Dynamic Aerial Base Station Placement for Minimum-Delay Communications

Tong Bai, Cunhua Pan, Jingjing Wang, Yansha Deng, Maged Elkashlan, Arumugam Nallanathan, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Queuing delay is of essential importance in the Internet-of-Things scenarios where the buffer sizes of devices are limited. The existing cross-layer research contributions aiming at minimizing the queuing delay usually rely on either transmit power control or dynamic spectrum allocation. Bearing in mind that the transmission throughput is dependent on the distance between the transmitter and the receiver, in this context we exploit the agility of the unmanned aerial vehicle (UAV)-mounted base stations for proactively adjusting the aerial base station (ABS)'s placement in accordance with wireless tele-traffic dynamics. Specifically, we formulate a minimum-delay ABS placement problem for UAV-enabled networks, subject to realistic constraints on the ABS's battery life and velocity. Its solutions are technically realized under three different assumptions in regard to the wireless tele-traffic dynamics. The backward induction technique is invoked for both the scenario where the full knowledge of the wireless tele-traffic dynamics is available, and for the case where only their statistical knowledge is available. By contrast, a reinforcement learning aided approach is invoked for the case when neither the exact number of arriving packets nor that of their statistical knowledge is available. The numerical results demonstrate that our proposed algorithms are capable of improving the system's performance compared to the benchmark schemes in terms of both the average delay and of the buffer overflow probability.**

***Index Terms*—UAV, delay-optimal, Markov decision process, dynamic programming, reinforcement learning.**

I. Introduction

Given their agility, the on-demand deployment and the bird's-eye perspective of unmanned aerial vehicles (UAV) [1], they have been exploited in diverse military and civilian areas [2], such as surveillance and environmental monitoring [3], data collection [4], mobile edge computing [5], wireless power transfer [6], and wireless networking [7]. Particular to wireless networking, the decreasing cost and increasing sophistication of consumer UAVs combined with miniaturization of BS electronics have made it technically feasible to deploy base stations (BSs) on flying UAVs [8]. In practice, UAVs may carry BSs for supporting emergency communications in scenarios, where the communications infrastructure is destroyed [9] and for assisting the terrestrial cellular network in remote areas and at hotspots (e.g. stadiums) [10]. In order to further exploit the potential of this three-dimensional (3-D) infrastructure, extensive research contributions have been made in air-ground channel modeling [11]–[21], propulsion power conservation [22], [23], link-level implementations [24]–[26] and in system-level designs [27]–[30].

Among the design metrics in UAV-enabled communications systems, the placement and trajectory planning of UAVs play a crucial role [7]. More explicitly, UAVs either have rotary [23] or fixed wings [22]. Rotary-wing UAVs are capable of moving in any direction or of hovering in the air [23]. Their positions have been optimized for meeting diverse requirements, such as throughput [31], coverage [32], privacy preservation [33] and ultra-reliable low-latency communications services [34]. Their locations may also be dynamically adjusted in accordance with the users' locations [35]. By contrast, fixed-wing UAVs have to maintain continuous forward motion for remaining aloft [22]. Their trajectory optimization can be appropriately adjusted for acting as a BS [36], a relay [37], a computing node [5] and a data collector [38]. The existing research contributions on UAV communications mainly focus on the static networks, where a set of constant throughput values are required by the users. In practice, however, the wireless tele-traffic may vary extensively over time, which leads to dynamically fluctuating queuing delay for the users. Therefore, it is imperative to address these wireless tele-traffic dynamics by conceiving a dynamic aerial BS (ABS) placement scheme relying on a cross-layer perspective.

Bearing in mind that the attainable throughput is directly dependent on the distance-related path-loss [39], we may resort to dynamically adjusting the distances between the ABS and the devices supported for the sake of adapting their transmission throughput to the prevalent wireless tele-traffic dynamics. Intuitively, the ABS can be moved close to the devices, which have numerous queuing packets in their buffers. However, this may increase the delay of other devices. Therefore, it is desirable to propose an appropriate dynamic ABS placement strategy for minimizing the average queuing delay of the overall network. The investigations specific to UAV communications concerning the queuing delay are still in their infancy in the open literature. Concretely, based on the analytical results obtained from queuing theory, a resource allocation scheme was proposed for UAV-aided networks for minimizing

This work was supported in part by the EPSRC under grants EP/N004558/1 and EP/N023862/1, in part by the European Research Council's Advanced Fellow Grant under the QuantCom Project, and in part by the Royal Society's Global Research Challenges Grant.

T. Bai, C. Pan, M. Elkashlan and A. Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: t.bai@qmul.ac.uk, c.pan@qmul.ac.uk, maged.elkashlan@qmul.ac.uk, a.nallanathan@qmul.ac.uk).

J. Wang is with the Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China. (e-mail: chinaeephd@gmail.com).

Y. Deng is with the Department of Informatics, King's College London, London, WC2R 2LS, U.K. (e-mail: yansha.deng@kcl.ac.uk).

L. Hanzo is with the School of Electronics and Computer science, University of Southampton, Southampton, SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

the time-averaged queuing delay [40]. The knowledge of the instantaneous wireless traffic was not exploited in this scheme, which inevitably limited the system performance attained. A dynamic trajectory control algorithm was conceived for multi-UAV-enabled networks [41], where the adjacent UAVs were moved one step closer to the specific UAV, whose queue length is higher than a pre-set threshold. However, since this threshold was not optimized, the proposed system was incapable of attaining the minimum delay. A queue-length-aware trajectory design of UAV-mounted computing nodes was proposed for serving multiple devices having limited computational capability [42], whilst relying on the classic Lyapunov optimization theory [43]. Similarly, the activity of UAVs was dynamically controlled by leveraging Lyapunov's optimization theory [43] in UAV-enabled caching systems for maximizing the long-term average revenue, while stabilizing the queuing system [44]. However, the stable queue status is only an indirect indicator of the delay and the policy derived from the Lyapunov optimization approach cannot achieve a satisfactory performance for devices equipped with small storage space [45]. By contrast, the Markov Decision Process (MDP) approach of [46] aims for minimizing the average queuing delay of communications networks. Although it has not been exploited in UAV communications from a wireless-traffic perspective as yet, it has been widely leveraged in terrestrial cellular networks to adapt the transmit power [47], [48], the spectral resource allocation [49], [50], and the user association [51], for accommodating the erratically fluctuating wireless tele-traffic dynamics. These impressive studies inspired us to pursue the minimum-delay design of UAV-enabled networks using the MDP approach.

Against this background, we have conceived a dynamic ABS placement scheme for minimizing the queuing delay of UAV-enabled networks. Specifically, the placement of the ABS is dynamically adjusted for adapting the distance-dependent transmission throughput in response to the fluctuating wireless tele-traffic dynamics. As a benefit, the average queue length in the buffer can be minimized. Given that the dynamic fluctuation of wireless tele-traffic is considered, we have investigated three different scenarios. Explicitly, the first one is, when the tele-traffic is predictable [52]. The second one is, when the specific probability density function of the packet arrival process is known by the ABS [53], while in the third case neither the exact number of arriving packets nor its statistical knowledge is known by the ABS. For each scenario, we have provided a specific dynamic ABS placement strategy for minimizing the average queuing delay. The main technical contributions of this work are summarized as follows:

- *Minimum-delay problem formulation for UAV-enabled networks:* We formulate a minimum-delay ABS placement problem, subject to realistic constraints on the ABS' battery charge and speed. In contrast to the state-of-the-art in UAV communications, where the throughput requirements of ground users are static, in this treatise we consider dynamically fluctuating wireless tele-traffic.
- *Transformation to a Markov decision process problem:* Since the dynamics are imposed by the fluctuating wireless tele-traffic, the conventional trajectory design [22] relying on the sequential convex optimization technique fails to solve this problem. Furthermore, the queue-aware UAV placement [42] based on the Lyapunov optimization theory is incapable of finding the minimum-delay solution. In this context, owing to their one-to-one correspondence, we transform the original minimum-delay problem to the corresponding constrained Markov decision process (MDP), which constitutes an appropriate mathematical framework for solving this stochastic control problem. Then, relying on the classic Lagrangian approach, we reformulate the constrained MDP to an unconstrained MDP.
- *Strategies under diverse assumptions of wireless tele-traffic dynamics:* We holistically consider three assumptions concerning the apriori information of the wireless tele-traffic. To address the problems under these assumptions, we provide solutions for the first and the second scenarios relying on the technique of backward induction, whereas a reinforcement learning aided approach is conceived for the third problem.
- *Numerical validations and evaluations:* Our numerical results quantify the performance of the solutions for the above three scenarios. Specifically, we compare the proposed algorithms to two benchmark schemes, in terms of both the average delay per user and the buffer overflow probability, under various settings of both the ABS' total energy and of the wireless tele-traffic dynamics as well as of ground devices' locations.

The rest of this paper is organized as follows. In Section II, we elaborate on the system model and formulate a minimum-delay ABS placement problem for UAV-enabled networks. Section III details the transformation from the primal minimum-delay problem both to the constrained and then to an unconstrained MDP problem. In Section IV, we provide solutions to these problems in the three scenarios. In Section V, we evaluate the proposed strategies through numerical analysis. Finally, we conclude this study in Section VI.

II. System Models and Problem Formulation

As illustrated in Fig. 1, we consider the uplink of a UAV-enabled network in a 3-D Cartesian coordinate system, where a single-antenna rotary-wing ABS [1] maintaining aloft serves K devices on the ground over T time slots (TSs) [2]. We assume that the ABS is linked with the core network via terrestrial base stations using high-capacity millimeter-wave communications [57]. The backhaul link is assumed to be capable of fully supporting the ABS-enabled network. We use $\boldsymbol{U}[t] = \big(u_x[t], u_y[t], h[t]\big)$ and $(d_x^k, d_y^k, 0)$ to represent the coordinates of the ABS at the t-th TS and of the k-th device, respectively. Each TS lasts τ seconds. The whole spectrum is

[1]This paper aims for verifying the effectiveness of the proposed delay-minimum ABS placement strategies in a single-ABS scenario. Multi-ABS systems can be realized by appropriately designing both the user association and the resource management [27], [54] with the aid of a multi-agent MDP framework [55], [56], which is beyond our current scope.

[2]In this paper, the locations of the ground users are assumed to be static for simplicity, which is applicable to the nodes of wireless sensor networks.

partitioned into K non-overlapping equal-bandwidth subchannels, while each device is connected to the network via a single subchannel relying on the classic frequency-division multiple access (FDMA). Each device is equipped with a buffer for storing its queuing packets to be transmitted. The placement of the ABS is updated on the temporal basis of a TS. In the following, we detail the system model from the perspectives of the physical layer as well as of the wireless tele-traffic dynamics and then formulate a minimum average-delay ABS placement problem.

A. Physical Layer Model of the UAV-Enabled Network

The air-to-ground channel has been investigated in various research contributions [13]–[21], where it was assumed to obey a probabilistic line-of-sight (LoS) channel [13]–[17], Rician fading [18]–[20] and Nakagami-m fading [21]. Here we assume that the channel between the ABS and devices obeys the probabilistic LoS model, where the link can be either of LoS or of non-LoS (NLoS) nature. The probability of the LoS link is given by [13]

$$P_{\text{LoS}}^{k} = \frac{1}{1+\psi \exp\left[-\beta(\theta_k - \psi)\right]}, \tag{1}$$

where ψ and β are constant values that are determined by the carrier frequency and the surrounding environments; $\theta_k = \frac{180}{\pi} \times \sin^{-1}\left(\frac{h[t]}{D_k[t]}\right)$ denotes the elevation angle; $D_k[t] = \sqrt{\left(u_x[t]-d_x^k\right)^2 + \left(u_y[t]-d_y^k\right)^2 + h[t]^2}$ corresponds to the distance between the ABS and the k-th device at the t-th TS. Then, the NLoS probability can be calculated as $P_{\text{NLoS}}^{k} = 1 - P_{\text{LoS}}^{k}$. Hence the channel gain between the ABS and the k-th device at the t-th TS is readily given by [27]

$$g_k[t] = \left\{\varrho D_k[t]\right\}^{-2}\left(P_{\text{LoS}}^{k}\mu_{\text{LoS}} + P_{\text{NLoS}}^{k}\mu_{\text{NLoS}}\right)^{-1}, \tag{2}$$

where we have $\varrho = \frac{4\pi f_c}{c}$; f_c and c represent the carrier frequency and the speed of light, respectively; μ_{LoS} and μ_{NLoS} are the attenuation factors considered for LoS and NLoS links, respectively. Here we use $\boldsymbol{G}[t] = \left(g_1[t], \ldots, g_K[t]\right)$ to denote the channel state information (CSI) between the ABS and these K devices.

For the sake of simplicity, we assume that an idealized-capacity-achieving-coding scheme [58] is invoked and that the coordinates of the ABS and those of the devices are perfectly known. Then, the transmission rate of the communications link between the ABS and the k-th device at the t-th TS is given by

$$R_k[t] = B\log_2\left(1 + \frac{P_t g_k[t]}{B\sigma^2}\right), \tag{3}$$

where B, P_t and σ^2 denote the subchannel bandwidth, the transmit power and the power spectral density of the zero-mean white Gaussian noise at the receiver, respectively.

B. Queuing Model and System Dynamics

The number of packets in the buffer (also termed by the queue length) at the beginning of the $(t+1)$-th TS is jointly determined by the queue length at the beginning of the t-th TS as well as by the number of arriving and departing packets during the t-th TS. Here let us denote the number of the k-th device's arriving packets during the t-th TS by $B_k[t]$. We assume that the packet arrival process is independent and identical distributed (i.i.d.) over the TSs and its mean value is denoted by $\lambda_k = \mathbb{E}\{B_k[t]\}$, for $\forall k \in \{1, 2, \ldots, K\}$. Here we use $Q_k[t]$ to denote the queue length of the k-th device at the beginning of the t-th TS and $\boldsymbol{Q}[t] = \left(Q_1[t], \ldots, Q_K[t]\right)$ to represent the joint queue length state information (QSI) for these K devices. More particularly, $Q_k[t]$ evolves by obeying the equation below:

$$Q_k[t+1] = \min\left\{\left(Q_k[t] - \frac{R_k[t]\tau}{N_k}\right)^{+} + B_k[t], N_Q\right\}, \tag{4}$$

for $\forall t \in \{0, 1, \ldots, T-1\}$, where we define $x^{+} = \max(x, 0)$; N_k is the packet size of the k-th user; N_Q represents the maximum number of packets that can be stored in the buffer. Without loss of generality, we assume $Q_k[0] = 0$, for $\forall k \in \{1, 2, \ldots, K\}$.

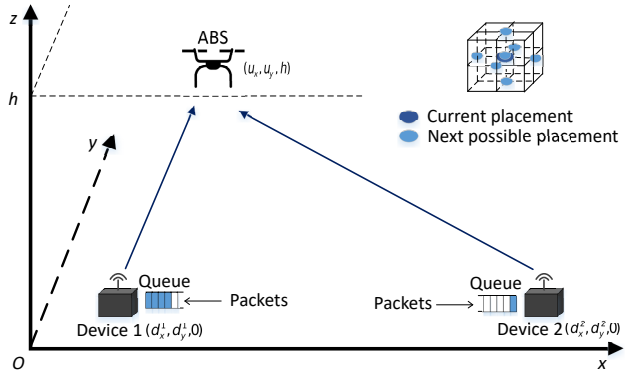

Figure 1: Illustration of the system model. A UAV-enabled network comprises an ABS located at (u_x, u_y, h) and a number of devices located at $(d_x^k, d_y^k, 0)$. Each device is equipped with a buffer for storing its queuing packets. The movement of the ABS follows the rule as depicted in the top-right figure.

C. ABS Placement Scheduling

The ABS is equipped with a placement scheduler, which is capable of dynamically adjusting its placement in accordance with the joint CSI and QSI on the temporal basis of a time slot τ. Specifically, we realize the scheduling strategy relying on a 3-D grid associated with the horizontal basis of $\delta_h = v_h\tau$ and with the vertical basis of $\delta_v = v_v\tau$, where v_h and v_v represent the horizontal and vertical velocity of the ABS, respectively. As depicted in Fig. 1, at each TS the ABS can be scheduled to stay at the previous location or to move to one of its six surrounding points on the grid [3].

D. Minimum-Delay Control Problem

1) Constraints: Our minimum-delay ABS placement problem is formulated under the following constraints:

[3]Without any loss of generality, our proposed framework and solutions are also applicable to the ABS scheduling strategies that are based on more complex movement patterns, e.g. following the hexagonal grid, where the delay can be further reduced.

- *Power consumption constraint:* since the ABS is typically equipped with a capacity-limited battery [7], our proposed scheduling scheme has to satisfy a realistic energy consumption constraint. Specifically, the total energy consumption of the ABS is the sum of both the communications-related and decision-making signal processing dissipation as well as of the mechanical propulsion. Here let us use E_{tot} to represent the total energy in the battery of the ABS. We denote its power consumption of communications, horizontal movement, vertically up movement, vertically down movement, maintaining aloft and decision making by P_c, P_m^h, P_m^{v+}, P_m^{v-}, P_h and P_d, respectively. To elaborate, P_c represents the power consumption on communications, including the carrier frequency down conversion, power amplifying and baseband signal processing; if the ABS stays at the previous location, its power consumption on mechanical propulsion is given by P_h, whereas if the ABS moves from the previous location to one of four possible locations in the same horizon, its power consumption on mechanical propulsion is formulated by $P_h + P_m^h$; if the ABS moves one step vertically up, the power consumption on mechanical propulsion is $P_h + P_m^{v+}$. To this end, we may formulate the average power constraint as (7) in the next page, where we define $P_{avg} = \frac{E_{tot}}{T\tau} - P_h - P_c - P_d$ and the operations as

$$(x)^+ = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise;} \end{cases} \tag{5}$$

and

$$(x)^- = \begin{cases} |x|, & \text{if } x < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

- *Grid constraint:* The ABS dynamically moves among the points on the grid associated with the basis of $\{\delta_h, \delta_v\}$. Hence at each TS the coordinate of the ABS has to satisfy the constraint $u_x[t] \in \{\underline{x}, \ldots, -\delta_h, 0, \delta_h, \ldots, \overline{x}\}$, $u_y[t] \in \{\underline{y}, \ldots, -\delta_h, 0, \delta_h, \ldots, \overline{y}\}$, and $h[t] \in \{\underline{h}, \underline{h}+\delta_v, \ldots, \overline{h}-\delta_v, \overline{h}\}$, where the bounds $\underline{x}$, $\overline{x}$, $\underline{y}$, $\overline{y}$, $\underline{h}$ and $\overline{h}$ restrict the 3-D placement of the ABS.
- *Speed constraint:* As illustrated in Section II-C, at each TS the ABS can stay at the location of the previous TS or move to other horizontally surrounding locations at the velocity of v_h or other vertical surrounding locations at the velocity of v_v. Given that we have defined $\delta_h = v_h\tau$ and $\delta_v = v_v\tau$, the speed constraint is formulated by $\left|\left|\boldsymbol{U}[t] - \boldsymbol{U}[t-1]\right|\right|^2 \in \{0, \delta_h^2, \delta_v^2\}, \forall t \in \{1, 2, \ldots, T\}$.

2) Problem Formulation: We focus our attention on the queuing delay in the network layer, which is defined as the temporal interval between the instant when a packet arrives at the transmitter and the instant when it is delivered [59]. Our objective is to minimize the average queuing delay over T TSs in the ABS-enabled network. By Little's Law [60], the relationship among the average delay denoted by $\overline{D}$, average queue length and packet arrival rate is given by [49]:

$$\overline{D}_k = \frac{\mathbb{E}\{Q_k[t]\}}{\lambda_k} = \mathbb{E}\left\{\frac{Q_k[t]}{\lambda_k}\right\}. \tag{8}$$

Now, given an initial state of $\boldsymbol{X}[0] = \{\boldsymbol{U}[0], \boldsymbol{Q}[0]\}$, the minimum-delay ABS placement problem is readily formulated as:

$$\mathcal{P}0: \arg\min_{\boldsymbol{U}[t]} \frac{1}{TK} \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{E}\left\{\frac{w_k Q_k[t]}{\lambda_k} \middle| \boldsymbol{X}[0]\right\}$$

s.t. (7)

$$\left|\left|\boldsymbol{U}[t] - \boldsymbol{U}[t-1]\right|\right|^2 \in \{0, \delta_h^2, \delta_v^2\}, \forall t \in \{1, 2, \ldots, T\}, \tag{9a}$$

$$u_x[t] \in \{\underline{x}, \ldots, -\delta_h, 0, \delta_h, \ldots, \overline{x}\}, \forall t \in \{0, 1, \ldots, T\}, \tag{9b}$$

$$u_y[t] \in \{\underline{y}, \ldots, -\delta_h, 0, \delta_h, \ldots, \overline{y}\}, \forall t \in \{0, 1, \ldots, T\}, \tag{9c}$$

$$h[t] \in \{\underline{h}, \underline{h} + \delta_v, \ldots, \overline{h} - \delta_v, \overline{h}\}, \forall t \in \{0, 1, \ldots, T\}, \tag{9d}$$

where the positive weighting factor w_k indicates the relative importance of the queuing delay of the k-th device, which is dependent on the device's priority.

III. Markov Decision Process Transformation

In a stochastic dynamic control problem, the controller can make decisions in accordance with the environment's state. If the states satisfies the Markov property, this decision process can be termed as a Markov decision process (MDP) [61]. As a special case of the MDP, the constrained MDP (CMDP) [62] has multiple objectives. It enables minimizing one of the objectives, while satisfying the constraints imposed on the others. In this section, the minimum-delay control problem formulated in Section II-D is transformed to its corresponding CMDP problem. By further exploiting the Lagrangian approach, we then reformulate this CMDP problem as an unconstrained MDP problem, which can then be readily solved using various approaches, as detailed in Section IV.

A. Constrained Markov Decision Process

In general, a CMDP can be characterized by four elements, namely the state space, the action space, the state transition probability and the constrained optimization problem [62], which are specified for our minimum-delay control problem formulated in Section II-D as follows:

- **State space:** Let us denote the system state at the t-th TS by $\boldsymbol{X}[t]$, which is defined as the aggregation of the ABS's location and the global QSI, i.e., $\boldsymbol{X}[t] = \{\boldsymbol{U}[t], \boldsymbol{Q}[t]\}$, where $\forall t \in \{0, 1, \ldots, T\}$. Here let us use $\boldsymbol{\mathcal{X}}$ to represent the set including all possible states. The number of elements in $\boldsymbol{\mathcal{X}}$ is denoted by $|\boldsymbol{\mathcal{X}}|$. It can be readily inferred that the future states of both the ABS location $\boldsymbol{U}[t]$ and of the joint QSI $\boldsymbol{Q}[t]$ are only dependent on their current states, but not on the states at the previous TSs. Hence, the states evolved as a controlled Markov chain and this decision process is an MDP.
- **Action space:** We use $A[t]$ to represent the action taken at the beginning of the t-th TS. For $\forall t \in \{1, 2, \ldots, T-1\}$, one of the seven actions can be taken from the action space denoted by $\boldsymbol{\mathcal{A}} = \{A_0, A_1, A_2, A_3, A_4, A_5, A_6\}$, where A_0 means $\boldsymbol{u}[t] = (u_x[t-1], u_y[t-1], h[t-1])$; A_1, A_2, A_3, A_4, A_5 and A_6 refer to $\boldsymbol{u}[t] = (u_x[t-1] - \delta_h, u_y[t-1], h[t-1])$, $\boldsymbol{u}[t] = (u_x[t-1] + \delta_h, u_y[t-1], h[t-1])$, $\boldsymbol{u}[t] = (u_x[t-1], u_y[t-1] - \delta_h, h[t-1])$, $\boldsymbol{u}[t] = (u_x[t-1], u_y[t-1] + \delta_h, h[t-1])$, $\boldsymbol{u}[t] =$

$$\frac{P_m^h}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{\frac{|u_x[t]-u_x[t-1]|}{\delta_h}+\frac{|u_y[t]-u_y[t-1]|}{\delta_h}\Big\}+\frac{P_m^{v+}}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{\frac{(h[t]-h[t-1])^+}{\delta_v}\Big\}+\frac{P_m^{v-}}{T}\sum_{t=1}^{T}\mathbb{E}\Big\{\frac{(h[t]-h[t-1])^-}{\delta_v}\Big\}\leq P_{avg}. \quad (7)$$

$(u_x[t-1], u_y[t-1], h[t-1]-\delta_v)$, $\boldsymbol{u}[t]=(u_x[t-1], u_y[t-1], h[t-1]+\delta_v)$, respectively. The total number of actions is denoted by $|\boldsymbol{\mathcal{A}}|$. Furthermore, the CMDP formulated in this section aims for finding the optimal ABS placement policy $\Omega^*: \boldsymbol{\mathcal{X}} \mapsto \boldsymbol{\mathcal{A}}$ that minimizes the average delay subject to the power constraint, from the set denoted by $\boldsymbol{\Omega}$, which includes all possible policies.

- **Transition kernel:** Again, since the MDP is a controlled Markov chain, the state transition is determined by both (4) and by the actions taken at the beginning of the TSs. The state transition probability is formulated by:

$$\begin{aligned}&\Pr\Big[\boldsymbol{X}[t+1]\Big|\boldsymbol{X}[t],\Omega\big(\boldsymbol{X}[t]\big)\Big] \quad (10)\\ &=\Pr\Big[\boldsymbol{U}[t+1]\Big|\boldsymbol{U}[t],\Omega\big(\boldsymbol{X}[t]\big)\Big]\Pr\Big[\boldsymbol{Q}[t+1]\Big|\boldsymbol{Q}[t],\Omega\big(\boldsymbol{X}[t]\big)\Big].\end{aligned}$$

- **Constrained optimization problem:** Commencing from an initial state $\boldsymbol{X}[0]=\mathbf{X}_0\in\boldsymbol{\mathcal{X}}$ and following a policy $\Omega\in\boldsymbol{\Omega}$, the average expected delay and power consumption are defined as

$$\overline{D}^{\Omega}(\mathbf{X}_0)=\frac{1}{TK}\sum_{t=1}^{T}\sum_{k=1}^{K}\mathbb{E}^{\Omega}\Big\{\frac{w_kQ_k[t]}{\lambda_k}\Big|\boldsymbol{X}[0]=\mathbf{X}_0\Big\} \quad (11)$$

and (12) in the next page, respectively, where $\mathbb{E}^{\Omega}(\bullet)$ represents the expectation value of $\bullet$ under the policy Ω. Then, we may formulate the constrained optimization version of the minimum-delay ABS placement control problem as

$$\mathcal{P}1:\ \min_{\Omega\in\boldsymbol{\Omega}}\overline{D}^{\Omega}(\mathbf{X}_0)\ \text{subject to}\ \overline{P}^{\Omega}(\mathbf{X}_0)\leq P_{avg}. \quad (13)$$

Remark 1. *There is a one-to-one correspondence between Problem* $\mathcal{P}0$ *and Problem* $\mathcal{P}1$. *Specifically, the constraint* (7) *in Problem* $\mathcal{P}0$ *corresponds to the constraint in Problem* $\mathcal{P}1$, *while the constraints* (9a), (9b), (9c) *and* (9d) *restrict the action taken at each TS into the afore-specified action space in Problem* $\mathcal{P}1$.

B. The Lagrangian Approach

As analyzed in [63], the MDP problem falls into the class of convex programming problems. Therefore, solving the constrained MDP problem is equivalent to solving the unconstrained MDP associated with its Lagrangian dual problem [62]. In this subsection, we reformulate the CMDP in (13) to be an unconstrained MDP by introducing a Lagrange multiplier γ, following the theorem below.

Theorem 1. *There is a one-to-one correspondence between the constrained MDP formulated in* (13) *and the unconstrained MDP formulated below:*

$$\begin{aligned}\overline{C}^*(\mathbf{X}_0)&=\inf_{\Omega\in\boldsymbol{\Omega}}\sup_{\gamma\geq 0}\Big\{\overline{D}^{\Omega}(\mathbf{X}_0)+\gamma\Big[\overline{P}^{\Omega}(\mathbf{X}_0)-P_{avg}\Big]\Big\}\\ &=\sup_{\gamma\geq 0}\inf_{\Omega\in\boldsymbol{\Omega}}\Big\{\overline{D}^{\Omega}(\mathbf{X}_0)+\gamma\Big[\overline{P}^{\Omega}(\mathbf{X}_0)-P_{avg}\Big]\Big\} \quad (14)\end{aligned}$$

and a policy Ω^* *is optimal for the CMDP if and only if*

$$\overline{C}^*(\mathbf{X}_0)=\sup_{\gamma\geq 0}\Big\{\overline{D}^{\Omega^*}(\mathbf{X}_0)+\gamma\Big[\overline{P}^{\Omega^*}(\mathbf{X}_0)-P_{avg}\Big]\Big\}. \quad (15)$$

Proof: See Appendix A. ∎

With the aid of Theorem 1, we may transform the CMDP of Section III-A to an unconstrained MDP. Specific to a fixed γ, we use $C_\gamma(\boldsymbol{X}[t],\boldsymbol{X}[t+1],A[t])$ to denote *the per-stage cost function* of the corresponding unconstrained MDP that emerges from the state $\boldsymbol{X}[t]$ to the state $\boldsymbol{X}[t+1]$ following the action $A[t]$. Its expression is given by (16) in the next page. To this end, we may readily formulate the corresponding unconstrained MDP problem as:

$$\overline{C}^*(\mathbf{X}_0)=\sup_{\gamma\geq 0}\inf_{\Omega\in\boldsymbol{\Omega}}\sum_{t=0}^{T-1}\mathbb{E}^{\Omega}\Big\{C_\gamma\big(\boldsymbol{X}[t],\boldsymbol{X}[t+1],A[t]\big)\Big|\boldsymbol{X}[0]=\mathbf{X}_0\Big\}, \quad (17)$$

where we have $\Omega:\boldsymbol{\mathcal{X}}\mapsto\boldsymbol{\mathcal{A}}$.

IV. Solutions to the Minimum-Delay MDP Problem

Since the actions described in Section III-A may impose causality constraints on the consecutive states, we have to solve this MDP problem by using dynamic programming instead of solving it in each TS independently. In general, an MDP problem can be solved by the techniques of backward induction, policy iteration, value iteration and reinforcement learning [4] [61]. Specifically, backward induction is appropriate for the problem associated with a finite number of TSs, while both the policy iteration and value iteration aim for providing the policy for the MDP problem associated with an infinite number of TSs. The apriori knowledge of the state transition probability is required by all of these three techniques. By contrast, reinforcement learning interacts with the environment through the "trial and error" mechanism and hence does not require the exact mathematical model of the MDP problem. Given the finite number of TSs in the problem formulated in (17), we provide a set of solutions for diverse assumptions of the prior knowledge of wireless tele-traffic in this section, by using the techniques of backward induction or of reinforcement learning.

The key idea both of the backward induction and of the reinforcement learning techniques is to invoke the so-called *value functions* for finding appropriate policies [61]. Given a

[4]Machine learning techniques can be generally classified into supervised and unsupervised learning as well as reinforcement learning [64]. Specifically, both the supervised and unsupervised learning techniques are typically used for classification and clustering, while reinforcement learning aims for assisting decision making for the MDP problem, where the state transition probability is unknown.

$$\overline{P}^{\Omega}(\mathbf{X}_0)=\sum_{t=1}^{T}\mathbb{E}^{\Omega}\Big\{\frac{P_m^h}{T}\Big(\frac{\big|u_x[t]-u_x[t-1]\big|}{\delta_h}+\frac{\big|u_y[t]-u_y[t-1]\big|}{\delta_h}\Big)+\frac{P_m^{v+}}{T}\frac{\big(h[t]-h[t-1]\big)^{+}}{\delta_v}+\frac{P_m^{v-}}{T}\frac{\big(h[t]-h[t-1]\big)^{-}}{\delta_v}\Big|\boldsymbol{X}[0]=\mathbf{X}_0\Big\}. \quad (12)$$

$$C_{\gamma}(\boldsymbol{X}[t],\boldsymbol{X}[t+1],A[t])=\sum_{k=1}^{K}\frac{w_kQ_k[t+1]}{TK\lambda_k}+\gamma\Big\{\frac{P_m^h}{T}\Big(\frac{\big|u_x[t+1]-u_x[t]\big|}{\delta_h}+\frac{\big|u_y[t+1]-u_y[t]\big|}{\delta_h}\Big)+\frac{P_m^{v+}}{T}\frac{\big(h[t+1]-h[t]\big)^{+}}{\delta_v}+\frac{P_m^{v-}}{T}\frac{\big(h[t+1]-h[t]\big)^{-}}{\delta_v}-P_{avg}\Big\}. \quad (16)$$

fixed γ and a policy Ω, let us denote the state-value function at the state $\mathbf{X}$ by $v_{\gamma}^{\Omega}(\mathbf{X})$, which satisfies the Bellman expectation equation as follows [61]:

$$v_{\gamma}^{\Omega}(\mathbf{X})=\sum_{A\in\boldsymbol{\mathcal{A}}}\Pr^{\Omega}(\mathbf{X},A)\sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}}\Pr(\mathbf{X}'|\mathbf{X},A)\cdot\big[C_{\gamma}(\mathbf{X},\mathbf{X}',A)+v_{\gamma}^{\Omega}(\mathbf{X}')\big], \quad (18)$$

where $\Pr^{\Omega}(\mathbf{X},A)$ refers to the probability of choosing the action A at the state $\mathbf{X}$ under the policy Ω. To elaborate a little further, since $\Pr^{\Omega}(\mathbf{X},A)$ is determined by a specific policy and the term $C_{\gamma}(\mathbf{X},\mathbf{X}',A)$ in (18) can be readily calculated by (16), we focus our attention on $\Pr(\mathbf{X}'|\mathbf{X},A)$. Specifically, given a state and an action taken at the t-th TS, the state at the $(t+1)$-st TS is solely determined by $B_k[t]$ in (4). Hence, once we have the knowledge of $B_k[t]$, the state transition probability of $\Pr(\mathbf{X}'|\mathbf{X},A)$ can be readily determined. Without any loss of generality, again we consider three different assumptions in terms of the knowledge of $B_k[t]$ as detailed below:

- **Case 1:** The exact value of $B_k[t]$ for $\forall t\in\{0,1,\ldots,T-1\}$ and $\forall k\in\{1,2,\ldots,K\}$ is known by the ABS placement scheduler;
- **Case 2:** The specific probability density function of $B_k[t]$ for $\forall k\in\{1,2,\ldots,K\}$ is available at the ABS placement scheduler;
- **Case 3:** Neither the exact value nor the statistical information of $B_k[t]$ for $\forall k\in\{1,2,\ldots,K\}$ is known at the ABS placement scheduler.

Remark 2. *The stochastic process of wireless tele-traffic dynamics is assumed to be accurately predicted in Case 1. It can be realized relying on wireless tele-traffic prediction techniques [65]. Additionally, some periodic data transmission schemes also fall into this category. Case 2 is suitable for the scenario, where the devices' wireless tele-traffic dynamics are accurately modeled. Finally, Case 3 considers the scenario, where we do not have any prior information.*

In the following, a set of solutions are provided for these three cases, respectively.

A. Solution to the Problem in Case 1

Given the exact value of $B_k[t]$ for $\forall t\in\{0,1,\ldots,T-1\}$ and $\forall k\in\{1,2,\ldots,K\}$, the problem formulated in (17) can be simplified to a deterministic form

$$\overline{C}^{*}(\mathbf{X}_0)=\sup_{\gamma\geq 0}\inf_{\Omega\in\boldsymbol{\Omega}}\frac{1}{T}\sum_{t=0}^{T-1}\Big\{C_{\gamma}\big(\boldsymbol{X}[t],\boldsymbol{X}[t+1],A[t]\big)\Big|\boldsymbol{X}[0]=\mathbf{X}_0,\Omega\Big\}. \quad (19)$$

Algorithm 1 Backward induction based approach for Case 1

Input: Simultaneous knowledge of wireless tele-traffic dynamics
Output: Minimum-delay ABS placement policy
1. Initialization
initialize γ arbitrarily
for each state $\mathbf{X}\in\boldsymbol{\mathcal{X}}$ **do**
  $v_{\gamma}^{T}(\mathbf{X})\leftarrow 0$
end for
2. State-value table generation
for each TS t from $T-1$ to 0 **do**
  $v_{\gamma}^{t}(\mathbf{X})\leftarrow\min_{A\in\boldsymbol{\mathcal{A}}}\big[C_{\gamma}(\mathbf{X},\mathbf{X}',A)+v_{\gamma}^{t+1}(\mathbf{X}')\big]$
end for
3. Average cost evaluation and γ update
if $\overline{C}_{\gamma}(\mathbf{X}_0)$ cannot be improved **then**
  $v_{*}^{t}(\mathbf{X})\leftarrow v_{\gamma}^{t}(\mathbf{X}),\forall t\in\{0,1,\ldots,T\}$
  go to step 4)
else
  update γ using the bisection search algorithm and go to step 2)
end if
4. Policy output
for each TS t from 0 to $T-1$ **do**
  $\Omega^{t}(\mathbf{X})\leftarrow\arg\min_{A\in\boldsymbol{\mathcal{A}}}\big[C_{\gamma}(\mathbf{X},\mathbf{X}',A)+v_{*}^{t+1}(\mathbf{X}')\big]$
end for

This can be solved by the technique of backward induction [66]. The basic idea is to establish a trellis that comprises all possible states at each TS and then obtain a table that contains the aforementioned state-value functions for all the states in each TS sequentially from the final TS to the beginning. Based on this table, the policy can be implemented, commencing from the initial TS. Here let us denote the state-value function of the state $\mathbf{X}$ at the t-th TS by $v_{\gamma}^{t}(\mathbf{X})$, given a specific γ. We may follow the steps below for establishing a table containing all $v_{\gamma}^{t}(\mathbf{X})$.

1) Since the delay is characterized in terms of a finite number of TSs T as formulated in (19), the states at the T-th TS do not impose further cost [5]. Therefore, we set the state-value functions at the T-th TS to 0, i.e., $v_{\gamma}^{T}(\mathbf{X})=0,\forall\mathbf{X}\in\boldsymbol{\mathcal{X}}$.
2) Given that $B_k[t]$ is pre-acknowledged in each TS, its randomness vanishes in the decision making process. In other words, the state at the next TS is solely determined both by the state at the current TS and by the action taken. Therefore, in this case we may update $v_{\gamma}^{t}(\mathbf{X})$ for the TS $t=T-1,\ldots,0$ sequentially relying on the Bellman optimality equation [61] as follows:

$$v_{\gamma}^{t}(\mathbf{X})=\min_{A\in\boldsymbol{\mathcal{A}}}\big[C_{\gamma}(\mathbf{X},\mathbf{X}',A)+v_{\gamma}^{t+1}(\mathbf{X}')\big]. \quad (20)$$

[5]This is in accordance with the convention that costs occur at the next TS. Therefore, the states at the T-th TS do not impose further cost.

Algorithm 2 Backward induction based approach for Case 2

Input: Statistical knowledge of wireless tele-traffic dynamics
Output: Minimum-delay ABS placement policy
1. Initialization
initialize γ arbitrarily
for each state $\mathbf{X} \in \boldsymbol{\mathcal{X}}$ **do**
$v_\gamma^T(\mathbf{X}) \leftarrow 0$
end for
2. State-value table generation
for each TS t from $T-1$ to 0 **do**
$v_\gamma^t(\mathbf{X}) \leftarrow \min_{A\in\boldsymbol{\mathcal{A}}} \sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}} \Pr(\mathbf{X}'|\mathbf{X},A)\cdot\big[C_\gamma(\mathbf{X},\mathbf{X}',A) + v_\gamma^{t+1}(\mathbf{X}')\big]$
end for
3. Average cost evaluation and γ update
if $\overline{C}_\gamma(\mathbf{X}_0)$ cannot be improved **then**
$v_*^t(\mathbf{X}) \leftarrow v_\gamma^t(\mathbf{X})$, $\forall t \in \{0,1,\ldots,T\}$
go to step 4)
else
update γ using the bisection search algorithm and go to step 2)
end if
4. Policy output
for each TS t from 0 to $T-1$ **do**
$\Omega^t(\mathbf{X}) \leftarrow \arg\min_{A\in\boldsymbol{\mathcal{A}}} \sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}} \Pr(\mathbf{X}'|\mathbf{X},A)\cdot\big[v_*^{t+1}(\mathbf{X}') + C_\gamma(\mathbf{X},\mathbf{X}',A)\big]$
end for

3) For a fixed γ, the average cost $\overline{C}_\gamma(\mathbf{X}_0)$ can be readily obtained by $\overline{C}_\gamma(\mathbf{X}_0) = v_\gamma^0(\mathbf{X}_0)$.
4) Following (19), we update γ using the classic bisection search algorithm and go to Step 2) until a maximum average cost $\overline{C}^*(\mathbf{X}_0)$ is obtained. Here we denote its corresponding optimal state-value function by $v_*^t(\mathbf{X})$, for $\forall t \in \{0,1,\ldots,T\}$ and $\forall \mathbf{X} \in \boldsymbol{\mathcal{X}}$.

Based on the optimal average cost $\overline{C}^*(\mathbf{X}_0)$ and the corresponding $v_*^t(\mathbf{X})$, we may then carry out the policy by setting $\boldsymbol{X}[0] = \mathbf{X}_0$ and then by solving the equation below

$$\Omega^t(\mathbf{X}) = \arg\min_{A\in\boldsymbol{\mathcal{A}}} \big[C_\gamma(\mathbf{X},\mathbf{X}',A) + v_*^{t+1}(\mathbf{X}')\big], \tag{21}$$

from the TS $t=0$ to $t=T-1$ sequentially. The Pseudocode of the backward induction based solution of the problem in Case 1 is given by Algorithm 1.

B. Solution to the Problem in Case 2

Given the finite number of TSs and the statistical information concerning the packet arrival process, the problem of Case 2 can be solved by the technique of backward induction. However, instead of following the deterministic formulations specified in (20) and (21) in Case 1, in Case 2 we have to update the state-value functions and implement the policies using the statistical information available. The steps required for establishing a table containing all $v_\gamma^t(\mathbf{X})$ are detailed as follows.

1) Again, we set the state-value functions at the T-th TS to 0, i.e., $v_\gamma^T(\mathbf{X}) = 0, \forall \mathbf{X} \in \boldsymbol{\mathcal{X}}$.
2) Given the probability density function of $B_k[t]$, we are ready to calculate $\Pr(\mathbf{X}'|\mathbf{X},A)$. Then, the state value functions $v_\gamma^t(\mathbf{X})$ can be updated from the TS $t=T-1$ to $t=0$ by sequentially invoking the Bellman optimality equation of [61] as follows:

$$v_\gamma^t(\mathbf{X}) = \min_{A\in\boldsymbol{\mathcal{A}}} \sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}} \Pr(\mathbf{X}'|\mathbf{X},A) \cdot\big[C_\gamma(\mathbf{X},\mathbf{X}',A) + v_\gamma^{t+1}(\mathbf{X}')\big]. \tag{22}$$

3) For a fixed γ, the average cost $\overline{C}_\gamma(\mathbf{X}_0)$ can be readily obtained by $\overline{C}_\gamma(\mathbf{X}_0) = v_\gamma^0(\mathbf{X}_0)$.
4) Following (17), we update γ using the bisection search algorithm and go to Step 2) until a maximum average cost $\overline{C}^*(\mathbf{X}_0)$ is obtained. Here we denote its corresponding optimal state-value function by $v_*^t(\mathbf{X})$ for $\forall t \in \{0,1,\ldots,T\}$ and $\forall \mathbf{X} \in \boldsymbol{\mathcal{X}}$.

Based on the optimal average cost $\overline{C}^*(\mathbf{X}_0)$ and the corresponding $v_*^t(\mathbf{X})$, we may then implement the policy by setting $\boldsymbol{X}[0] = \mathbf{X}_0$ and by solving the equation below:

$$\Omega^t(\mathbf{X}) = \arg\min_{A\in\boldsymbol{\mathcal{A}}} \sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}} \Pr(\mathbf{X}'|\mathbf{X},A) \cdot\big[C_\gamma(\mathbf{X},\mathbf{X}',A) + v_*^{t+1}(\mathbf{X}')\big], \tag{23}$$

from the TS $t=0$ to $t=T-1$ sequentially. Note that $\Pr(\mathbf{X}'|\mathbf{X},A) = 0$ for the set of $\{\mathbf{X},\mathbf{X}',A\}$, where the state $\mathbf{X}$ cannot reach state $\mathbf{X}'$ after executing the action A. The Pseudocode of the backward induction based solution to the problem in Case 2 is given by Algorithm 2.

Algorithm 3 R-learning based approach for Case 3

Input: Neither simultaneous nor statistical knowledge of wireless tele-traffic dynamics
Output: Reduced-delay ABS placement policy
1. Initialization
initialize γ arbitrarily, $\boldsymbol{X}[0] \leftarrow \mathbf{X}_0$
for each state $\mathbf{X} \in \boldsymbol{\mathcal{X}}$ and $A \in \boldsymbol{A}$ **do**
$R_\gamma^0(\mathbf{X},A) \leftarrow 0$
end for
2. Action-value table generation
for each TS t **do**
set $\mathbf{X} \leftarrow \boldsymbol{X}[t]$
select an action A following the ϵ-greedy method
execute the action A
observe the next state $\mathbf{X}'$
receive the immediate cost $C_\gamma(\mathbf{X},\mathbf{X}',A)$
update the $R_\gamma(\mathbf{X},A)$ by $R_\gamma^{t+1}(\mathbf{X},A) \leftarrow (1-\eta)R_\gamma^t(\mathbf{X},A) + (\mathbf{X},A) + \eta\big[C_\gamma(\mathbf{X},\mathbf{X}',A) - \rho_\gamma^t + \min_{A'\in\boldsymbol{\mathcal{A}}} R_\gamma^t(\mathbf{X}',A')\big]$
update ρ_γ by $\rho_\gamma^{t+1} \leftarrow (1-\alpha)\rho_\gamma^t + \alpha\big[C_\gamma(\mathbf{X},\mathbf{X}',A) + \min_{A'\in\boldsymbol{\mathcal{A}}} R_\gamma^t(\mathbf{X}',A') - \min_{A\in\boldsymbol{\mathcal{A}}} R_\gamma^t(\mathbf{X},A)\big]$
update $\boldsymbol{X}[t+1] \leftarrow \mathbf{X}'$
end for
3. Policy output
$\Omega_\gamma(\mathbf{X}) = \arg\min_{A\in\boldsymbol{\mathcal{A}}} R_\gamma(\mathbf{X},A)$
4. Constraint satisfaction evaluation and γ update
if the equality of (7) holds **then**
$R_*^t(\mathbf{X},A) \leftarrow R_\gamma^t(\mathbf{X},A)$
$\Omega_*(\mathbf{X}) \leftarrow \Omega_\gamma(\mathbf{X})$
else
update γ using the bisection search algorithm and go to step 2)
end if

C. Solution to the Problem in Case 3

As for the case where $\Pr(\mathbf{X}|\mathbf{X}',A)$ is unknown to the scheduler, we have to rely on the technique of reinforcement

learning [61], which enables the scheduler to carry out the policy by interacting with the environment. As a classic reinforcement learning technique, Q-learning [61, Ch. 6] has been leveraged in diverse research areas. However, its cost is accumulated in a discounted manner for future TSs and may not solve the MDP problem formulated in (17) that is associated with undiscounted costs. To address this issue, we invoke the technique of R-learning [67], which has been tailored for the problem associated with the undiscounted costs.

In generally, the horizon of a reinforcement learning problem is assumed to be infinite. The problem formulated in (17), however, is readily observed to be a finite-horizon MDP. As detailed in Remark 3, we have to train the scheduler in an offline manner and hence it can be trained over a large number of episodes for approaching the infinite-horizon performance. In infinite-horizon problems associated with undiscounted costs, the state-value function in (18) becomes infinite, which cannot be used as a comparative basis, when we implement the policy. In order to tackle this issue, the concept of the action value $R^{\Omega}(\mathbf{X}, A)$ is introduced into the technique of R-learning [67], which represents the average adjusted value of carrying out an action A in State $\mathbf{X}$ once and then following the policy Ω [67]. Mathematically, $R^{\Omega}(\mathbf{X}, A)$ is given as follows [67]:

$$R_{\gamma}^{\Omega}(\mathbf{X}, A) = C_{\gamma}(\mathbf{X}, \mathbf{X}', A) - \rho_{\gamma}^{\Omega} + \sum_{\mathbf{X}'} \Pr(\mathbf{X}'|\mathbf{X}, A) v_{\gamma}^{\Omega}(\mathbf{X}'), \tag{24}$$

where ρ^{Ω} is the average cost of the policy Ω. Given a specific value of γ, let us detail the steps of generating a table containing $R^{\Omega}(\mathbf{X}, A)$ as follows.

1) We set the initial average-adjusted value to $R_{\gamma}^{0}(\mathbf{X}, A) = 0$, $\forall \mathbf{X} \in \boldsymbol{\mathcal{X}}$ and $\forall A \in \boldsymbol{\mathcal{A}}$. The initial state is set to $\boldsymbol{X}[0] = \mathbf{X}_0$.
2) The actions are chosen using *the exploration/exploitation selection mechanism* [61]. Specifically, the term *exploitation* means that we opt for an action following the policy, which minimizes the average-adjust value functions. Mathematically, we have $\Omega_{\gamma}(\mathbf{X}) = \arg\min_{A \in \boldsymbol{\mathcal{A}}} R_{\gamma}^{\Omega}(\mathbf{X}, A)$. However, before obtaining a set of reliable $R_{\gamma}^{\Omega}(\mathbf{X}, A)$ values, the action taken following this policy is not deemed to be satisfactory. To overcome this hindrance, the concept of *exploration* is introduced to randomly select an action in $\boldsymbol{\mathcal{A}}$. This is capable of discovering better policies and of improving the estimate of $R_{\gamma}^{\Omega}(\mathbf{X}, A)$. In particular, we invoke the ϵ-greedy action selection method, which either takes actions randomly (*exploration*) with a probability of ϵ or follows the policy (*exploitation*) with probability $(1-\epsilon)$ at each TS, where $0 < \epsilon < 1$ [68].
3) After executing an action A at a state $\mathbf{X}$, we may observe its subsequent state $\mathbf{X}'$ and the immediate cost $C_{\gamma}(\mathbf{X}, \mathbf{X}', A)$, both of which are used for updating the average cost ρ_{γ}^{t+1} and the average-adjusted value $R_{\gamma}^{t+1}(\mathbf{X}, A)$. Specifically, the average-adjusted value is updated by [67]

$$R_{\gamma}^{t+1}(\mathbf{X}, A) = \eta\big[C_{\gamma}(\mathbf{X}, \mathbf{X}', A) - \rho_{\gamma}^{t} + \min_{A' \in \boldsymbol{\mathcal{A}}} R_{\gamma}^{t}(\mathbf{X}', A')\big] + (1-\eta) R_{\gamma}^{t}(\mathbf{X}, A), \tag{25}$$

where η is the learning rate for the average-adjusted value. Furthermore, if the action A obeys $\Omega_{\gamma}(\mathbf{X}) = \arg\min_{A \in \boldsymbol{\mathcal{A}}} R_{\gamma}^{t}(\mathbf{X}, A)$, i.e, a non-exploratory action is taken, the average cost is updated by [67]

$$\rho_{\gamma}^{t+1} = \alpha\big[C_{\gamma}(\mathbf{X}, \mathbf{X}', A) + \min_{A' \in \boldsymbol{\mathcal{A}}} R_{\gamma}^{t}(\mathbf{X}', A') - \min_{A \in \boldsymbol{\mathcal{A}}} R_{\gamma}^{t}(\mathbf{X}, A)\big] + (1-\alpha)\rho_{\gamma}^{t}, \tag{26}$$

where α is the learning rate of the average cost.
4) Set the current state to $\mathbf{X}'$ and go to Step 2.

Based on the R value $R_{\gamma}(\mathbf{X}, A)$ obtained, we may carry out a stationary policy by solving the equation below:

$$\Omega_{\gamma}(\mathbf{X}) = \arg\min_{A \in \boldsymbol{\mathcal{A}}} R_{\gamma}(\mathbf{X}, A). \tag{27}$$

Given a sequence of wireless tele-traffic over T TSs, we dynamically adjust the ABS' placement following the policy $\Omega_{\gamma}(\mathbf{X})$ and observe the total number of movements, which can be used for checking the satisfaction of constraint (7). If the equality of constraint (7) does not hold, we then update the value of γ using the bisection search algorithm and then carry out the policy $\Omega_{\gamma}(\mathbf{X})$ until the equality holds.

Remark 3. *Policies designed for the problems in both Case 1 and Case 2 belong to the Markov policy [62, Ch. 2], where the action taken at the t-th TS is a function of the state at the t-th TS. The policy conceived for the problem in Case 3 belongs to the stationary deterministic policy [62, Ch. 2], where the action taken at a specific state is only determined by this state, regardless of which TS it is. Furthermore, since the satisfaction of Constraint* (7) *has to be checked before carrying out the policy, the R-learning aided scheduler proposed for Case 3 has to be trained in an offline manner.*

D. Analysis of Computational Complexity

The computational complexity of the backward induction method and the R-learning method is dominated by generating the tables of the state value function $\{v^t(\boldsymbol{X})\}$ and of the action value function $R(\boldsymbol{X}, \boldsymbol{\mathcal{A}})$, respectively. Hence we focus our attention on analyzing the computational complexity of the table generation for each case as follows:

- *Solution to the Problem in Case 1:* The Bellman optimality equation (20) represents a series of equations, whose total number is determined by the number of states, $|\boldsymbol{\mathcal{X}}|$. In each equation, all $|\boldsymbol{\mathcal{A}}|$ actions have to be tried for finding the appropriate action that maximizes the value function in (20). As a result, the computational complexity is on the order of $\mathcal{O}(|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|)$ at each TS. Given that the total number of TSs is T, the overall computational complexity is $\mathcal{O}(T|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|)$.
- *Solution to the Problem in Case 2:* Similar to Case 1, the Bellman optimality equation (22) comprises $|\boldsymbol{\mathcal{X}}|$ equations and all $|\boldsymbol{\mathcal{A}}|$ actions have to be tried for each

state. Note that at the t-th TS, all $|\boldsymbol{\mathcal{X}}|$ state value functions at $(t+1)$-st have to be accessed for calculating $\sum_{\mathbf{X}'\in\boldsymbol{\mathcal{X}}} \Pr(\mathbf{X}'|\mathbf{X},A)\cdot\left[C_\gamma(\mathbf{X},\mathbf{X}',A)+v_\gamma^{t+1}(\mathbf{X}')\right]$. Hence the overall computational complexity is given by $\mathcal{O}(T|\mathbf{X}|^2|\boldsymbol{\mathcal{A}}|)$.

- *Solution to the Problem in Case 3:* At each training TS, the operation $\arg\min_{A\in\boldsymbol{\mathcal{A}}} R_\gamma^\Omega(\mathbf{X},A)$ dominates the computational complexity, which is the order of $\mathcal{O}(|\boldsymbol{\mathcal{A}}|)$. Then, upon setting the total number of training TSs to T_{train}, we may obtain the overall computational complexity as $\mathcal{O}(T_{\text{train}}|\boldsymbol{\mathcal{A}}|)$.

Remark 4. *The size of the table including all action values of the learning approach proposed for Case 3 equals $|\boldsymbol{\mathcal{X}}||\boldsymbol{\mathcal{A}}|$, which increases along with the joint queuing state information space $|\boldsymbol{\mathcal{Q}}|$, the ABS' location space $|\boldsymbol{\mathcal{U}}|$, and the action space $|\boldsymbol{\mathcal{A}}|$. It can be readily seen that both the ABS' location space $|\boldsymbol{\mathcal{U}}|$ and the action space $|\boldsymbol{\mathcal{A}}|$ are limited in the problem considered, while the joint queuing state information space $|\boldsymbol{\mathcal{Q}}|$ increases exponentially along with the number of ground devices. Three approaches can be considered for addressing this complexity issue. Firstly, by using the value-function approximation [69], the original value-function can be replaced by a value-function approximator, which may help to find a sub-optimal policy associated with a reduced complexity. The second approach is deep reinforcement learning [61], where the original action-state value function is replaced by the value function weighted by the deep neural-network having multiple layers, which is capable of handling very large state spaces. The third approach is multi-agent reinforcement learning [70]. Specifically, if dense ground devices have to be served, a single ABS may not be able to accommodate the erratically fluctuating wireless tele-traffic dynamics due to its limited buffer space and agility. Alternatively, the ground devices can be clustered into a number of groups, each of which is assigned to an ABS. Under the framework of multi-agent reinforcement learning, we may view each ABS as an agent. As a benefit, the space size of each agent is reduced.*

V. Performance Evaluation

In this section, we characterize the performance of our proposed minimum-delay dynamic ABS placement strategies by numerical results, in terms of the average delay per user and of the buffer overflow probability. Specifically, the average delay per user is given by $\sum_1^K \mathbb{E}\{\omega_k Q_k[t]\}/K$. It reflects the overall delay performance of the system. The buffer overflow probability characterizes the probability that the buffer size is incapable of storing the queue length and it plays a crucial role in devices equipped with limited buffer sizes. For comparison, we also consider two benchmark schemes, detailed as follows:

- *CSI-only scheme:* The placement of the ABS is optimized for maximizing the summation of the ground devices' throughput. Mathematically, the objective function of Problem $\mathcal{P}0$ is replaced by $-\sum_{k=1}^K R_k$. This algorithm represents the state-or-the-art schemes, where the wireless traffic dynamics are not considered when scheduling the ABS' placement.
- *MaxWeight scheme [71], [72]:* This is a classic delay-aware scheduling algorithm in wireless communications, which has hitherto not been investigated in UAV communications. We dynamically schedule the ABS placement to the specific points in harmony with the dynamic wireless traffic using this scheme, for maximizing the sum of the queue-length-weighted throughput. Mathematically, the objective function of Problem $\mathcal{P}0$ is replaced by $-\sum_{k=1}^K Q_k[t]R_k[t]$. This scheme is capable of achieving throughput-optimal performance, while maintaining the queue's stability [71].

Without loss of generality, the ground devices are located on the rectangular area bounded by its vertexes $[\underline{x},\underline{y},0]$, $[\underline{x},\overline{y},0]$, $[\overline{x},\underline{y},0]$ and $[\overline{x},\overline{y},0]$. The height of the ABS is adapted in the range of $[\underline{h},\overline{h}]$. These minimum and maximum heights have to comply with relevant regulations [23], e.g. FAA. The energy consumption of the UAV mobility is based on the model proposed in [23]. As illustrated in Section II-A, a probabilistic LoS model is considered for the link between the ground devices and the ABS, while data transmission is contaminated by additive white Gaussian noise associated with a zero mean and a power spectral density of σ^2. The default settings are specified in Table I. Under this parameter setting, the received signal-to-noise ratio at the vertex on the ground can be tuned from -6.92 dB to 24.31 dB by dynamically adjusting the placement of the ABS in this 3-D space. As for the wireless tele-traffic, we model the packet arrival process of each ground device by a two-state hidden Markov process, where State S_1 and S_2 represent the states of a low packet arrival rate and of a high packet arrival rate, respectively. The transition probability between two states is set as $p_{th}=0.1$. We assume that the packet arrival process of both states obeys the Poisson distribution [60] and the packet arrival rates of the two states are $\lambda_k^{S_1}$ and $\lambda_k^{S_2}$, respectively. The implementation of reinforcement learning is comprised of two steps, namely offline training and online policy operation. As for the offline training step, we initialize the parameter settings as $\epsilon=0.8$, $\alpha=0.1$, and $\beta=0.2$. The value of ϵ is gradually reduced during the training process. We stop the training when neither the average delay nor the overflow probability can be reduced during the performance evaluation. Note that we set $\epsilon=0$ during the performance evaluation, because our reinforcement learning algorithm is trained in an offline manner. Using the R-table obtained, we may carry out the policy by solving (27) at each TS. Let us now study the performance of the proposal in various simulation environments, compared to that of benchmark schemes.

A. Impact of the ABS' Total Energy

Fig. 2 shows the average delay per user and the buffer overflow probability versus the total battery energy of the ABS. Specific to the parameter settings, as detailed in Table I, the total power required for remaining airborne is $P_c+P_h+P_d=180$ W. Hence, given that the total service time is 30 min, the ABS cannot be moved when its total battery charge is 90 Wh. By contrast, it can be inferred that the ABS may be scheduled for travel in any direction at each time slot, when

Table I: Default simulation parameter settings

Description	Parameter and Value
Bandwidth	$B = 500$ KHz
Scheduling slot	$\tau = 2$ s
Carrier frequency	$f_c = 2.4$ GHz
Path-loss exponent [27]	2
LoS probability setting	$\beta = 0.14$ $\psi = 11.95$
Attenuation factor [27]	$\mu_{\text{LoS}} = 3$ dB $\mu_{\text{NLoS}} = 23$ dB
Noise	$\sigma^2 = -170$ dBm/Hz
Power consumption [23]	$P_t = 0.1$ mW $P_c = 5$ W $P_m^h = 10$ W $P_m^{v+} = 20$ W $P_m^{v-} = 15$ W $P_h = 170$ W $P_d = 5$ W
Buffer size	$N_Q = 5$ packet
Packet size [49]	$N_k = 290$ Kbyte/packet
Service time	$T\tau = 30$ min
UAV moving speed	$v_h = 20$ m/s $v_v = 5$ m/s
UAV altitude The area of ground devices	$[\underline{h}, \overline{h}] = [60, 80]$ m $\underline{x} = \underline{y} = 0$, $\overline{x} = \overline{y} = 160$ m
Weight factor	$w_1 = w_2 = 1$

its battery life is 100 Wh. Our observations are as follows. Firstly, the performance of the CSI-only scheme does not change upon increasing the ABS' total energy. This is because the maximum-throughput placement pursued by the CSI-only scheme is static, once the locations of the ground devices are fixed. Secondly, upon increasing the ABS' total energy, both a lower average delay and a lower overflow probability are achieved by using the MaxWeight scheme and using our proposed algorithms for the three cases. This implies that these queue-aware dynamic ABS placement scheduling schemes are indeed capable of reducing both the system delay and the overflow probability, when the battery energy is sufficient for the ABS' movement. Thirdly, equipped with sufficient battery energy for movement, the delay is the lowest for the backward induction aided scheme in Case 1, followed by the backward induction aided scheme in Case 2, the reinforcement learning aided scheme in Case 3 and the MaxWeight scheme. As illustrated in [45], the MaxWeight scheme aims for achieving the maximum throughput, while maintaining a stable queue, whose delay is higher than that of the minimum-delay schemes conceived for Case 1, 2 and 3. As for the order in Case 1, 2, and 3, this is due to their different apriori knowledge of the wireless tele-traffic dynamics. Specifically, the exact number of arriving packets is known in Case 1, and the probability mass function of the arrival packets is known in Case 2, while in Case 3 the wireless tele-traffic dynamics have to be learned during the training process. Fourthly, the increase of the ABS' total energy drastically reduces both the average delay and the overflow probability in the queue-aware ABS placement scheduling schemes, when the ABS' total energy is below a certain threshold, say 94 Wh, while the reduction becomes much smaller afterwards. It can be inferred that the minimum delay can be achieved without adjusting the ABS' placement for every TS.

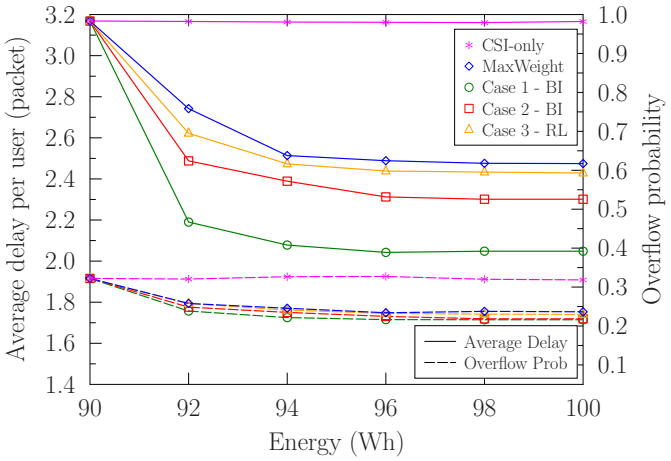

Figure 2: Simulation results of the average delay per user and of the buffer overflow probability versus the total battery energy of the ABS in a two-device system, where the devices' locations are $(0, 80, 0)$ m and $(160, 80, 0)$ m. The packet arrive rate of two states are $\lambda_k^{S_1} = 0.2$ pck/τ and $\lambda_k^{S_2} = 3.0$ pck/τ, respectively.

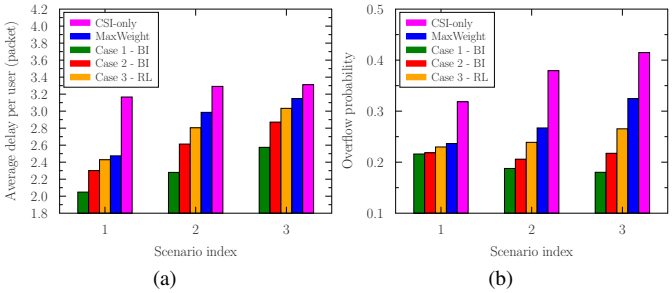

Figure 3: Simulation results of the average delay per user (a) and of the buffer overflow probability (b) in various wireless traffic scenarios in a two-device system, where the devices' locations are $(0, 80, 0)$ m and $(160, 80, 0)$ m.. The total battery energy of the ABS is 100 Wh. The wireless traffic scenarios are specified as follows. Scenario 1: $\lambda_k^{S_1} = 0.2$ pck/τ, $\lambda_k^{S_2} = 3.0$ pck/τ; Scenario 2: $\lambda_k^{S_1} = 0.8$ pck/τ, $\lambda_k^{S_2} = 2.4$ pck/τ; Scenario 3: $\lambda_k^{S_1} = 1.6$ pck/τ, $\lambda_k^{S_2} = 1.6$ pck/τ.

B. Impact of the Asymmetry Wireless Tele-Traffic

Fig. 3 presents both the average delay and the buffer overflow probability of a two-device system, where various packet arrival rates are set for the two traffic states in three different scenarios. The expectation values of the packet arrival rates in these three scenarios remain the same. Having a higher difference between the values of $\lambda_k^{S_1}$ and $\lambda_k^{S_2}$ implies a more asymmetric packet arrival process in the simulations. We have the following observations. Firstly, as for the average delay, the advantage of the queue-aware dynamic ABS placement schemes over the CSI-only scheme becomes higher upon increasing the difference between the values of $\lambda_k^{S_1}$ and $\lambda_k^{S_2}$. This is because a higher difference between the values of $\lambda_k^{S_1}$ and $\lambda_k^{S_2}$ implies having more substantially fluctuating wireless tele-traffic dynamics, while the queue-aware schemes are capable of tracking these dynamic fluctuations. Secondly, although the average delay performance increases upon reducing the difference between $\lambda_k^{S_1}$ and $\lambda_k^{S_2}$ for both our proposed algorithms and for the benchmark schemes,

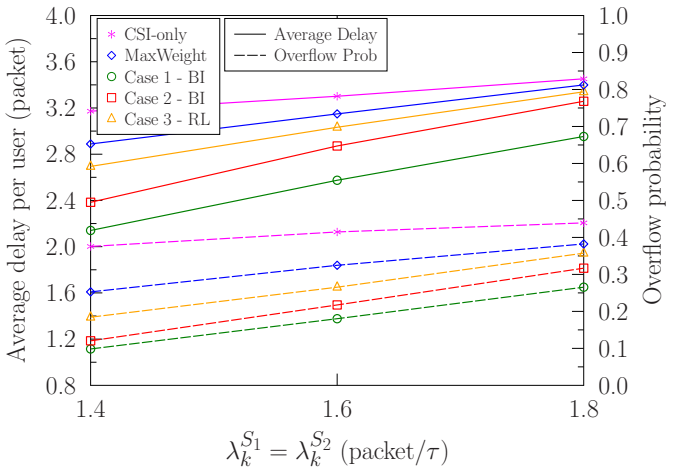

Figure 4: Simulation results of the average delay per user and of the buffer overflow probability versus the packet arrival rate in a two-device system, where the devices' locations are $(0, 80, 0)$ m and $(160, 80, 0)$ m. The total battery energy of the ABS is 100 Wh.

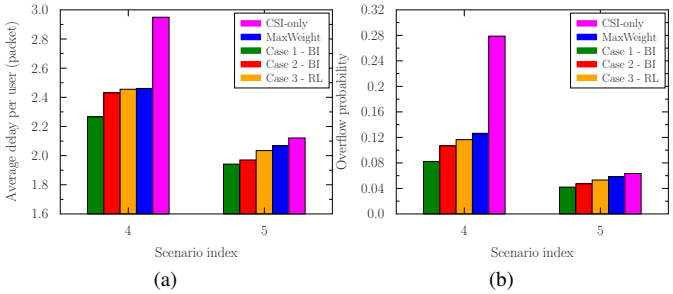

Figure 5: Simulation results of the average delay per user (a) and of the buffer overflow probability (b) for a three-device system under two different devices' locations settings. The locations of three devices are specified as follows. Scenario 4: Device 1 $(3, 76, 0)$ m, Device 2 $(4, 7, 0)$ m, and Device 3 $(38, 75, 0)$ m; Scenario 5: Device 1 $(3, 76, 0)$ m, Device 2 $(4, 47, 0)$ m, and Device 3 $(38, 75, 0)$ m. The total battery energy of the ABS is 100 Wh. The packet arrival rate is set as $\lambda_k^{S_1} = \lambda_k^{S_2} = 1.6$ pck/τ.

the overflow probability of our proposed algorithms remains almost the same, which demonstrates the efficiency of our proposed algorithms.

C. Impact of the Wireless Tele-Traffic Rate

Fig. 4 plots the average delay per user and the buffer overflow probability versus the packet arrival rates in a two-device system. Our observations are as follows. Firstly, as expected, both the average delay per user and the buffer overflow probability increase upon increasing the packet arrival rate. Specific to the average delay per user, with reference to (4), the value ranges from the mean value of the packet arrival process λ and N_Q. Secondly, the advantage of our proposed algorithms over the CSI-only and MaxWeight schemes becomes lower, upon increasing the packet arrival rate. This is because the average delay is saturated by N_Q, when the packet arrival rate is high.

D. Impact of the Ground Devices' Location

Fig. 5 illustrates both the average delay and the buffer overflow probability in two different device location settings for a three-device system. Specifically, the distance among the devices in Scenario 4 is higher than that in Scenario 5. It can be observed that both the average delay and the buffer overflow probability can be significantly reduced, if the locations of the ground devices are closer, because in this case the transmission throughput of ground devices may be beneficially adjusted by adapting the ABS placement in a single TS. This provides an important insight for engineering design. For a system where a large number of ground devices have to be served, we may cluster the devices based on their distance and assign an ABS for each device cluster for attaining a reduced delay. Multi-ABS systems can be realized by appropriately designing both the user association and the resource management [27], [54] with the aid of a multi-agent MDP framework [55], [56], which is beyond our current scope.

VI. Conclusions

A beneficial architecture has been proposed for a UAV-aided network from a delay-minimization perspective. We have formulated a minimum-delay ABS placement problem, subject to practical constraints imposed on the ABS' battery life and velocity. We then transformed the primal problem to the corresponding CMDP problem, and provided solutions to the problems formulated under various assumptions concerning our knowledge about wireless tele-traffic. The numerical results demonstrated that our proposed solutions are capable of reducing the delay compared to the benchmark scheme under the various scenarios considered.

As for future work, the mobility of the ground users will also be addressed, for rendering our cross-layer optimization framework applicable to general cellular networks. Furthermore, the increasing number of ground users imposes a higher complexity on the learning-aided approach, which will be tackled with the aid of value-function approximation, of deep reinforcement learning, and of multi-agent reinforcement learning.

Appendix A
The Proof of Theorem 1

The problem formulated in Section III-A is readily observed to be a CMDP problem associated with the average expected cost. The feasibility of applying the Lagrangian approach to this type of problems has been richly documented in [62, Ch. 12]. Here we aim for proving that the immediate costs, i.e. $\sum_{k=1}^{K}\left\{\frac{w_k Q_k[t]}{\lambda_k}\right\}$ and $P_m\left(\frac{|u_x[t]-u_x[t-1]|}{\delta}+\frac{|u_y[t]-u_y[t-1]|}{\delta}\right)$, are *bounded from below* and the average expected cost of the objective satisfies the so-called *grow condition* of [62, Ch. 12], which are the prerequisites of Theorem 1. Specifically, since w_k, Q_k and λ_k are all non-negative, we have $\frac{w_k Q_k}{\lambda_k} \geq 0$ and hence $\sum_{k=1}^{K}\left\{\frac{w_k Q_k[t]}{\lambda_k}\right\} \geq 0$. Furthermore, given an action specified in Section III-A, $\left(\frac{|u_x[t]-u_x[t-1]|}{\delta}+\frac{|u_y[t]-u_y[t-1]|}{\delta}\right)$ equals either 0 or 1. Bearing in mind that P_m is positive, we have $P_m\left(\frac{|u_x[t]-u_x[t-1]|}{\delta}+\frac{|u_y[t]-u_y[t-1]|}{\delta}\right) \geq 0$. In this case, we have proved that the immediate costs related both to the objective and to the constraint are lower bounded by 0. In

terms of the so-called growth condition [62, Ch. 12], since the state space $\boldsymbol{X}[t] = \{\boldsymbol{U}[t], \boldsymbol{Q}[t]\}$ is finite, we have

$$\text{the set } \left\{ \boldsymbol{X}[t] \in \boldsymbol{\mathcal{X}} : \inf_{A} \left\{ \frac{w_k Q_k[t+1]}{\lambda_k} \right\} < \ell \right\} \text{ is finite,} \tag{28}$$

$\forall \ell \in \mathbb{R}$. This is a sufficient condition for the so-called growth condition [62, Ch. 12]. Hence, the two prerequisite conditions have been proved to be true and the proof is complete.

REFERENCES

[1] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, pp. 2241–2263, April 2019.

[2] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, pp. 36–42, May 2016.

[3] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV-based IoT platform: A crowd surveillance use case," *IEEE Commun. Mag.*, vol. 55, pp. 128–134, Feb. 2017.

[4] D. Ebrahimi, S. Sharafeddine, P. Ho, and C. Assi, "UAV-aided projection-based compressive data gathering in wireless sensor networks," *IEEE Internet Things J.*, vol. 6, pp. 1893 – 1905, April 2019.

[5] S. Jeong, O. Simeone, and J. Kang, "Mobile edge computing via a UAV-mounted cloudlet: Optimization of bit allocation and path planning," *IEEE Trans. Veh. Technol.*, vol. 67, pp. 2049–2063, Mar. 2018.

[6] L. Li, Y. Xu, Z. Zhang, J. Yin, W. Chen, and Z. Han, "A prediction-based charging policy and interference mitigation approach in the wireless powered Internet of Things," *IEEE J. Sel. Areas Commun.*, vol. 37, pp. 439–451, Feb. 2019.

[7] Q. Wu, L. Liu, and R. Zhang, "Fundamental trade-offs in communication and trajectory design for UAV-enabled wireless network," *IEEE Wireless Commun.*, vol. 26, pp. 36–44, Jan. 2019.

[8] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surv. Tutor.*, vol. 21, pp. 3417 – 3442, Fourthquarter 2019.

[9] N. Zhao, W. Lu, M. Sheng, Y. Chen, J. Tang, F. R. Yu, and K.-K. Wong, "UAV-assisted emergency networks in disasters," *IEEE Wireless Commun.*, vol. 26, pp. 45–51, Jan. 2019.

[10] L. Wang, Y. L. Che, J. Long, L. Duan, and K. Wu, "Multiple access mmwave design for UAV-aided 5G communications," *IEEE Wireless Commun.*, vol. 26, pp. 64–71, Feb. 2019.

[11] A. A. Khuwaja, Y. Chen, N. Zhao, M.-S. Alouini, and P. Dobbins, "A survey of channel modeling for UAV communications," *IEEE Commun. Surv. Tutor.*, vol. 20, pp. 2804–2821, Apr. 2018.

[12] A. Al-Hourani and K. Gomez, "Modeling cellular-to-UAV path-loss for suburban environments," *IEEE Wireless Commun. Lett.*, vol. 7, pp. 82–85, Jan. 2018.

[13] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, pp. 569–572, June 2014.

[14] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, "Modeling air-to-ground path loss for low altitude platforms in urban environments," in *Proc. 2014 IEEE Global Commun. Conf.*, pp. 2898–2904, IEEE, 2014.

[15] S. Chandrasekharan, K. Gomez, A. Al-Hourani, S. Kandeepan, T. Rasheed, L. Goratti, L. Reynaud, D. Grace, I. Bucaille, T. Wirth, *et al.*, "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, 2016.

[16] H. Kang, J. Joung, J. Ahn, and J. Kang, "Secrecy-aware altitude optimization for quasi-static UAV base station without eavesdropper location information," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 851–854, 2019.

[17] H. Ren, C. Pan, K. Wang, Y. Deng, M. Elkashlan, and A. Nallanathan, "Achievable data rate for URLLC-enabled UAV systems with 3-D channel model," *IEEE Wireless Commun. Lett.*, vol. 8, pp. 1587 – 1590, Dec. 2019.

[18] N. Goddemeier and C. Wietfeld, "Investigation of air-to-air channel characteristics and a UAV specific extension to the Rice model," in *Proc. 2015 IEEE Globecom Workshops*, pp. 1–5, IEEE, 2015.

[19] X. Ye, X. Cai, X. Yin, J. Rodríguez-Piñeiro, L. Tian, and J. Dou, "Air-to-ground big-data-assisted channel modeling based on passive sounding in LTE networks," in *Proc. 2017 IEEE Globecom Workshops*, pp. 1–6, IEEE, 2017.

[20] X. Cai, A. Gonzalez-Plaza, D. Alonso, L. Zhang, C. B. Rodríguez, A. P. Yuste, and X. Yin, "Low altitude UAV propagation channel modelling," in *Proc. 2017 European Conf. on Antennas Propagation*, pp. 1443–1447, IEEE, 2017.

[21] W. Khawaja, I. Guvenc, and D. Matolak, "Uwb channel sounding and modeling for uav air-to-ground propagation channels," in *Proc. 2016 IEEE Global Commun. Conf.*, pp. 1–7, IEEE, 2016.

[22] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 3747–3760, June 2017.

[23] Y. Zeng, J. Xu, and R. Zhang, "Energy minimization for wireless communication with rotary-wing UAV," *IEEE Trans. Wireless Commun.*, vol. 18, pp. 2329–2345, April 2019.

[24] P. Chandhar, D. Danev, and E. G. Larsson, "Massive MIMO for communications with drone swarms," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 1604–1629, Mar. 2018.

[25] N. Rupasinghe, Y. Yapıcı, I. Güvenç, and Y. Kakishima, "Non-orthogonal multiple access for mm-wave drone networks with limited feedback," *IEEE Trans. Commun.*, vol. 67, pp. 762–777, Jan. 2019.

[26] C. Xu, J. Zhang, T. Bai, P. Botsinis, R. G. Maunder, R. Zhang, and L. Hanzo, "Adaptive coherent/non-coherent single/multiple-antenna aided channel coded ground-to-air aeronautical communication," *IEEE Trans. Commun.*, vol. 67, pp. 1099–1116, Feb. 2019.

[27] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications," *IEEE Trans. Wireless Commun.*, vol. 16, pp. 7574–7589, Nov. 2017.

[28] T. Hou, Y. Liu, Z. Song, X. Sun, and Y. Chen, "Multiple antenna aided NOMA in UAV networks: A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 67, pp. 1031–1044, Feb. 2019.

[29] T. Bai, J. Wang, Y. Ren, and L. Hanzo, "Energy-efficient computation offloading for secure UAV-edge-computing systems," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 6074–6087, June 2019.

[30] 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 Trans. Commun.*, vol. 67, pp. 3723–3735, May 2019.

[31] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, "Joint UAV hovering altitude and power control for space-air-ground IoT networks," *IEEE Internet Things J.*, vol. 6, pp. 1741–1753, April 2019.

[32] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Commun. Lett.*, vol. 6, pp. 434–437, Apr. 2017.

[33] H. Kim, J. Ben-Othman, and L. Mokdad, "UDiPP: A framework for differential privacy preserving movements of unmanned aerial vehicles in smart cities," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 3933–3943, April 2019.

[34] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," *IEEE Commun. Lett.*, 2019.

[35] L. Liu, S. Zhang, and R. Zhang, "CoMP in the sky: UAV placement and movement optimization for multi-user communications," *IEEE Trans. Commun.*, pp. 1–1, 2019.

[36] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 2109–2121, Mar. 2018.

[37] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, pp. 4983–4996, Dec. 2016.

[38] J. Gong, T.-H. Chang, C. Shen, and X. Chen, "Flight time minimization of UAV for data collection over wireless sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 36, pp. 1942–1954, Sep. 2018.

[39] C. E. Shannon, "Communication in the presence of noise," *Proc. IEEE*, vol. 86, pp. 447–457, Feb. 1998.

[40] J. Li and Y. Han, "Optimal resource allocation for packet delay minimization in multi-layer UAV networks," *IEEE Commun. Lett.*, vol. 21, pp. 580–583, March 2016.

[41] Z. M. Fadlullah, D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "A dynamic trajectory control algorithm for improving the communication

throughput and delay in UAV-aided networks," *IEEE Netw.*, vol. 30, pp. 100–105, Jan. 2016.

[42] J. Zhang, L. Zhou, Q. Tang, E. C. . Ngai, X. Hu, H. Zhao, and J. Wei, "Stochastic computation offloading and trajectory scheduling for UAV-assisted mobile edge computing," *IEEE Internet Things J.*, vol. 6, pp. 3688–3699, April 2019.

[43] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, pp. 1–211, Jan. 2010.

[44] A. Asheralieva and D. Niyato, "Game theory and lyapunov optimization for cloud-based content delivery networks with device-to-device and UAV-enabled caching," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 10094 – 10110, Oct. 2019.

[45] Y. Cui, V. K. Lau, R. Wang, H. Huang, and S. Zhang, "A survey on delay-aware resource control for wireless systems—large deviation theory, stochastic Lyapunov drift, and distributed stochastic learning," *IEEE Trans. Inf. Theory*, vol. 58, pp. 1677–1701, Mar. 2012.

[46] R. A. Howard, *Dynamic programming and Markov processes.* John Wiley, 1960.

[47] X. Chen, W. Chen, J. Lee, and N. B. Shroff, "Delay-optimal buffer-aware scheduling with adaptive transmission," *IEEE Trans. Commun.*, vol. 65, pp. 2917–2930, July 2017.

[48] M. Wang, J. Liu, W. Chen, and A. Ephremides, "Joint queue-aware and channel-aware delay optimal scheduling of arbitrarily bursty traffic over multi-state time-varying channels," *IEEE Trans. Commun.*, vol. 67, pp. 503–517, Jan. 2019.

[49] V. K. Lau and Y. Cui, "Delay-optimal power and subcarrier allocation for OFDMA systems via stochastic approximation," *IEEE Trans. Wireless Commun.*, vol. 9, pp. 227–233, Jan. 2010.

[50] L. Lei, Y. Kuang, N. Cheng, X. S. Shen, Z. Zhong, and C. Lin, "Delay-optimal dynamic mode selection and resource allocation in device-to-device communications—Part I: Optimal policy," *IEEE Trans. Veh. Technol.*, vol. 65, pp. 3474–3490, May 2016.

[51] R. Zhang, Y. Cui, H. Claussen, H. Haas, and L. Hanzo, "Anticipatory association for indoor visible light communications: Light, follow me!," *IEEE Trans. Wireless Commun.*, vol. 17, pp. 2499–2510, Apr. 2018.

[52] J. Du, C. Jiang, Y. Qian, Z. Han, and Y. Ren, "Traffic prediction based resource configuration in space-based systems," in *Proc. 2016 IEEE Int. Conf. Commun.*, (Kuala Lumpur, Malaysia), pp. 1–6, May 2016.

[53] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Commun. Mag.*, vol. 32, no. 3, pp. 70–81, 1994.

[54] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 8036–8049, Aug 2019.

[55] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. 17th Int. Conf. Machine Learning*, Citeseer, 2000.

[56] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2019.

[57] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, pp. 66–73, May 2016.

[58] D. J. Costello and G. D. Forney, "Channel coding: The road to channel capacity," *Proc. IEEE*, vol. 95, pp. 1150–1177, June 2007.

[59] I. Bettesh and S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Trans. Inf. Theory*, vol. 52, pp. 4115–4141, Sep. 2006.

[60] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. McGraw-Hill Education, 2002.

[61] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, vol. 135. MIT Press Cambridge, 1998.

[62] E. Altman, *Constrained Markov Decision Processes*, vol. 7. CRC Press, 1999.

[63] V. S. Borkar, "Convex analytic methods in Markov decision processes," in *Handbook of Markov Decision Processes*, pp. 347–375, Springer, 2002.

[64] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, pp. 98–105, Feb. 2016.

[65] N. Bui, M. Cesana, S. A. Hosseini, Q. Liao, I. Malanchini, and J. Widmer, "A survey of anticipatory mobile networking: Context-based classification, prediction methodologies, and optimization techniques," *IEEE Commun. Surv. Tutor.*, vol. 19, pp. 1790–1821, Mar. 2017.

[66] R. J. Aumann, "Backward induction and common knowledge of rationality," *Games and Economic Behavior*, vol. 8, no. 1, pp. 6–19, 1995.

[67] A. Schwartz, "A reinforcement learning method for maximizing undiscounted rewards," in *Proc. 10th Int. Conf. Machine Learning*, vol. 298, pp. 298–305, 1993.

[68] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, pp. 1872–1882, Apr. 2013.

[69] G. Konidaris, S. Osentoski, and P. S. Thomas, "Value function approximation in reinforcement learning using the fourier basis," in *Proc. 25th AAAI Conf. Artif. Intell.*, vol. 6, Aug. 2011.

[70] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for uav networks," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, 2020.

[71] L. Georgiadis, M. J. Neely, L. Tassiulas, *et al.*, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1–144, 2006.

[72] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, pp. 150–154, Feb. 2001.