le P_{\max}, \quad \forall m \in \mathcal{M}. \tag{1f}$$

Constraint (1b) ensures that each user's minimum data rate should be above the defined threshold, $\tilde{q}$. Furthermore, we introduce the constraint (1c) that ensures that user u lies within the distance of ρ from ABS center. Constraint (1d) ensures that there is no overlap between the coverage areas of deployed $\mathcal{M}$ ABS. To meet this constraint, the horizontal distance among two ABS should be larger than 2ρ. The constraints (1e) and (1f) ensure that the ABS deployment and transmit power, respectively are within the specified limits.

The problem presented in (1a) is non-convex in nature because of the several non-convex constraints presented in (1c)-(1f), and hence difficult to solve directly. The formulated sum-capacity maximization problem (1a) relates to Markov decision process (MDP), where a UAV acts as an agent. MDP consists of four main components; a) set of finite states (here, we map the number of ABS to states S, which are finite), b) set of finite actions (here, we map discrete choice of power allocation to set of actions as $\boldsymbol{A}$, which are also finite), c) state transition probability $\mathbb{P}^a_{s_t, s_{t+1}}$, which describes how current state and action influence the future state by taking action a (here, we change the state based on maximum Q-value, and d) the reward function (here, we adopt the reward function given in (4)). Therefore, based on the mentioned similarities with MDP formulation, the proposed problem is

considered an MDP problem. Moreover, the state-transition also satisfies the Markov property as the decision to change the state only depends on the values in the current state rather than past states.

To solve this we divide the problem into two sub-problems; 1) optimization of ABS horizontal positions $\boldsymbol{w}^* = (x_m^{\mathrm{B}}, y_m^{\mathrm{B}})$ with fix ABS height H_m and transmit power P_m, 2) optimization of ABS power for varying ABS height for the optimized horizontal $\boldsymbol{w}^*$ positions.

A. ABS Horizontal Placement Optimization

With fixed altitude H_m and $\boldsymbol{\mathcal{P}}$, the problem in (1) is formulated as

$$\max_{\boldsymbol{w_l}^m} \sum_{m=1}^{M} \sum_{u=1}^{U} C_u^m \tag{2a}$$

$$\text{s.t. } (1b),\ (1c),\ (1d). \tag{2b}$$

The presented sub-problem in (2a) is simplified as we fix the transmit power and the ABS height. Since, there are infinite number of horizontal positions in the region that requires exhaustive search to find the optimal horizontal positions, which in turn will be infeasible to implement. To intelligently deploy ABS in the region that can cover large number of users and reduce the search window, we adopt K-means clustering algorithm, which organize users into K-clusters. The partition of each cluster is calculated using the sum-of-squared-error criterion as $e = \sum_{k=1}^{K} \sum_{u=1}^{U_k} \|\boldsymbol{\Phi}_u - \zeta_k\|^2$ where $\boldsymbol{\Phi}_u$ stores locations of users and ζ_k stores the center location of k-th cluster. We obtain the cluster center by calculating the mean value of all user points classified to that cluster, and is calculated as $\zeta_k = [\zeta_{kx}, \zeta_{ky}] = \left\{ \frac{1}{U_k} \sum_{u=1}^{U_k} x_u, \frac{1}{U_k} \sum_{u=1}^{U_k} y_u \right\}$ where U_k represents the total number of users associated with k-th cluster. Hence, the new search window for ABS 2D positions $\boldsymbol{w}$ lies in the range $[\zeta_k, \mathrm{v_{pos}}]$ rather than $[1, \mathrm{v_{pos}}]$.

Thus, we discretize those positions to have feasible finite points, where we first form a vector of $\mathrm{v_{pos}}$ 2D positions after adjusting the starting point based on cluster center ζ_k rather than the complete search space to further reduce the search space. That is, $\boldsymbol{w_l}^m = \{(x_{\zeta_k}^{\mathrm{B}}, y_{\zeta_k}^{\mathrm{B}}), (x_{\zeta_{k+1}}^{\mathrm{B}}, y_{\zeta_{k+1}}^{\mathrm{B}}), \ldots, (x_{\mathrm{v_{pos}}}^{\mathrm{B}}, y_{\mathrm{v_{pos}}}^{\mathrm{B}})\}, \zeta_k \leq l \leq \mathrm{v_{pos}}$, where $\boldsymbol{w_l}^m$ is the search space for the horizontal placement of m-th ABS. The $\boldsymbol{w_1}^m = \boldsymbol{w}_{\min}{}^m = (x_{\zeta_k}^{\mathrm{B}}, y_{\zeta_k}^{\mathrm{B}})$ and $\boldsymbol{w}_{\mathrm{v_{pos}}}^m = \boldsymbol{w}_{\max}{}^m = (x_{\mathrm{v_{pos}}}^{\mathrm{M}}, y_{\mathrm{v_{pos}}}^{\mathrm{M}})$ are the minimum and maximum boundaries, respectively for m-th ABS. The displacement between two consecutive points is calculated as $\kappa = ||\mathbf{w}_{l+1}^m - \mathbf{w}_l^m||$. We solve this problem by substituting each $\mathbf{w}_l^m$ into (2a) and evaluate utility $Y[l] = \max_{\boldsymbol{w_l}^m} \sum_{m=1}^{M} \sum_{u=1}^{U} C_u^m(w_l^m)$. Finally, we obtain the optimum or sub-optimum $\boldsymbol{w}^*$ that results in max sum-capacity when iterated over the positions $\zeta_k \leq l \leq \mathrm{v_{pos}}$. We summarize those steps in the Algorithm 1.

B. ABS Power Optimization

After getting the optimal 2D horizontal position ($\boldsymbol{w}^*$) by solving (2), we calculate the optimal ABS power P_m^* by iteratively varying ABS height H_m in range $[H_{\min}, H_{\max}]$ and using ($\boldsymbol{w}^*$). The problem in (1) is reformulated as

$$\max_{\boldsymbol{\mathcal{P}}, H_m} \sum_{m=1}^{M} \sum_{u=1}^{U} C_u^m \tag{3a}$$

$$\text{s.t. } (1e),\ (1f). \tag{3b}$$

The problem in (3) is solved using the ϵ-greedy-based Q-learning algorithm [10] with key steps summarized in Algorithm 2. In the proposed algorithm, each ABS as an agent maintains the Q-matrix to record the best possible action in the current state. In this letter, the action ($\boldsymbol{A}$) is the allocated power to the ABS from a vector $\boldsymbol{A} = \{a_1, a_2, \ldots, a_{N_{\text{power}}}\}$, which covers the range between minimum power $P_{\min}$ and maximum power $P_{\max}$ with a length N_{power}. In the current proposal, as we are not receiving any particular information from the environment, thus we equally divide the action (i.e., power) space.

The reward plays the main role in achieving the desired sum-capacity maximization objective. Hence, we define the reward r_t^m for m-th ABS at t-th iteration as

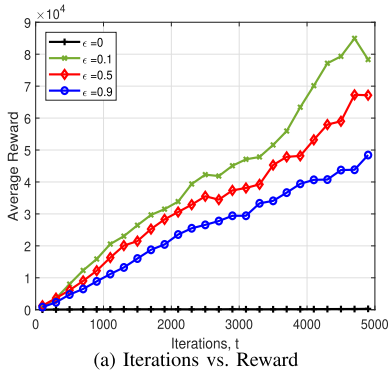
$$r_t^m = \beta^m C_{u,t}^m - \frac{1}{\beta^m}(C_{u,t}^m - \tilde{q})^2, \tag{4}$$

where $C_{u,t}^m$ is the user u capacity associate with m-th ABS capacity at t-th iteration and $\tilde{q}$ is the minimum required data rate for user u associated to m-th ABS. The first term in the reward expression ensures that a higher capacity of users improves the reward. However, if the capacity decreases from the threshold, the second part of the equation ensures that the reward reduces. β^m provides the fairness to the algorithm, that is, β^m assists in giving more reward to the ABS that increases the rate and decreases the ABS reward when the rate reduces below the threshold $\tilde{q}$, and we calculate $\beta^m = \frac{\rho}{\rho_{\text{th}}}$, with $\rho_{\text{th}} = 50\text{m}$. We iteratively update the state-action value function for each agent m by using the simple temporal difference-based one step Q-learning approach as follows $Q^m(s_t^m, a_t^m) \leftarrow (1-\alpha)Q^m(s_t^m, a_t^m) + \alpha\left(r_t^m + \max_a \gamma Q^m(s_{t+1}^m, a)\right)$ [10], where α is the learning rate and γ is the discount factor.

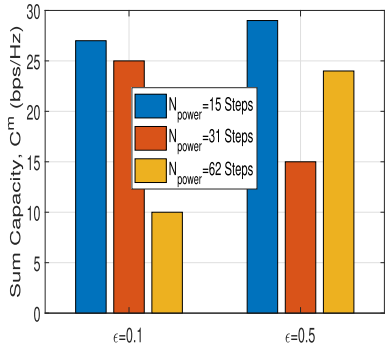
Let us analyze the worst case complexity of the proposed KQPP algorithm. The complexity of Algorithm 1 which implements K-means and discret serach is $O(|\mathcal{M}||\mathcal{U}|T + (\frac{\mathrm{v_{pos}} - \zeta_k}{\kappa})|\mathcal{M}|)$. However, the complexity of Algorithm 2 that implements Q-learning is $O((\frac{H_{\max} - H_{\min}}{\Delta H})|\mathcal{M}||\mathcal{U}|T)$. Therefore, the complexity of the KQPP is $O(|\mathcal{M}||\mathcal{U}|T + (\frac{\mathrm{v_{pos}} - \zeta_k}{\kappa})|\mathcal{M}| + (\frac{H_{\max} - H_{\min}}{\Delta H})|\mathcal{M}||\mathcal{U}|T)$.

IV. Simulation Results

We deploy the varying number of users $U \in (50, 90)$ and ABS $M \in (4, 8)$, respectively, in the region of area $1000 \times 1000\text{m}^2$. We randomly deploy the users in each ABS, with a coverage radius of 100m, that follows a uniform distribution, and at least a single user is attached to each ABS. We model the co-channel interference scenario among the ABS cells by reusing the same frequency band. The ABS transmission power varies in the range of $[-20, 25]$dBm and divided equally into $N_{power} = 31$ steps with a step-size of 1.5dBm. These 31 steps represent the actions against each state in a Q-matrix. The case with $M \in (4, 8)$ results in 4 and 8 states, respectively. We fix the ABS height after calculating the w^* for each ABS



(a) Iterations vs. Reward



(b) Impact of N_{power} steps variation

Fig. 1. Impact of N_{power} and ϵ: (a) Average reward comparison for varying ϵ for $M = 4$, whereas it is equally valid for other number of M; (b) Impact of exploration vs exploitation by varying ϵ and N_{power} steps.

Algorithm 1 K-Means and Discrete Search-Based ABS 2D Placement Optimization Algorithm

K-means clustering: Deploy U users in the deployment grid having total $\mathrm{v_{pos}}$ number of positions;
Deploy K UAVs in the deployment grid;
Calculate cluster partition using sum-of-squared-error criterion
$e = \sum_{k=1}^{K} \sum_{u=1}^{U_k} \|\mathbf{\Phi}_u - \zeta_k\|^2$;
Get cluster center by calculating the mean value as
$\zeta_k = [\zeta_{kx}, \zeta_{ky}] = \left\{ \frac{1}{U_k} \sum_{u=1}^{U_k} x_u, \frac{1}{U_k} \sum_{u=1}^{U_k} y_u \right\}$;
Discrete Search Input: Use K-means assisted ABS 2D positions vector $\boldsymbol{w_l^m}(\zeta_k \leq l \leq \mathrm{v_{pos}})$ of spacing κ with start point as $w_{\min}$ and end point $w_{\mathrm{v_{pos}}}$. Keep fix height H_m and power P_m for each ABS;
Output: ABS 2D optimal position $\boldsymbol{w}^*$;
for $(l = \zeta_k : \Delta l : \mathrm{v_{pos}})$ **do**
  Calculate sum capacity using (2), i.e.,
  $Y[l] = \max_{\boldsymbol{w_l}^m} \sum_{m=1}^{M} \sum_{u=1}^{U} C_u^m(w_l^m)$;
end
Get the output Y that maximize the sum capacity; $Y = \max_l(Y[l])$;
Return optimum 2D ABS $\boldsymbol{w}^*$ that maximize Y in the previous step;

during the number of iterations, and then optimal power is calculated against each (w^*, H_m). This process is repeated for varying ABS height in the range $[H_{\min}, H_{\max}] = [100, 700]$m with a step size of 40 m. In turn, we calculate the sum capacity against each height to show which ABS height gives us the maximum sum capacity.

We verify the effectiveness of the proposed 3D ABS placement scheme by comparing it with the baseline schemes; 1) Exhaustive Search (ES) for power allocation, 2) Equal Power Allocation (EPA), and 3) PSO-based power allocation (PSO-PA) schemes. To have a fair comparison, ABS are deployed in the coverage area using K-means clustering algorithm for those schemes. For power allocation, in ES, the power is allocated among multiple ABS using each available power combination. In EPA, the power is allocated equally among ABSs. In PSO-PA, the power is allocated using PSO algorithm [11] with the key parameters, i.e., population size ($\nu = 200$), inertial minimum and maximum weight $(\varpi_{\min}, \varpi_{\max}) = (0.8, 1)$, acceleration factors $(c_1, c_2) = (2, 2)$. To implement Q-learning algorithm, the key parameters values used are; discount factor $\gamma = 0.5$, learning rate $\alpha = 0.9$, and $\epsilon = 0.1$ because these values gives us maximum average reward and quick convergence as shown in Fig. 1 (a) and Fig. 2.

We calculate average reward in Fig. 1 (a) using $r_t = \frac{1}{M} \sum_{m \in \mathcal{M}} r_t^m$ for varying ϵ when $M = 4$. The results

Algorithm 2 Q-Learning Assisted ABS Height and Power Allocation for Sum Capacity Maximization

Input: Declare ABS 3D positions vector $(\boldsymbol{w_m^*}, H_m)$ by using optimal $\boldsymbol{w_m^*}$ from Algorithm 1 and initialize Q matrices. Also initialize the action vector $A^m = \{a_1^m, a_2^m, \ldots, a_{N_{\text{power}}}^m\}$ and set $Q_t^m(s_t^m, a_t^m) = 0$;
Output: Optimal $Q^*(s_t, a_t)$ that maximizes the sum capacity for varying ABS height H_m;
for $(H = H_{min} : \Delta H : H_{max})$ **do**
  while $(t \leq T) \parallel (L \leq 1\%)$ **do**
    for ***all*** *ABS* m, $m \in \mathcal{M}$ **do**
      Choose action a_t^m (i.e., power P^m for each ABS) from set of actions;
      Take action a_t^m (i.e., allocate power using ϵ-greedy approach) and calculate reward r_t^m and capacity against each action;
      Update the state to s^{t+1} of each ABS;
      Update **Q**-matrix $Q^m(s_t^m, a_t^m)$ for current state using $Q^m(s_t^m, a_t^m) \leftarrow (1-\alpha) Q^m(s_t^m, a_t^m) + \alpha \left(r_t^m + \max_a \gamma Q^m(s_{t+1}^m, a) \right)$;
    end
    Update $t = t + 1$;
  end
  Store **Q**-matrix for each ABS that maximizes sum-capacity against each height H_m;
end
Return optimum $Q^*(s_t, a_t)$ against each height that maximizes sum-capacity against varying ABS height and sub-optimum horizontal positions $\boldsymbol{w}^*$;

clearly indicate that reward increases by increasing iterations as the long-term reward increases with iterations. We select the $\epsilon = 0.1$ for the simulations because it gives maximum reward. Whereas we can notice that the reward not increases when $\epsilon = 0$ as the agent will not explore other actions to maximize the reward. Moreover, we verify the impact of increasing N_{power} steps on sum-capacity in Fig. 1 (b). We notice that increasing N_{power} steps is only beneficial by increasing exploration (i.e., higher value of $\epsilon = 0.5$) rather than for higher exploitation (lower value of $\epsilon = 0.1$).

We select learning rate $\alpha = 0.9$ for simulations as it is noticeable from the Fig. 2 that it converges quickly in 36.1 sec in 100 iterations. The main reason is its low loss $L = |Q^m(s_t^m, a_t^m) - Q(s_{t+1}^m, a_{t+1}^m)|$ in the beginning of the iterations and capability to avoid under-and over-fitting. We implement the ϵ-greedy algorithm for the first 80% of the iterations out of the total iterations of $T = 50,000$ as it was proved [10] to converge faster and provide results near optimal. The environment considered here is suburban, and the key related parameters are $f_c = 900$MHz, $(\zeta_{LoS}, \zeta_{NLoS}) = (0.1, 21)$dB [8], $(\kappa, \lambda) = (4.88, 0.43)$ [8], $\bar{q}^m = 0.5$bps/Hz.

To obtain the optimal ABS horizontal positions w^*, we simulate the K-means algorithm by deploying M number of ABS into the area with discrete available positions $\mathrm{v_{pos}} = 100$.

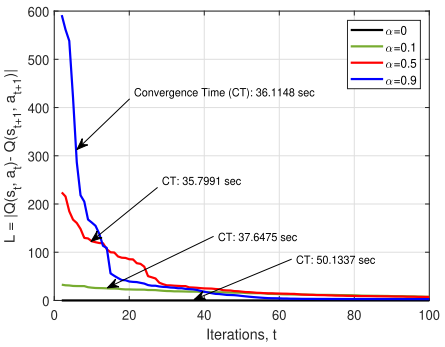


Fig. 2. Learning rate α selection for quick convergence of the proposed algorithm.

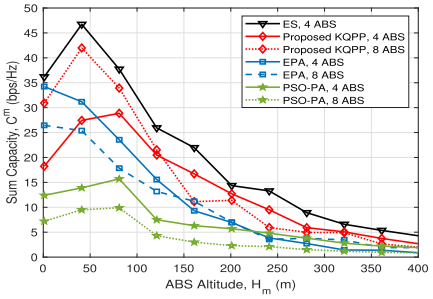


Fig. 3. Effect of varying ABS altitude on users' sum-capacity for different deployment situations.

K-means algorithm is the iterative clustering method that works by finding the best center. Moreover, we select the number of clusters K based on the number of users and their capacity requirements. We verify by simulations that increasing the cluster size K improved the users' sum capacity while placing ABSs at low altitudes. However, ABS with high altitude and more clusters will decrease the sum capacity due to high path loss and more interference generation because of increased ABS's coverage, as shown in Fig. 3. Hence, the choice of K depends on the number of users, their rate requirements, and ABS altitude.

To verify the efficacy of the proposed algorithm in terms of achieving the system sum-capacity, we compare the performance of the proposed algorithm with ES, EPA, and PSO-PA algorithms in Fig. 3. Simulation results show that the proposed KQPP learning algorithm improves the sum capacity for varying numbers of deployed ABS. We compare the performance of KQPP learning algorithm at ABS height of $H_m = 100$m for $M = 4$ with EPA and PSO-PA, the results shows that the proposed KQPP has 6bps/Hz and 16bps/Hz high sum-capacity as compare to EPA and PSO-PA, respectively. However, the sum-capacity of ES is higher than the propose KQPP but at the expense of high processing time. The processing time of the proposed KQPP, ES, EPA, and PSO-PA is around 39.99, 100.18, 23.87, and 80.45sec, respectively. We also notice that the sum-capacity decreases when number of agents (i.e., ABS) increases because it generates high interference. Moreover, the sum-capacity also decreases after certain height because of increasing channel gain.

We observe steady increase in the sum capacity from Fig. 4 as maximum transmit power $P_{\max}$ increases. It is noticeable

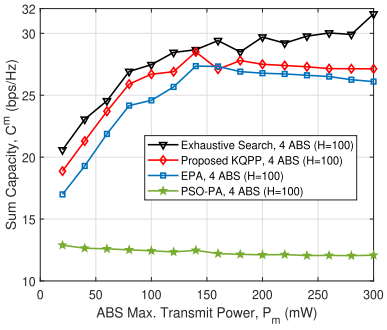


Fig. 4. Effect of varying ABS transmit power on users' sum-capacity for different deployment situations.

that the proposed KQPP scheme clearly outperforms other schemes when compared at different $P_{\max}$ values. However, there is not much sum capacity improvement for all the schemes after transmit power reaches 150mW because of significant increase in the co-channel interference beyond this power.

V. Conclusion

We maximized the user's sum capacity under the numerous QoS constraints by proposing the intelligent KQPP placement and power allocation algorithm in a multi-ABS system. To achieve the sum capacity, we adopted a K-means algorithm to optimize the ABS horizontal positions w^* and Q-learning to optimize the power allocation. The simulation results proved that the proposed KQPP algorithm achieved 6bps/Hz and 16bps/Hz higher sum-capacity gain compared to equal power allocation and PSO-based power allocation schemes, respectively.

References

[1] M. A. Ali and A. Jamalipour, "UAV placement and power allocation in uplink and downlink operations of cellular network," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4383–4393, Jul. 2020.
[2] I. Valiulahi and C. Masouros, "Multi-UAV deployment for throughput maximization in the presence of co-channel interference," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3605–3618, Mar. 2021.
[3] M.-A. Lahmeri, M. A. Kishk, and M.-S. Alouini, "Artificial intelligence for UAV-enabled wireless networks: A survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 1015–1040, 2021.
[4] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.
[5] J.-M. Kang and C.-J. Chun, "Joint trajectory design, Tx power allocation, and Rx power splitting for UAV-enabled multicasting SWIPT systems," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3740–3743, Sep. 2020.
[6] J.-M. Kang, "Reinforcement learning based adaptive resource allocation for wireless powered communication systems," *IEEE Commun. Lett.*, vol. 24, no. 8, pp. 1752–1756, Aug. 2020.
[7] X. Liu *et al.*, "Placement and power allocation for NOMA-UAV networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 965–968, Jun. 2019.
[8] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
[9] S. Shakoor, Z. Kaleem, D.-T. Do, O. A. Dobre, and A. Jamalipour, "Joint optimization of UAV 3-D placement and path-loss factor for energy-efficient maximal coverage," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9776–9786, Jun. 2021.
[10] R. S. Sutton and A. G. Barto, *Reinforcement Learning: AN Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[11] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. Int. Conf. Neural Netw.*, Nov. 1995, pp. 1942–1948.