UAV-Assisted Wireless Energy and Data Transfer With Deep Reinforcement Learning

Zehui Xiong, *Member, IEEE*, Yang Zhang, *Member, IEEE*, Wei Yang Bryan Lim, Jiawen Kang, Dusit Niyato, *Fellow, IEEE*, Cyril Leung, *Life Member, IEEE*, and Chunyan Miao, *Senior Member, IEEE*

***Abstract*—As a typical scenario in future generation communication network applications, UAV-assisted communication can perform autonomous data delivery for massive machine type communication (mMTC), where the data generated from Internet of Things (IoT) devices can be carried and delivered to the corresponding locations with no direct communication channels to the IoT devices. Wireless energy transfer technique can recharge the UAV when the system is in operation, assisting the UAV to continuously collect and deliver data. In this work, we formulate a Markov decision process (MDP) model to describe the energy and data transfer optimization problem for the UAV. To maximize the long-term utility of the UAV, the MDP model is solved by value iteration algorithm to obtain the optimal strategies of the UAV to collect data, deliver data, and receive transferred energy to replenish on-device battery energy storage. Furthermore, to tackle the issues of system state uncertainties, partially observable states, and large state space in UAV-assisted communication systems, we extend the MDP model and solve it by using a *Q*-learning and a deep reinforcement learning (DRL) schemes. Simulations and numerical results validate that, compared with baseline schemes, the proposed MDP model with DRL based scheme can achieve better wireless energy and data transfer strategies in terms of the higher long-term utility of the UAV.**

***Index Terms*—Unmanned aerial vehicle, wireless energy transfer, Internet of Things, Markov decision process, deep reinforcement learning.**

Manuscript received April 27, 2020; revised August 11, 2020; accepted September 21, 2020. Date of publication September 29, 2020; date of current version March 8, 2021. This research is supported in part by National Natural Science Foundation of China (Grant No. 62071343), the National Research Foundation (NRF), Singapore, under Singapore Energy Market Authority (EMA), Energy Resilience, NRF2017EWT-EP003-041, Singapore NRF2015-NRF-ISF001-2277, Singapore NRF National Satellite of Excellence, Design Science and Technology for Secure Critical Infrastructure NSoE DeST-SCI2019-0007, A*STAR-NTU-SUTD Joint Research Grant on Artificial Intelligence for the Future of Manufacturing RGANS1906, Wallenberg AI, Autonomous Systems and Software Program and Nanyang Technological University (WASP/NTU) under grant M4082187 (4080), Singapore Ministry of Education (MOE) Tier 1 (RG16/20), Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), and the Open Research Project of the State Key Laboratory of Industrial Control Technology, Zhejiang University, China (No. ICT20044). The work of Zehui Xiong is supported by Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI), Nanyang Technological University, Singapore. The associate editor coordinating the review of this article and approving it for publication was C. Jiang. *(Corresponding author: Yang Zhang.)*

Zehui Xiong is with the Alibaba-NTU Joint Research Institute and School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Yang Zhang is with the Hubei Key Laboratory of Transportation Internet of Things, School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430070, China (e-mail: yangzhang@whut.edu.cn).

Wei Yang Bryan Lim is with the Alibaba Group and Alibaba-NTU Joint Research Institute, Nanyang Technological University, Singapore.

Jiawen Kang and Dusit Niyato are with the School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Cyril Leung is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada, and also with the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore.

Chunyan Miao is with the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY) and School of Computer Science and Engineering, Nanyang Technological University, Singapore.

Digital Object Identifier 10.1109/TCCN.2020.3027696

I. Introduction

INNOVATIVE techniques will be required to significantly enhance coverage and connectivity in future generation communication networks, such as 5G, 6G and beyond. One of the efforts is to enable massive amount of heterogeneous Internet of Things (IoT) devices to communicate and share information from various locations. To improve communication efficiency, unmanned aerial vehicles (UAVs) have been deployed and applied as a promising approach to achieve physically extended coverage and long-distance communication capabilities [1], especially in the situations where there is no overhaul on the existing architectures and components of communication systems.

A typical UAV application in future generation communication networks is UAV-assisted network data delivery for massive machine type communication (mMTC). The mMTC scenario requires autonomous and continuous data delivery from IoT devices of different locations to their corresponding destinations, e.g., base stations and data sinks. However, existing mMTC communication techniques suffer from low data throughput and low data transmission frequency. Moreover, in a practical large-scale system with enormous network participants (e.g., over 8,000 access points are deployed in merely a WiFi system to cover all the network users in an experimental study in [2]), dynamic network architecture design (e.g., cell design and base station deployment) and network resource allocation (e.g., energy, data, and spectrum management) are required to optimize network performance [3]. Comparing with conventional fixed base stations, UAV-assisted communication has several advantages. For example, as UAVs can approach and access the IoT devices within a close geographical range from the sky, UAV-assisted communication can provide line-of-sight (LoS) connections between the UAV and IoT devices [4]. Moreover, with the features of high mobility and data storage capacity, UAVs can collect and move data from IoT devices to their corresponding destinations in an

offline fashion [5], [6], which enables relatively long distance data transmission without implementing extra communication infrastructures. In this regard, UAV-assisted communication can improve the efficiency and reliability of IoT data collecting by alternatively providing an aerial-based physical link.

Simultaneous wireless information and power transfer (SWIPT) [7] is considered for UAVs to switch between energy and data transfer decisions using the same communication channel. By carrying a SWIPT device on the UAV, the UAV can charge other IoT devices and be charged with external wireless energy sources. For example, the UAV can receive energy from ambient energy sources as well as BSs to replenish energy storage. Moreover, as a method for collecting data, the UAV can also activate or compensate IoT devices in the environment to encourage the devices to sense and transfer data to the UAV. We therefore consider that the UAV-assisted data delivery system supported by wireless energy charging can be a promising energy management method. At the IoT device side, a UAV can charge the IoT devices to collect data. Energy consumption of an IoT device can be compensated by wireless energy charged from the UAV, which potentially extends the device battery life. The UAV then carry and deliver the data, which consumes energy storage for communication on the UAV. When the UAV reaches a location with energy source, e.g., a base station or a charging station, the UAV can be also charged. By strategically managing the wireless energy charging and data delivery actions, the UAV can perpetually operates in the network with minimal human intervention, thereby providing efficient services to serve machine type communications.

Drawbacks still exist when applying UAVs-assisted IoT data collecting. Firstly, future generation communication and vehicular systems, i.e., 5G and beyond, requires certain metrics such as delay and quality of service/experience [8], [9]. However, UAVs require some time to reach the data destinations, e.g., base stations. Each UAV is required to strategically make data delivery decisions upon reaching the data destinations. Otherwise, severe delay in data delivery may be caused, especially for IoTs with no other connection channels to the corresponding data destinations. Secondly, the UAVs behaviours, such as cruising, hovering and wireless communication, are solely powered by on board batteries. To avoid losing control or forced landing because of battery depletion, only a limited portion of battery energy will be allocated for data communication. Once energy outage for data transmission happens, delay or data transmission at an extremely high cost can be incurred. To this end, the energy management and charging approaches for controlling the energy usage and supply are of vital importance for persistent operations of UAV-assisted data delivery for IoTs in future communication networks.

In this work, we employ a Markov decision process (MDP) [10] to model the dynamics of a UAV with wireless energy transfer to deliver data in an IoT system. By solving the MDP with the methods of value iteration and Q-learning, the UAV manages to optimize the energy charging and data delivery strategies to maximize the revenue obtained from data delivery, while taking into account of the energy efficiency and the data delivery delay. Moreover, we propose a deep reinforcement learning (DRL) [11] approach for the UAV to determine the optimized data delivery and wireless energy transfer strategies in the case of unknown system states and large system state space in practical real-world applications and services.

The contributions of this work are summarized as follows:

- We propose a UAV-assisted data collecting and delivery system in IoT systems, where the UAV is equipped with wireless energy transfer functionalities to charge IoT devices for collecting data, as well as receiving energy from base stations to charge the UAV battery.
- We model the UAV data delivery and energy management as a Markov decision process (MDP) problem, which captures the dynamics and stochastic behaviours of the UAV in the system. By solving the optimization problem, the long-term expected reward received by the UAV is maximized.
- We develop a Q-learning and a DRL based schemes to determine the optimal energy transfer and data collecting strategies under the conditions of state uncertainties, partially observable states, and large state space.
- Through extensive numerical simulations, the proposed MDP-based optimization scheme is validated and shown to outperform the baseline schemes in terms of achieving higher long-term expected utility of the UAV. Simulation results also confirm that the proposed Q-learning and DRL based optimization schemes yield the same result as conventional value iteration algorithm to solve the MDP model.

II. Related Work

A. UAV as a Communication Network Component

Unmanned aerial vehicles (UAVs) communications and networking recently have been increasingly attracting attention from researchers and engineers due to their wide coverage and high agility in various promising applications. By integrating UAVs into existing terrestrial infrastructure, many of the wireless communications performance gains can be achieved and even enhanced without being restricted by geographical constraints [1]. For example, UAVs can be deployed as flying "base stations" to improve the wireless coverage [12]. The authors in [13], [14] studied the placement optimization for UAVs to achieve higher coverage. UAVs can also be treated as flying "relays" to support cooperative communication when the direct communication link is severely blocked [15]. In [16], the authors designed an energy-efficient scheme to optimize the UAV's trajectory for data collection, where the objective is to maximize the communication throughput while minimizing the energy consumption of UAV. The authors in [17] studied a similar trajectory design problem but the focus is to minimize the UAV's mission completion time subject to the energy constraints. The authors in [18] proposed a joint trajectory and power control design to maximize the minimum throughput by optimizing the user scheduling and association.

B. Wireless Energy Charging and Energy Management for UAVs

UAVs can act as energy sources to charge the wireless nodes in certain sensing applications where wire charging is

impractical [19]. In [20], the authors considered that UAVs employ the radio frequency wireless power transfer to charge the users, and developed the optimal transmission resource allocation given with the energy constraints of users. In [21], a UAV is implemented with a mounted wireless energy transmitter is dispatched to transfer energy to replenish energy receivers at particular ground locations. Trajectory strategy of the UAV is optimized for maximizing the total amount of energy charged to all the energy receivers within a certain time period. Joint management of UAV energy and spectrum access has been proposed in [22], where a UAV can hover and make the optimal decision to decide whether to perform RF wireless energy transfer (WET) to charge the corresponding ground terminal, or to access primary spectrum for wireless information transfer (WIT) to serve the ground terminal. A bipartite matching model has been modeled in [23] to describe and optimize the multi-stage energy charging process between UAVs and devices receiving wireless energy from chargers carried by the UAVs.

Despite the various advantages brought by the introduction of UAVs, the UAV's operations are often restricted by their limited battery capacity and power to support the necessary mobility and communication. As such, UAVs need to return to its base, i.e., charging stations for recharging [24]. In [25], the authors developed a scheduler policy by scheduling energy replenishment operations and leveraging the charging stations. Both centralized and distributed deployment strategies of UAVs to meet the requirements of UAV coverage to target locations, as well as fast return to ground-based UAV charging stations. The authors in [26] considered a scenario where multiple UAVs that are remotely powered by a wireless charging station on the ground. In such a mobile network with wireless-powered UAVs, the optimal resource allocation scheme is investigated for joint scheduling between wireless charging and downlink transmission. The authors in [27] argued that UAVs have to make the decisions of serving the users along the designed trajectory or to return to the charging stations based on a real-time observation accordingly and autonomously. The authors in [28] proposed an optimal charge scheduling algorithm among UAVs, where the game-theoretic approach to model the energy trading interactions between the UAVs and charging station in a cost-effective manner. A market-oriented framework for reliable energy trading has been proposed in [29], where the UAV charging process is not free. UAVs need to buy energy from charging stations using blockchain tokens. A UAV has to borrow tokens from the charging station if the UAV does not have sufficient token at the moment of charging. The UAV pays back the borrowed token as well as interest later. By the means of monetary transactions, the UAVs rationally choose their energy trading actions by employing a game-theoretic approach. As such, energy efficiency and sustainability are achieved.

C. System Optimization, Markov Decision Process, and Reinforcement Learning Approaches

Markov process based models for optimal energy and data management have been discussed and studied recently. Studies in [30], [31] have found that vehicular communication users are able to sense information from environment and adjacent users and optimize the system performance based on the sensed information. Beaconing signals are regularly broadcasted by communication users. The beaconing signal transmission behaviours of users are modeled as a two-state Markov chain [30]. As investigated in [32], for large-scale complex environments, it is essential for UAVs to make online actions to execute their tasks.

Simple solution methods for MDP approaches, e.g., value iteration and policy iteration, requires complete information to be transmitted to a decision maker for optimal decision making. As argued in [33], UAV systems with mobility feature and distributed nature are usually trust prone. it is also discussed in [34] that optimizing resource allocation strategies in distributed network and communication systems relies on information exchange to improve system and individual performance, which is not always feasible as generated and sensed data by network participants can be ususally considered as private and incomplete information. As comprehensively reviewed in [35]–[37], machine learning and deep learning approaches are already proved to be efficient in analyzing and exploiting complete or limited network information and system dynamics to improve the performance of resource management in future network and communication systems (including UAV assisted systems).

For example, A three-layered complex satallite communication network with heterogeneous satallite communication users is discussed in [38], where Q-learning is applied as an effective and low-complexity method to optimize the long-term utility of the system. A liquid state machine method is employed in [39] for cloud servers (i.e., data sources) to predict future content requests made by data users with limited information. In [40], UAVs operate as mobile edge computing servers for ground mobile users, where the task offloading problem is formulated as a semi-Markov decision process (SMDP) without determined state transition probabilities (i.e., model-free). A DRL scheme has been proposed to maximize the user throughput. Similarly, a practical model-free DRL scheme is proposed in [41] for UAV-assisted sensing data collection in smart city applications. In [42], a UAV plans its trajectory according to the situations of ground terminal users, who offload computational tasks to the UAV. A MDP model with a double deep Q-network (DDQN) schemes have been developed to optimize trajectory and UAV-user association parameters under the quality of service constraint. By employing DDQN, the UAV system can achieve more accurate and intelligent trajectory planning strategies [43]. A UAV-based security attack strategy is studied in [44] to jam the ground users in the system. The UAV jammer and the users act as leader and follower players of a Stackelberg game, respectively, during the jam and anti-jam interactions. As the UAV has limited knowledge of user actions, a partially observable MDP (POMDP) problem is formulated to model the UAV behaviours to handle the missing observations [23], which is then solved by using a deep recurrent Q-network (DRQN) in the three dimension space.

To the best of our knowledge, there is no existing works in the literature proposing the deep reinforcement learning

approach for wireless energy and data transfer in UAV-assisted communication systems.

III. System Description

As shown in Fig. 1, we consider a UAV-assisted data delivery system consisting of a single UAV equipped with energy storage and data storage facilities, which delivers data from different locations to different base stations. Specifically, there are N local base stations (BSs) in the system, denoted as $b_1, b_2, \ldots, b_N$. Each local BS b_i aims at collecting a certain type of data generated from remote IoT nodes located at a particular location. The UAV moves among different locations to collect data from different IoT nodes. The UAV is equipped with a group of differentiated queues to separately store different types of data collected from IoT nodes. When the UAV carries the data back to the target BS, the UAV can deliver the corresponding data from the queue to the BS, providing an over-the-air end-to-end delay tolerant data transmission to the BS. To incentivize the UAV to help BSs obtain data from remote IoT nodes, each local BS also provides a wireless energy charging service, which transfers energy to the communication module of the UAV as a compensation of data delivery.

When the system is in operation, the UAV senses internal parameter states and environmental states. Based on all the states sensed, the UAV makes energy and data management. A utility-based system model is applied to the UAV that each decision action of energy and data management results in a reward or a cost to the UAV. All the system and internal states, decision strategies, and rewards (as well as costs) potentially correlated mutually. As a result, to maximize the system performance in terms of the long-term utility of the UAV. In this work, the operation and decision making of the UAV is modeled as a Markov Decision Process (MDP), which can be formally denoted as a 5-tuple $<\mathbb{S}, \mathbb{A}, P, U, \gamma>$, where

- $\mathbb{S}$ denotes the state space including all possible system states of the UAV assisted communication system;
- $\mathbb{A}$ denotes the action space including all actions that the UAV can take;
- P is the transition probability indicating the chance that the UAV transit to a particular future state once an action is taken;
- U is the immediate utility as the reward in response to the current action taken by the UAV;
- γ is the discount factor of all the future reward received by the UAV. The discount factor indicates that the UAV as a decision maker is more in favor of short-term reward because of the volatility, long-term risks and uncertainties.

Based on the formal definition of MDP, the state space of the UAV assisted communication system is defined as follows:

$$\mathbb{S} = \left\{\left(\mathcal{L}, \mathcal{Q}^M, \mathcal{E}\right)\right\}, \tag{1}$$

where $\mathcal{S} = (\mathcal{L}, \mathcal{Q}^M, \mathcal{E}) \in \mathbb{S}$ is a composite system state including all the component system state variables $\mathcal{Q}^M$, $\mathcal{E}$, and $\mathcal{L}$, defined as follows:

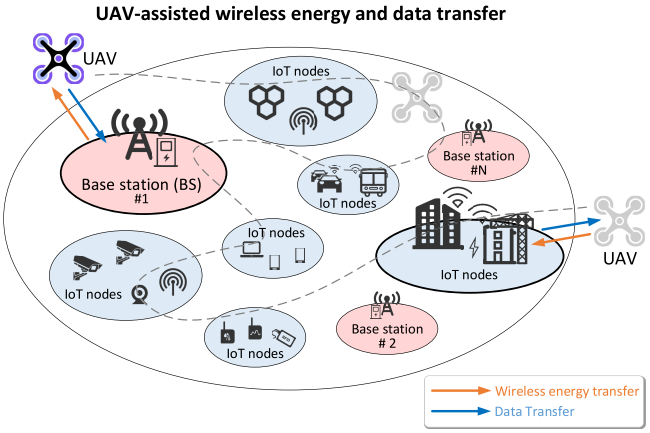


Fig. 1. UAV-assisted data delivery system with wireless energy charging.

- $\mathcal{L} \in \mathbb{L} = \{0, \ldots, L\}$ denotes the location that the UAV is currently visiting, where $\mathbb{L} = \{0, \ldots, L\}$ is the set of all the locations that the UAV can visit.
- $\mathcal{Q}^M$ is a composite state denoting the states of all the queues for M different data packets, where $\mathcal{Q}^M = (\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_M) \in \prod_{i=1}^{M} \mathbb{Q}_i$. The set $\mathbb{Q}_i = \{0, \ldots, Q_i\}$, which includes all the possible queue state of the i^{th} queue. Without loss of generality, we let $Q_i = Q$, $\forall i$, that is, all the queues have the same maximum capacity Q. From an information and human centric perspective, Each queue can represent a particular demand for a type of data, as in [39], [45].
- $\mathcal{E} \in \mathbb{E}$ is the energy state (i.e., the current energy level of the battery) of the UAV, where E is the maximum capacity of the stored energy in the battery.

By observing the current state $\mathcal{S}$ that the UAV is in, the UAV takes an action $\mathcal{A} \in \mathbb{A}$ in response, aiming at maximizing the expected utility (i.e., expected reward). The action space is defined as $\mathbb{A} = \{\mathsf{IDLE}, \mathsf{RECV}, \mathsf{CHRG}, \mathsf{TRAN}\}$, where each element denotes the actions of being idle, receiving IoT data from the current location, charging the UAV battery with wireless energy transfer, and transferring data to the BS from the corresponding queue, respectively.

IV. Markov Decision Process Formulation

In this section, we model the transition of all the system states that the UAV observes as Markovian. A Markov Decision Process (MDP) is formulated and solved to optimize the reward from the perspective of the UAV.

A. Location State Transition

In the system, the UAV moves among locations to deliver data packets. The set of locations can be further categorized into two subsets $\mathbb{L}_{BS}$ and $\mathbb{L}_N$, indicating the locations with BSs and IoT nodes, respectively. The subset $\mathbb{L}_{BS}$ is defined as follows:

$$\mathbb{L}_{BS} = \mathbb{L}_1^{\text{Rx}} \cup \mathbb{L}_2^{\text{Rx}} \cup \ldots \cup \mathbb{L}_M^{\text{Rx}}, \tag{2}$$

where each element $\mathbb{L}_i^{\text{Rx}}$ is a set of locations that request for the data $i \in \{1, 2, \ldots, M\}$. In similar manner, the subset $\mathbb{L}_N$

is defined by

$$\mathbb{L}_N = \mathbb{L}_1^{\mathrm{Tx}} \cup \mathbb{L}_2^{\mathrm{Tx}} \cup \ldots \cup \mathbb{L}_M^{\mathrm{Tx}}, \tag{3}$$

where $\mathbb{L}_i^{\mathrm{Tx}}$ is a set of all the locations that have the data $i \in \{1, 2, \ldots, M\}$. The UAV aims at deliver packets from each pair of sets $\mathbb{L}_i^{\mathrm{Tx}}$ and $\mathbb{L}_i^{\mathrm{Rx}}$, $\forall i$.

We assume that $\mathbb{L} = \mathbb{L}_{BS} \cup \mathbb{L}_N$ and $\mathbb{L}_{BS} \cap \mathbb{L}_N = \varnothing$. That is, each location has either a BS requesting for data packets, or IoT nodes generating data packets that a particular BS needs to collect their data. otherwise the IoT nodes can be directly activated by receiving energy from the BS, and deliver corresponding data packets to the BS. The total number of locations is $L = |\mathbb{L}_{BS}| + |\mathbb{L}_N|$. The movement pattern of the mobile IoT node can be general in the model. We define the transition probability of the UAV moving from the location $\mathcal{L}$ to $\mathcal{L}'$ as $\psi_{\mathcal{L},\mathcal{L}'} = P_L(\mathcal{L}, \mathcal{L}')$. The transition matrix of the location state $\mathcal{L}$ of the UAV can be expressed by the following transition matrix:

$$\mathbf{L} = \begin{bmatrix} \psi_{0,0} & \cdots & \psi_{0,L} \\ \vdots & \ddots & \vdots \\ \psi_{L,0} & \cdots & \psi_{L,L} \end{bmatrix}. \tag{4}$$

B. Queue-State Transition Matrix

When the UAV is located at a location $\mathcal{L} \in \mathbb{L}_N$ where IoT nodes generate data packets, the UAV has an option to receive data generated at the IoT nodes, i.e., $\mathcal{A} = \mathsf{RECV}$, or do nothing, i.e., $\mathcal{A} = \mathsf{IDLE}$. Note that the UAV also has an option to take the actions $\mathcal{A} = \mathsf{CHRG}$ or $\mathcal{A} = \mathsf{TRAN}$. However, the action of charging or transferring data will yield no reward at all at a certain cost. In that case, the UAV will not take the actions $\mathcal{A} = \mathsf{CHRG}$ and $\mathcal{A} = \mathsf{TRAN}$. After the action $\mathcal{A} = \mathsf{RECV}$ is taken, the received data will be stored in the corresponding queue in the UAV. For example, when the UAV is at a location $\mathcal{L} \in \mathbb{L}_i^{\mathrm{Tx}}$, the received data will be inserted into the queue $\mathcal{Q}_i$.

We denote that the i^{th} queue may receive k units of data packets, given the action $\mathcal{A} = \mathsf{RECV}$ taken by the UAV at the location $\mathcal{L} \in \mathbb{L}_i^{\mathrm{Tx}}$. The probability of k can be denoted as $\rho(k)$, $k = 0, 1, \ldots, +\infty$, indicating that the IoT nodes at the location may generate a random number of data packets. The transition matrix for the i^{th} queue is defined as a $(Q+1) \times (Q+1)$ upper triangular matrix $\mathbf{Q}_i^+$, as follows:

$$\mathbf{Q}_i^+ = \begin{bmatrix} \rho(0) & \rho(1) & \cdots & \rho(Q-1) & \sum_{k=Q}^{\infty} \rho(k) \\ & \rho(0) & \cdots & \rho(Q-2) & \sum_{k=Q-1}^{\infty} \rho(k) \\ & & & \ddots & \vdots \\ & & & & 1 \end{bmatrix}, \tag{5}$$

where $\sum_{k=0}^{+\infty} P(k) = 1$. Note that each queue has a limited capacity that the obtained data exceeding the capacity will be discarded.

As the queue state $\mathcal{Q}$ is a composite state, when the i^{th} queue of the UAV increases, the other queues will remain the same, denoted by $\mathbf{Q}_j^0 = I_{Q+1}$, $j \neq i$, where I_{Q+1} is an $(Q+1) \times (Q+1)$ identity matrix. As a result, the overall transition matrix for the queue state $\mathcal{Q}$ when there is an increased queue size is denote by $\mathcal{Q}^+(\mathcal{L} \in \mathbb{L}_i^{\mathrm{Tx}}))$, as follows:

$$\begin{aligned} &\mathbf{Q}^+\Big(\mathcal{L} \in \mathbb{L}_i^{\mathrm{Tx}}, \mathcal{A} = \mathsf{RECV}\Big) \\ &\quad = \mathbf{Q}_0^0 \otimes \mathbf{Q}_1^0 \otimes \cdots \otimes \mathbf{Q}_i^+ \otimes \cdots \otimes \mathbf{Q}_M^0. \end{aligned} \tag{6}$$

When the UAV moves to a BS, the UAV can take two actions, i.e., energy charging and data transferring. The UAV transfers a group of δ data packets to the corresponding BS. The queue state of the i^{th} queue can decrease when the UAV takes the action to transfer the corresponding stored data packets at the location $\mathcal{L} \in \mathbb{L}_{BS}$. The data transmission process is not always successful. The probability of successful transmission at the current location is set as β. The queue length remains the same if the data transfer fails. The transition matrix of the queue state $\mathbf{Q}_i$ in this case is denoted as follows:

$$\mathbf{Q}_i^- = \begin{bmatrix} \Delta_{(\delta+1)\times(\delta+1)} & & & & \\ \beta_{\mathcal{L}} & \mathbf{0}_{1\times(\delta-1)} & 1-\beta_{\mathcal{L}} & & \\ & \ddots & & \ddots & \\ & & \beta_{\mathcal{L}} & \mathbf{0}_{1\times(\delta-1)} & 1-\beta_{\mathcal{L}} \end{bmatrix}, \tag{7}$$

where the matrix $\Delta_{(\delta+1)\times(\delta+1)}$ denotes the case that the UAV has less than δ data packets for transmission. The UAV may handle such a case in two different manners, as follows:

- Firstly, the UAV delivers only a whole group of δ packets together. In this case, the matrix $\Delta_{(\delta+1)\times(\delta+1)} = \mathbf{I}_{(\delta+1)\times(\delta+1)}$, which is an identity matrix. The UAV will transfer no packet to the BS even if the action $\mathcal{A} = \mathsf{TRAN}$ is taken.
- Secondly, the UAV transfers all packets in the corresponding queue. The matrix becomes

$$\Delta_{(\delta+1)\times(\delta+1)} = \begin{bmatrix} 1 & & & \\ \beta_{\mathcal{L}} & 1-\beta_{\mathcal{L}} & & \\ \vdots & & \ddots & \\ \beta_{\mathcal{L}} & & & 1-\beta_{\mathcal{L}} \end{bmatrix}, \tag{8}$$

  which denotes that the queue state $\mathcal{Q}_i$ is emptied with the probability $\beta_{\mathcal{L}}$ when there are less than δ data packets stored in the i^{th} queue.

The i^{th} queue state decreases as indicated by $\mathbf{Q}_i^-$. However, the states of other queues remain the same owing to the fact that the UAV only delivers the corresponding data packets from $\mathcal{Q}_i$ which is requested by the current BS. The composite queue state's transition matrix $\mathbf{Q}^-(\mathcal{L} \in \mathbb{L}_i^{\mathrm{Rx}})$ is given in the following equation:

$$\begin{aligned} &\mathbf{Q}^-\Big(\mathcal{L} \in \mathbb{L}_i^{\mathrm{Rx}}, \mathcal{A} = \mathsf{TRAN}\Big) \\ &\quad = \mathcal{Q}_0^0 \otimes \mathcal{Q}_1^0 \otimes \cdots \otimes \mathcal{Q}_i^- \otimes \cdots \otimes \mathcal{Q}_M^0. \end{aligned} \tag{9}$$

Lastly, in other cases, the composite queue state remains the same, e.g., when the UAV does not receive data at the location $\mathcal{L} \in \mathbb{L}_N$ or transfer data to base station at location $\mathcal{L} \in \mathbb{L}_{BS}$. The queue state transition matrix for the i^{th} queue is I_{Q+1}. As a result, the transition matrix $\mathbf{Q}^0$ of the composite queue

state is denoted by

$$\mathbf{Q}^0 = \mathbf{Q}_0^0 \otimes \mathbf{Q}_1^0 \otimes \cdots \otimes \mathbf{Q}_M^0, \tag{10}$$

where $\otimes$ is the Kronecker product.

C. Energy State Transition Matrix

The energy charged into the UAV is used for the data transmission. We consider that the UAV device has separated energy supply for movement and data delivery. We only discuss the energy charged for data delivery, e.g., separated batteries for basic movement and communication functionalities. In this section, the energy state transition matrix of the UAV is derived.

The energy state may increase. This occurs when the UAV is at location with an energy charger, i.e., base stations $\mathcal{L} \in \mathbb{L}_{BS}$, and the charging action $\mathbb{A} = \mathsf{CHRG}$ is also taken. The UAV receives E_{B} units of energy from a charger. We assume that the number of energy units received from the wireless energy charger cannot exceed the battery capacity of the UAV, i.e., E. The transition matrix for the energy state under the charging action is given as in (11), at the bottom of the page. The matrix $\mathbf{E}^+$ in (11) is $(E+1)\times(E+1)$ dimensional, with each row and each column of the matrix denoting the current energy state $\mathcal{E}$, and the energy state of the next decision period $\mathcal{E}'$, respectively. As the charging process is via wireless channels, $\xi_{\mathcal{L}}$ is the efficiency of energy charging at location $\mathcal{L} \in \mathbb{L}_{NC}$, i.e., the probability of successful charging. $\mathbf{0}_{(E_{\mathrm{B}}-1)\times 1}$ is a row vector of $E_{\mathrm{B}}-1$ zeros.

When the UAV transfers data to BSs, or receives data from IoT nodes, energy will be consumed. In this case, energy can be consumed to transfer data, or to charge and activate IoT devices in exchange of collecting data. That is, the energy state decreases when $\mathcal{A} = \mathsf{RECV}$ or $\mathcal{A} = \mathsf{TRAN}$ when $\mathcal{L} \in \mathbb{L}_{BS}$ or $\mathcal{L} \in \mathbb{L}_N$, respectively. The UAV consumes E_{D} units of energy for data delivery. In this case, the energy state may decrease by E_{D}, except that when there is less than E_{D} units of energy in the battery. Considering the energy efficiency, we assume that at least one unit of energy will be transferred. The transition matrix is given in (12), which has the dimension of $(E+1)\times(E+1)$, as follows:

$$\mathbf{E}^-(\mathcal{L},\mathcal{A}) = \left[\begin{array}{c|ccc} \mathbf{I}_{(E_{\mathrm{D}}+1)\times(E_{\mathrm{D}}+1)} & & & \\ \hline \mathbf{0}_{1\times E_{\mathrm{D}}} & 1 & & \\ & & \ddots & \\ & & & 1 \quad \mathbf{0}_{1\times E_{\mathrm{D}}} \end{array}\right], \tag{12}$$

where $\mathbf{0}_{1\times E_{\mathrm{D}}}$ is a row submatrix of zeros. The submatrix $\mathbf{I}_{(E_{\mathrm{D}}+1)\times(E_{\mathrm{D}}+1)}$ is an identity matrix with the dimension of $(E_{\mathrm{D}}+1)\times(E_{\mathrm{D}}+1)$, which indicates that the stored energy in the UAV's battery is less than E_{D} and thus is not enough for performing further data delivery. The energy state remains the same.

Finally, the energy state can remain the same. For example, when the UAV neither receives energy nor transfers data at location $\mathcal{L} \in \mathbb{L}_{BS}$, or the UAV does not receive data at location $\mathcal{L} \in \mathbb{N}$, we have the transition matrix $\mathbf{E}^0 = I_{E+1}$, where $\mathbf{I}_{E+1}$ is an $(E+1)\times(E+1)$ identity matrix.

D. Overall Energy State Transition Matrix

Taking all the location state, energy state, and queue state into consideration, the transition matrix of entire state space can be derived, denoted as $\mathcal{W}(\mathcal{S},\mathcal{S}'|\mathcal{A})$. The current system state is denoted as $\mathcal{S} = (\mathcal{E},\mathcal{L},\mathcal{Q})$, and the system state of the next decision period is denoted by $\mathcal{S}' = (\mathcal{E}',\mathcal{L}',\mathcal{Q}')$. The overall system state transition matrix can be derived as follows:

$$\mathbf{T}\big(\mathcal{S},\mathcal{S}',\mathcal{A}\big) = \mathbf{L} \otimes \mathbf{B}\big(\mathcal{L},(\mathcal{E},\mathcal{Q}),\big(\mathcal{E}',\mathcal{Q}'\big),\mathcal{A}\big). \tag{13}$$

The transition matrix for combined energy-queue state is denoted as follows: $\mathbf{B}(\mathcal{L},(\mathcal{E},\mathcal{Q}),(\mathcal{E}',\mathcal{Q}'),\mathcal{A})$, which is derived as in (14), at the bottom of the next page. In the first condition of (14), the UAV takes data receiving action $\mathcal{A} = \mathsf{RECV}$ at the IoT node side, i.e., $\mathcal{L} \in \mathbb{L}_N$, given that the UAV has enough energy to support data collection. The energy state can decrease and queue state can increase. The second condition indicates that the energy is charged from the BS of the current location. The third condition denotes the data delivery process at the current BS (i.e., $\mathcal{L}_i \in \mathbb{L}_{BS}$) which is interested in the data packets with data type i. In this case, the UAV must have enough energy to supply energy for data transmission. The last condition denotes the case that both the data and energy queues in the UAV are kept unchanged, without any transferred or received data and energy.

E. Reward Functions for the UAV Delivering Data

To incentivize the UAV to move data from IoT nodes to BSs, a reward is required to encourage and compensate the UAV whenever data packets are successfully delivered to the corresponding BS. We define an immediate utility function $U(\mathcal{S},\mathcal{A})$ of the UAV as the reward of the current decision period, given the current system state $\mathcal{S}$ and the action $\mathcal{A}$ taken by the UAV. Different types of costs also incur in the processes

$$\mathbf{E}^+(\mathcal{L}\in\mathbb{L}_{BS},\mathbb{A}=\mathsf{CHRG}) = \left[\begin{array}{ccccc} 1-\xi_{\mathcal{L}} & \mathbf{0}_{(E_{\mathrm{B}}-1)\times 1} & \xi_{\mathcal{L}} & & \\ \ddots & & & \ddots & \\ & 1-\xi_{\mathcal{L}} & \mathbf{0}_{(E_{\mathrm{B}}-1)\times 1} & & \xi_{\mathcal{L}} \\ \hline & & & & \vdots \\ & & 1-\xi_{\mathcal{L}} & 0 & \xi_{\mathcal{L}} \\ & & & 1-\xi_{\mathcal{L}} & \xi_{\mathcal{L}} \\ & & & & 1 \end{array}\right] \tag{11}$$

of charging, data storing and delivery. The reward function $U(\mathcal{S},\mathcal{A})$ consists of different direct reward/cost components when the UAV operates under different system conditions, as follows:

- Charging cost: The UAV has an option to request for the wireless energy charging service at the current location. The unit price of energy charging can be different. The UAV can even request for charging in the locations where no wireless energy sources exist. In this case, the charging cost will be very high or $+\infty$ to prevent the UAV from taking the action of energy charging.
- Data delivery delay: As the data packets start to age immediately after being generated by the IoT nodes, the UAV needs to deliver the data collected in the queues to the corresponding BSs. There is a cost icurred by delivery delay before each data packet is actually delivered, i.e., after the action $\mathcal{A}=\mathsf{TRAN}$ is taken. Specifically, during each time slot, each of the data packets stored in the UAV will increase the overall delay of the UAV by the unit of one time slot.
- Data delivery reward: A reward will be given to the UAV by the BS which receives the requested data packets. Note that data packets can only be successfully delivered with a probability $\beta_{\mathcal{L}}$ when the UAV is at a location $\mathcal{L}\in\mathbb{L}_{BS}$. Otherwise, the data still cannot be delivered and the UAV will receive no reward.

Reward can encourage the UAV to operate in the system to deliver data perpetually. Note that the reward can be direct and indirect. By giving the UAV direct reward in the form of monetary or other tokens, the UAV is incentivized to deliver data to BSs. By replenish battery energy of the UAV using wireless energy charging, the UAV has potential to deliver data packets and obtain reward in the future. As a result, energy charging works as a form of indirect reward. The UAV is encouraged to charge energy when necessary, even at a certain cost.

The utility function $U(\mathcal{S},\mathcal{A})$ is defined as follows:

$$U(\mathcal{S},\mathcal{A}) = \begin{cases} -a_1 E_B\, C(\mathcal{L}) - a_0 D(\mathcal{Q}), & \mathcal{L}\in\mathbb{L}_{BS}, \mathcal{A}=\mathsf{CHRG} \\ a_2\beta_{\mathcal{L}}\delta R(\mathcal{L}) - a_0 D(\mathcal{Q}), & \mathcal{L}\in\mathbb{L}_{BS}, \mathcal{A}=\mathsf{TRAN}, \\ & \mathcal{E}\geq E_D \\ -a_0 D(\mathcal{Q}), & \text{otherwise}; \end{cases} \tag{15}$$

where a_0, a_1, and a_2 are weight factors taking values from $[0,1]$. In (15), the first condition denotes the cost incurred by energy charging price and data delay at the BS side. $C(\mathcal{L})$ is the unit charging price at the current location $\mathcal{L}\in\mathbb{L}_{BS}$. Energy price varies at the different chargers. $D(\mathcal{Q})$ denotes the delay cost of the data in all queues at the moment. The delay can be a linear function with respect to the current queue size, since the delay caused by each data packet increases by 1 during each decision period. As a result, the delay cost can be defined as $D(\mathcal{Q})=\sum_{i=0}^{M} Q_i$ for the current decision period. The second condition in (15) represents the utility gained by transferring data packets to the BS (i.e., $\mathcal{L}\in\mathbb{L}_{BS}$) from the corresponding queue, i.e., $\mathcal{A}=\mathsf{TRAN}$, given that there are enough energy stored in the battery of UAV. $R(\mathcal{L})$ is the reward per unit of data delivery at location $\mathcal{L}$. In our system, a fixed δ units of data will be transferred.

V. Solving the MDP Model: Value Iteration, Q Learning, and Deep Reinforcement Learning

A. Value Iteration Algorithm

When the system has finite number of few states, the MDP problem can be solved by value iteration algorithm to obtain the optimal policy. A Bellman equation of the MDP formulation can be formulated in a recursive manner as follows:

$$V(\mathcal{S}) = \max_{\pi(\mathcal{A}|\mathcal{S})} Q(\mathcal{S}|\mathcal{A}), \tag{16}$$

where

$$Q(\mathcal{S},\mathcal{A}) = U(\mathcal{S},\mathcal{A}) + \gamma \sum_{\mathcal{S}'} P\big(\mathcal{S},\mathcal{S}',\mathcal{A}\big)\, V\big(\mathcal{S}'\big). \tag{17}$$

The Bellman equation can be numerically solved by the value iteration algorithm. $V(\mathcal{S})$ denotes the discounted expected overall utility in the long-term, given the current system state $\mathcal{S}$ that the UAV observes. We denote $\pi^*(\mathcal{S})$ as the optimal strategy taken by the UAV, i.e., the optimal action $\mathcal{A}^*$ that the UAV takes in the current system state $\mathcal{S}$. The function $Q(\mathcal{S},\mathcal{A})$ defines the relation between the long-term expected overall utility and the current state-action pair. In $Q(\mathcal{S},\mathcal{A})$, the first term $U(\mathcal{S},\mathcal{A})$ is the current utility that the UAV can immediately obtain. The second term is the sum of expected future overall utility with respect to different future system states, settled into the current utility by a discount factor γ. The discount factor γ indicates the impacts of future risk when taking future utility into consideration.

B. Complexity Analysis

Based on analyzing the Bellman equation (16), the complexity of value iteration algorithm is $O(|\mathbb{A}|\cdot|\mathbb{S}|^2)$, where $|\mathbb{A}|$ and $|\mathbb{S}|$ are the size of action space and state space, respectively. As $|\mathbb{A}|$ is usually small and linearly correlated with the complexity, the size of state space $|\mathbb{S}|$ is the major factor that affects the complexity of value iteration algorithm in the proposed system.

$$\mathbf{B}\big(\mathcal{L},(\mathcal{E},\mathcal{Q}),\big(\mathcal{E}',\mathcal{Q}'\big),\mathcal{A}\big) = \begin{cases} \mathbf{Q}^{+}(\mathcal{L},\mathcal{A})\otimes\mathbf{E}^{-}(\mathcal{L},\mathcal{A}), & \mathcal{L}\in\mathbb{L}_N, \mathcal{E}\geq E_{\mathrm{D}}, \mathcal{A}=\mathsf{RECV} \\ \mathbf{Q}^{0}\otimes\mathbf{E}^{+}(\mathcal{L},\mathcal{A}), & \mathcal{L}\in\mathbb{L}_{BS}, \mathcal{A}=\mathsf{CHRG} \\ \mathbf{Q}^{-}(\mathcal{L},\mathcal{A})\otimes\mathbf{E}^{-}(\mathcal{L},\mathcal{A}), & \mathcal{L}\in\mathbb{L}_{BS}, \mathcal{E}\geq E_{\mathrm{D}}, \mathcal{A}=\mathsf{TRAN} \\ \mathbf{Q}^{0}\otimes\mathbf{E}^{0}, & \text{otherwise} \end{cases} \tag{14}$$

In practical UAV systems collecting information from various geographical locations with adequate energy quota for communication, as well as data storage, solving the optimal decisions of the UAV could result in *the curse of dimensionality*. That is, as the number of possible states increases, e.g., the numbers of locations, fine grained battery energy management, or data types in practical cases, solving the MDP problem by directly using the value iteration algorithm could be infeasible. Moreover, the value iteration algorithm requires complete information of system states, which may not be possible in practical system scenarios. For example, the UAV may not know the exact locations to visit before departure from the initial locations. In practical systems, UAVs are possibly required to build their knowledge and information about the system state space.

To deal with the aforementioned issues, in the next section, a reinforcement learning based approach and a DRL based approach will be introduced to solve the optimal UAV strategies in the cases of large system state space.

C. Optimizing UAV Strategies With Reinforcement Learning and Deep Reinforcement Learning

1) Reinforcement Learning Based Optimization Scheme: As discussed in the complexity analysis, conventional value iteration algorithm requires complete knowledge of all the system states, actions, as well as state transition probabilities to solve the MDP. As a result, value iteration suffers from the exponential complexity as the size of system increases. In this study, we revisit the MDP problem and solve it by using deep reinforcement learning (DRL) algorithm to avoid the curse of dimensionality.

Reinforcement learning (RL) method can assist a decision making agent (i.e., the UAV) to achieve optimal decisions *without knowledge on system states and transitions a priori*. In each decision period, the agent takes an action and receives a reward accordingly. Then the UAV transits to the next system state. The agent records and evaluates the optimal action with respect to each encountered system state, which leads to the long-term ultimate optimization objective of the agent. For example, a walking robot determines its optimal moving direction in every location to achieve the shortest path to the destination.

As a typical RL method, Q-learning employs a Q-function mapping the current state and possible actions of the agent to the expected total reward. A Q-table containing all the optimal reward values and state-action pairs is recorded by the decision making agent. The agent maximizes the Q-function by strategically comparing and selecting the optimal action in each possible state. ϵ-greedy approach is employed for balancing the exploitation and exploration tradeoff to determine the current action taken by the UAV. Given the current state, the UAV chooses the known action which results in already optimized results at a probability ϵ, and takes random actions for finding potentially improved actions that are not discovered at a probability $1-\epsilon$. Note that, different from the value iteration algorithm, the Q-table maintained by the Q-learning algorithm does not necessarily include all the state-action pairs. The Q-learning algorithm allows the agent to explore the system and update the Q-table given the record exists in the table. Otherwise, new records can be dynamically added to the Q-table.

Algorithm 1: Q-Learning Algorithm for UAV Energy and Data Management

Input: $\mathbb{S}, \mathbb{A}$
Output: $\phi^* : \mathcal{S} \mapsto \mathcal{A},\ \forall \mathcal{S} \in \mathbb{S},\ \forall \mathcal{A} \in \mathbb{A}$
Init: $Q(\mathcal{S}, \mathcal{A}) \leftarrow 0,\ \forall \mathcal{S},\ \forall \mathcal{A}$
repeat
  Select action $\mathcal{A}_t$ with ϵ-greedy method;
  Obtain U as reward; obtain next state $\mathcal{S}_{t+1}$;
  Update Q table:
    $Q(\mathcal{S}, \mathcal{A}) \leftarrow Q(\mathcal{S}, \mathcal{A}) + \alpha\big[U(\mathcal{S}, \mathcal{A}) + \gamma \max_a Q(\mathcal{S}', a) - Q(\mathcal{S}, \mathcal{A})\big]$;
  Update current state $\mathcal{S} \leftarrow \mathcal{S}'$
until $\mathcal{S}_t$ *is terminal state, or maximum iteration reached*;
return $\phi^*(\mathcal{S}) \triangleq \operatorname{argmax}_{\mathcal{A}} Q(\mathcal{S}, \mathcal{A})$

The updating process of Q-learning can also be described by using a Bellman equation

$$Q(\mathcal{S}, \mathcal{A}) \leftarrow Q(\mathcal{S}, \mathcal{A}) + \alpha\Big[U(\mathcal{S}, \mathcal{A}) + \gamma \max_a Q\big(\mathcal{S}', a\big) - Q(\mathcal{S}, \mathcal{A})\Big], \tag{18}$$

where α is learning rate ranging from 0 to 1, and $U(\mathcal{S}, \mathcal{A})$ is the immediate utility obtained by the agent (i.e., the UAV). The idea of the Q-function updating process is to gradually update the Q-value of each state-action pair in the Q-table with the difference between the optimal Q-value of the current decision period and the original Q-value before the current round of updating, i.e., the difference between $U(\mathcal{S}, \mathcal{A}) + \gamma \max_a Q(\mathcal{S}', a)$ and $Q(\mathcal{S}, \mathcal{A})$. Note that, Q-learning does not require any one-step action look-ahead calculation to obtain the optimal Q-value, compared with value iteration algorithm, which requires full transition probabilities $P(\mathcal{S}, \mathcal{S}', \mathcal{A})$. Q-learning only observes the consequent reward and state, and then selects the optimal action from the Q-table.

The decision making optimization algorithm with Q-learning for the wireless energy charging UAV is shown in Algorithm 1.

2) Deep Reinforcement Learning: Q-learning has its drawbacks when dealing with extremely large system state space. For example, when the UAV needs to precisely manage the energy storage, the UAV will set a large energy storage state space. In this case, even a slight change on the energy, e.g., charging, will result in a change of energy state. The UAV is required to include a large Q-table to store all the possible state-action pairs as well as Q-values accordingly, even if many of the states may not be encountered practically. A large volume of data storage will be consumed hence. As a result, Q-learning may not be feasible for devices with limited storage, e.g., sensors, mobile personal devices, and UAVs. Moreover, with a large amount of system states, the speed of convergence to optimal Q-values becomes low.

To overcome the drawbacks of Q-learning, deep Q network (DQN) based DRL has been introduced by combining deep learning techniques with reinforcement learning methods. Deep Q network (DQN), as one of the typical DRL algorithms, is a DRL counterpart of Q-learning, where the Q-table is replaced by a neural network with a small-sized memory pool for sampling the past experience of the agent. Unlike Q-table, the complexity of training a neural network does not increase when the number of system states increases, which enables DQN to be applied to large-scale UAV systems. The objective is to optimize the long-term utility $Q(\mathcal{S}, \mathcal{A}; \boldsymbol{\theta})$ by maximizing the weights $\boldsymbol{\theta}$ of the neural network. To train the neural network within DQN, the DQN inputs the neural network with observed system parameters $(\mathcal{S}, \mathcal{A}, U, \mathcal{S}')$, including the current system state $\mathcal{S}$, the action taken $\mathcal{A}$, the immediate utility $U(\mathcal{S}, \mathcal{A})$, and the next system state $\mathcal{S}'$ observed by the agent (i.e., the UAV). The weights $\boldsymbol{\theta}$ of the neural network will be updated during the training process. The neural network operates as a non-linear approximator $Q(\mathcal{S}, \mathcal{A}) = Q(\mathcal{S}, \mathcal{A}; \boldsymbol{\theta})$ which abstracts the potential features of the decision making tasks, i.e., data and energy transmission decisions in the proposed system.

The updating process of $\boldsymbol{\theta}$ may result in instability caused by the correlation of samples $(\mathcal{S}, \mathcal{A}, U, \mathcal{S}')$. In DQN, experience replay is employed to eliminate the correlation among all the samples. A fundamental assumption in DQN is that training samples are *i.i.d.* with each other. In each decision period, the agent (UAV) does not directly update $\boldsymbol{\theta}$ with the current sample $(\mathcal{S}, \mathcal{A}, U, \mathcal{S}')$. The current sample is stored in the pool. Instead, the agent randomly selects some historical samples for training. In the training process, the target function y^{DQN} is calculated by using Bellman equation, as follows:

$$y^{DQN} \triangleq U + \gamma \max_{\mathcal{A}} Q\left(\mathcal{S}', \mathcal{A}; \boldsymbol{\theta}^{-}\right) \tag{19}$$

where $\boldsymbol{\theta}^{-}$ is the weights that are not updated. The weights $\boldsymbol{\theta}$ are updated by minimizing the following loss function with gradient descent:

$$L(\boldsymbol{\theta}) = \mathbb{E}_{(\mathcal{S}, \mathcal{A}, U, \mathcal{S}')}\left[\left|y^{DQN} - Q(\mathcal{S}, \mathcal{A}; \boldsymbol{\theta})\right|^2\right]. \tag{20}$$

The parameters $\boldsymbol{\theta}^{-}$ in the target function will be updated by new values $\boldsymbol{\theta}$ every certain iterations of training to reduce correlation. By fixing the weights $\boldsymbol{\theta}$ for several training iterations, the instability of the training and weight updating processes can be reduced, thereby leading to lower risk of divergence in training.

Algorithm 2: DQN Algorithm for UAV Energy and Data Management

Input: $\mathbb{S}, \mathbb{A}$
Output: $\phi^* : \mathcal{S} \mapsto \mathcal{A},\ \forall \mathcal{S} \in \mathbb{S},\ \forall \mathcal{A} \in \mathbb{A}$
repeat
  Given initial state $\mathcal{S}_t$, $t = 0$;
  repeat
    Select action $\mathcal{A}_t$ with ϵ-greedy method;
    Obtain U as reward; obtain next state $\mathcal{S}_{t+1}$;
    Store $(\mathcal{S}_t, \mathcal{A}_t, U(\mathcal{S}_t, \mathcal{A}_t), \mathcal{S}_{t+1})$ to memory pool;
    Sample $(\mathcal{S}_k, \mathcal{A}_k, U(\mathcal{S}_k, \mathcal{A}_k), \mathcal{S}_{k+1})$ from memory pool;
    Gradient descent: $\left|y_k - \hat{Q}(\mathcal{S}_k, \mathcal{A}_k; \boldsymbol{\theta}^{-})\right|^2$, where $y_k \triangleq U + \gamma \max_{\mathcal{A}_{k+1}} Q(\mathcal{S}_{k+1}, \mathcal{A}_{k+1}; \boldsymbol{\theta}^{-})$;
    Update $\boldsymbol{\theta}^{-} \leftarrow \boldsymbol{\theta}$ every C steps;
  until $\mathcal{S}_t$ *is terminal state, or maximum iteration reached*;
until $\boldsymbol{\theta}^{-}$ *and* $\boldsymbol{\theta}$ *converge*;
return $\phi^*(\mathcal{S}) \triangleq \operatorname{argmax}_{\mathcal{A}} Q(\mathcal{S}, \mathcal{A}; \boldsymbol{\theta})$

VI. Numerical and Simulation Results

A. System Settings, Baseline Schemes, and Evaluation Metrics

The following system parameters are employed in the numerical evaluation of MDP, Q-learning, and DQN based schemes, unless otherwise stated.

- There are two pairs of locations in the system, $\mathbb{L} = \{1, 2, \ldots, 4\}$, in which $\mathbb{L}_{BS} = \{1, 2\}$ and $\mathbb{L}_N = \{3, 4\}$. We assume that the BSs in locations $\mathcal{L} = 1, 2$ request for data packets generated from IoT nodes in locations $\mathcal{L} = 3, 4$, respectively. The UAV moves among different location randomly with a uniform probability.
- The capacity of UAV battery is set as $E = 5$.
- The data queues facilitated in the UAV can receive 2 types of data packets, for BSs in the locations $\mathcal{L} = 1, 2$, respectively. In each data queue, the maximum queue capacity is set as 5, i.e., $Q_i = 5,\ \forall i$.
- The UAV receives $E_B = 2$ units of energy from each wireless energy charger, if the charging action $\mathcal{A} = \mathsf{CHRG}$ is successful. The UAV consumes $E_D = 1$ unit of energy by successfully transferring or collection a group of δ data packets.
- The probability of successful energy receiving is set as $\xi_{\mathcal{L}} = 0.85$ as energy transmission may fail, $\forall \mathcal{L} \in \mathbb{L}$. The probability of successful data packets delivery is set as $\beta_{\mathcal{L}} = 1.0,\ \forall \mathcal{L} \in \mathbb{L}$, i.e., guaranteed data delivery.
- In the utility function of the UAV, the coefficients are defined as $a_0 = 0.5,\ a_1 = 1.0,\ a_2 = 1.0$. The discount factor in Bellman equation is $\gamma = 0.95$.
- The wireless energy charging prices in different locations are not the same. The cost function of wireless energy charging is defined as $C(\mathcal{L}) = \frac{\mathcal{L}_i}{|\mathbb{L}|}$. The unit reward of data delivery in different locations are set as a uniform value $R(\mathcal{L}) \equiv 0.8$.

The MDP, Q-learning, and DQN based schemes are compared with three baseline schemes of UAV strategies, including

- *Greedy* scheme, where the UAV always takes an myopic action to maximize the current immediate utility function $\mathbf{C}(\mathcal{S}, \mathcal{A})$ in the current decision period, regardless of previous and possible future system states.
- *Random* scheme, where the UAV takes random actions from $\mathbb{A}$ in each location, with the probability of $\frac{1}{3}$ each.
- *Location-aware random* scheme, where the UAV rules out invalid actions in each location first, and select from remained actions randomly in each location. That is, the

UAV takes action $\mathcal{A} = \mathsf{RECV}$ in the region $\mathbb{L}_N$, and takes random action $\mathcal{A} = \mathsf{TRAN}$ and $\mathcal{A} = \mathsf{CHRG}$ in the region $\mathbb{L}_{BS}$ with the probability $\frac{1}{2}$ each.

The evaluation metrics compared in the numerical results are listed as follows:

- *Average long-term expected utility*: We assume that the UAV can be initialized (cold start) from any system state $\mathcal{S} \in \mathbb{S}$ uniformly. The average expected utility is defined as $\frac{\sum_{\mathcal{S}\in\mathbb{S}} V(\mathcal{S})}{|\mathbb{S}|}$.
- *Average delay*: The average delay is defined as the average single data queue length over all the decision periods that the UAV is in.
- *Relay energy efficiency*: We define the metric of relay efficiency to indicate whether the UAV has the ability to deliver data with enough energy storage. That is, the probability that there is data packets in the UAV and the UAV also has enough energy for data delivery to BSs, e.g., $\mathcal{Q}_i \geq \delta$ and $\mathcal{E} > E_{\mathrm{D}}$ in the experimental scenario. Otherwise, BSs will not be served by the UAV on all accounts due to not receiving data packets at all. A low relay energy efficiency indicates a low quality of data transmission service that the UAV has inefficient energy management strategy to serve the data sensing and delivery tasks.

B. System Performance Evaluation on Long-Term Utility of the UAV

In this section, performance metrics of the proposed MDP scheme are investigated. Impacts of several system parameters on the long-term utility of the UAV has been examined, which is also the optimization objective of the UAV. We employ simulation approach to investigate the performance of the proposed MDP based scheme. Firstly, the MDP model is solved by value iteration algorithm. The function of optimal strategies $\pi^* : \mathcal{S} \mapsto \mathcal{A}$ of the UAV is recorded for making data and energy transfer decisions in the system. The UAV starts the simulation from an initial status with the compound state of all zeros. The long-term expected utility of the UAV is measured from the initial state. In each round of simulation, the UAV experiences 200 decision periods.

1) UAV Parameters: The numerical experiment varies the maximum energy storage E of the UAV from 1 to 15. As observed in Fig. 2(a), the long-term expected utility of the UAV increases as the E increases. The reason is that, with more energy storage capacity, the UAV can store more energy to potentially support further successful data collection and data delivery. As E becomes large, the increasing rate of utility becomes lower, since the data delivery throughput of the UAV are limited in this case, although sufficient energy will be supplied with the increasing energy capacities.

The maximum data storage (i.e., queue length) Q increases from 1 to 15. As shown in Fig. 2(b), the long-term expected utility increases first and then tends to decrease. This is because, as the data capacity at the UAV increases, more data packets can be stored and delivered. However, as Q becomes too large, significant delay can be caused since the data delivery rates of the UAV to BSs are limited. There will be too many data packets held up in the queues of the UAV.

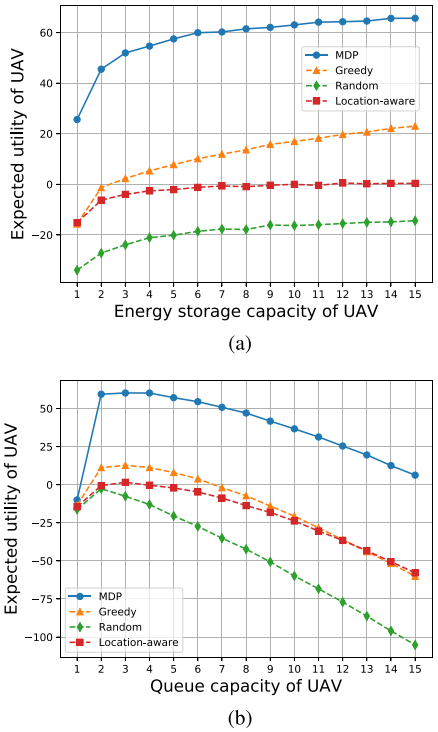


Fig. 2. (a) Impacts of UAV energy storage to utility, and (b) Impacts of UAV data storage to utility.

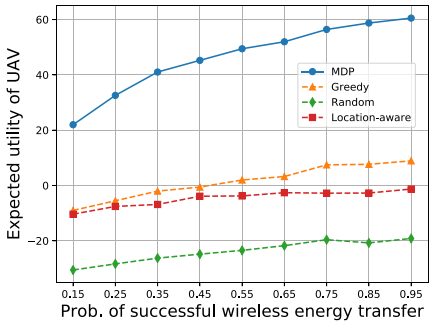
The proposed MDP scheme outperforms baseline schemes as compared in Figs. 2(a) and (b), since the MDP scheme can optimally select the strategies to support data delivery and energy charging to achieve the optimization goal, i.e., earning more revenue in the system. The MDP-based scheme also dynamically adjusts the optimal strategies in extreme system states. As shown in Fig. 2(b), when Q becomes too large, the MDP-based scheme suffers least from utility decrease by strategically reducing collecting data packets from IoT nodes to maximize the UAV utility.

2) Energy Charging, Data Delivery, and Data Arrival: In this section, impacts of wireless energy charging parameters are investigated.

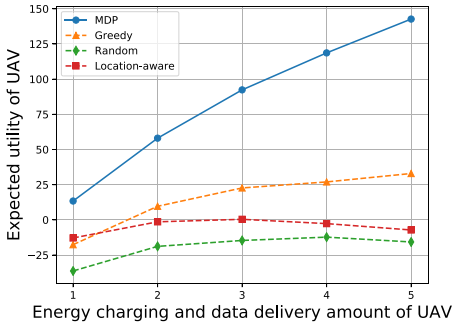
We vary the energy charging efficiency $\xi_{\mathcal{L}}$ from 0.15 to 0.95 in each location with an energy charging BS, i.e., $\mathcal{L} \in \mathbb{L}_{BS}$. As shown in Fig. 3(a), the expected utility increases as the UAV has higher chance to receive energy in each wireless energy charging process. Even when the energy charging efficiency is low (e.g., $\xi_{\mathcal{L}} = 0.15$), the MDP-based scheme still manages to utilize limited energy to achieve higher long-term utility compared with baseline schemes.

The energy charging action $\mathcal{A} = \mathsf{CHRG}$ will result in an increase of E_B energy units in the energy storage of the UAV. The UAV utilizes the stored energy to deliver δ data packets by taking the action $\mathcal{A} = \mathsf{TRAN}$. Firstly, we vary both E_B and δ from 1 unit to 5 units at the same time, i.e., increasing the energy charging and data delivery throughputs. As expected, the long-term expected utility increases as shown in Fig. 3(b).

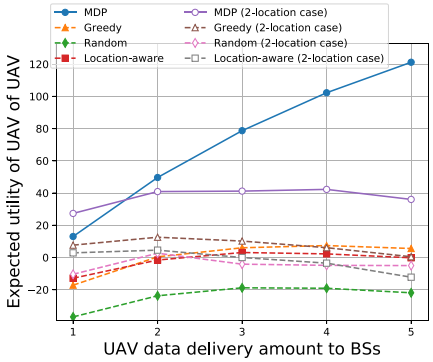
We also fix the charged energy amount $E_B = 2$, and increase δ only. That is, we investigate data delivery performance of the UAV under the situation of limited energy



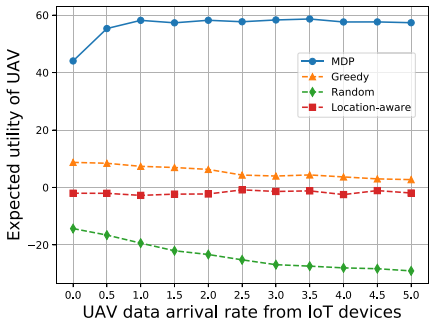
(a)



(b)



(c)



(d)

Fig. 3. (a) Impacts of changing the probability of wireless energy transfer at base station location $\xi_{\mathcal{L}}$, (b) Impacts of changing the energy transfer amount E_B to UAV and data delivery amount δ from UAV, (c) Impacts of only changing the data delivery amount δ from UAV, and (d) Impacts of Poisson data arrival rate of UAV data collecting at IoT nodes.

replenish rate. Two cases are examined, where the system has only one BS and two BSs, respectively. As shown in Fig. 3(c), in the case that there is only one BS, the UAV is not able to obtain enough energy for data delivery. The long-term utility is limited, which also suggests a downtrend. After the number of BSs only increases to 2, the UAV can obtain adequate energy to support further data transmission with an optimized increasing long-term expected utility, compared with the baseline schemes which fail to achieve the same optimal strategies.

When the UAV is at the IoT device side and takes the action of $\mathcal{A} = \mathsf{RECV}$, the collected number of data from the IoT devices follows a Poisson process. As shown in Fig. 3(d), we vary the Poisson data arrival rate λ from 0.0 to 5.0. The long-term utility of the UAV increases and then remain at a certain level. Compared with the baseline schemes where the fast arrived data packets result in higher delay and hence lower utility, the UAV refrains from taking the action of $\mathcal{A} = \mathsf{RECV}$ to reduce the impacts of data packet delay.

C. Data Delivery Delay

In this section, we investigate the impacts of system parameters to the average delay. We investigate the long-term average delay of the UAV by simulation, where the UAV starts the simulation with full load of data. In each round of simulation, the UAV experiences 200 decision periods.

As in depicted Fig. 4(a), by optimizing the UAV strategies, the average delay decreases first and then starts to increase as the maximum energy capacity E increases from 1 to 15. This indicates that a UAV with larger energy storage for data communication may suffer from delay, because more energy storage might incentivize the UAV to receive and store more data packets from IoT nodes whenever the UAV has access to the IoT nodes. In this case, data storage amount increases, leading to higher average delay. In Fig. 4(b), an increase in the maximum queue capacity state Q intuitively results in an increase in the average delay metric, since more data packets can be stored in the UAV for further delivery.

Figure 4(c) shows the performance of data delivery delay when the number of data packets δ increases, i.e., the amount of data packets that the UAV can send to BSs during the period that each data transferring action $\mathcal{A} = \mathsf{TRAN}$ is taken. However, as the energy charging rate still remains the same, the UAV is constrained by limited energy supply. To avoid failure in data delivery caused by inadequate energy storage, the UAV holds the data packets within data storage queues until enough energy is charged. As a result, the average delay significantly increases as δ increases. By taking the strategy of tolerating more delay, the optimization objective of maximizing the long-term utility, can be achieved, as also proved in Fig. 3(c).

D. Energy Efficiency

The system performance in terms of relay energy efficiency will be discussed in this section. The simulation setting is the same as Section VI-C.

From Figs. 5(a) and (b), the relay energy efficiency will increase in both the situations of increased energy capacity E and queue capacity Q. The reason is that the relay energy efficiency measures the probability of having adequate energy for data delivery. An increase in either E or Q ensures that there are enough energy and data for data delivery whenever the

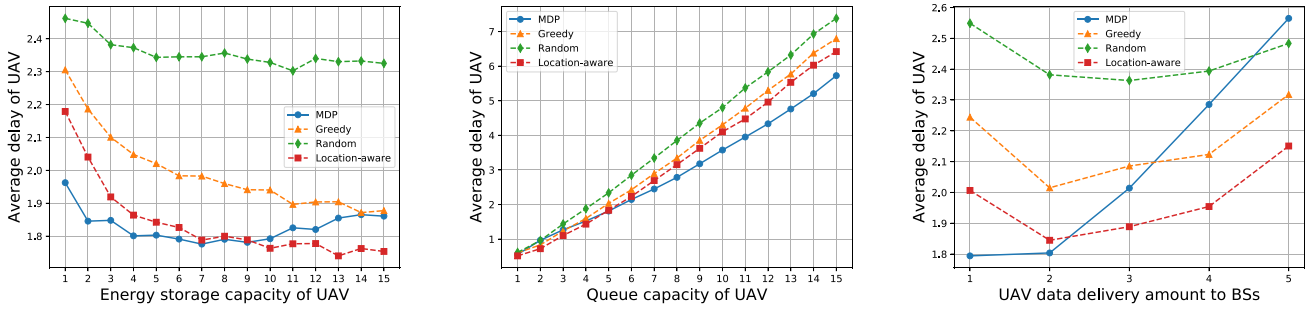


Fig. 4. (a) Impacts of UAV energy storage to utility to UAV average delay, (b) Impacts of UAV data storage to utility to UAV average delay, and (c) Impacts of only changing the UAV data delivery amount δ at base station.

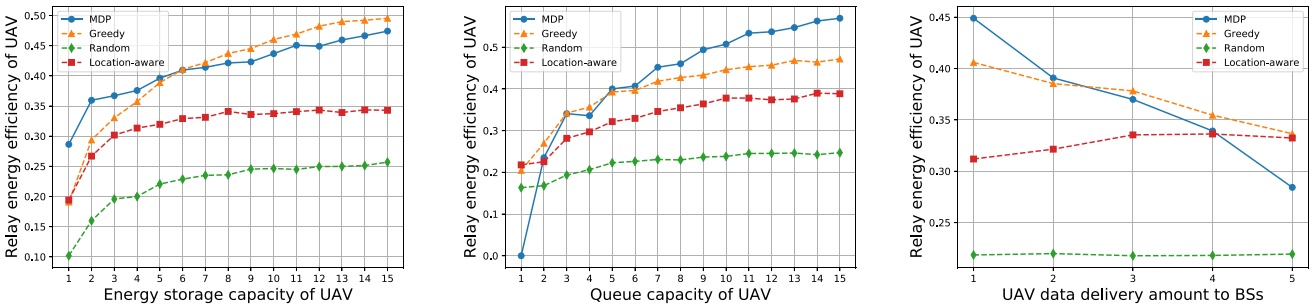


Fig. 5. (a) Impacts of Impacts of UAV energy storage to utility to UAV relay energy efficiency, (b) Impacts of UAV data storage to utility to UAV relay energy efficiency, and (c) Impacts of only changing the UAV data delivery amount δ at base station to UAV relay energy efficiency.

UAV arrives at a BS. However, as shown in Fig. 5(c), when the number of data packets δ increases, and the energy charging amount E_B remains the same, the relay energy efficiency decreases, since the data transferring requires more energy to deliver more data packets. In this case, the UAV may not have enough energy stored when data delivery is required. As shown in Fig. 3(c) and Fig. 5(c), with an MDP-based optimization scheme, relay energy efficiency is compromised in exchange of optimized long-term expected utility of the UAV.

E. DQN Results

To compare the proposed DRL based algorithm to solve the optimal UAV strategies of data delivery and energy transfer, the simulation results of long-term utility and average delay of the UAV are derived as shown in Fig. 6, respectively. In the simulation, the DQN and Q-learning based schemes are executed for 1500 episodes from initial system states. In each episode, the UAV experiences 200 decision making period, i.e., making 200 moves. A ϵ-greedy exploration is employed, such that the UAV has a probability of $1-\epsilon$ to choose an action with highest expected reward, and a probability of ϵ to randomly take new possible actions. With the ϵ-greedy exploration, the UAV takes different traces in each episode.

As shown in Fig. 6(a), in the simulation process of both DQN, the UAV trains the neural network model by employing historical data. After about 300 episodes of training, the average utility obtained by the DQN scheme converges to the maximum and stable value. Compared with the Q-learning scheme, the DQN scheme converges slower, since the neural

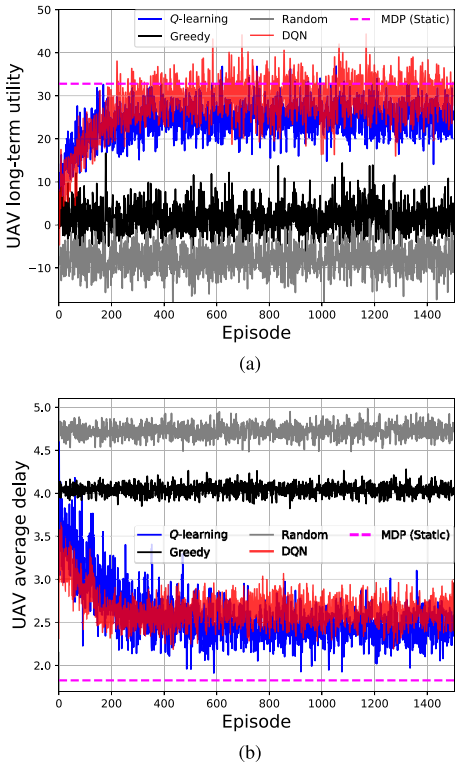
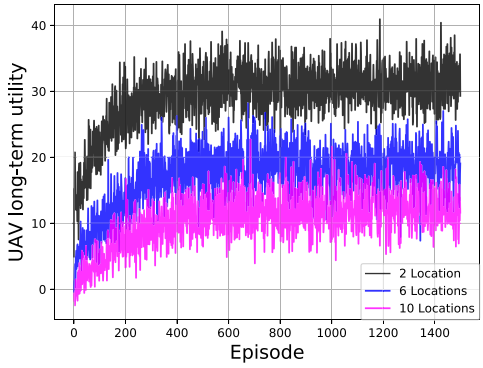
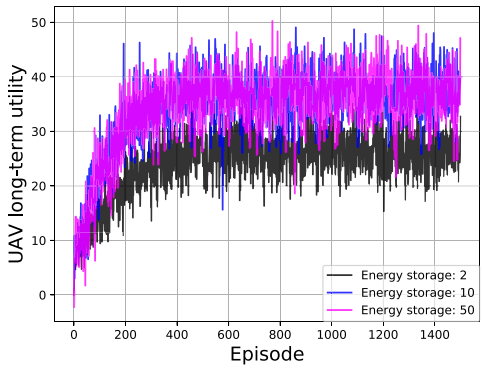


Fig. 6. Simulation results of DRL and baseline schemes in terms of (a) UAV utility, and (b) UAV average delay of data delivery.

network as an approximator does not record system state-action trajectories. However, as shown in Fig. 6(a), the DQN scheme outperforms the Q-learning scheme and other baseline



(a)



(b)

Fig. 7. DRL simulation results under different (a) location numbers, and (b) UAV average delay of data delivery.

schemes, i.e., greedy scheme and random scheme. The reason is that DQN adequately exploits uncorrelated historical state-action-reward data to obtain more information on the system parameters and features. The approximated results from DQN scheme is close to the static optimization results from numerically solving MDP, where the UAV is supposed to know all the system states, state transition probabilities and rewards.

Similarly, Fig. 6(b) illustrates the average delay occurred to the UAV. In the simulation, we set the energy storage of the UAV as a full state, i.e., fully charged battery. The UAV also shows significant performance on reducing the UAV average delay by strategically optimizing the data delivery and energy charging actions. After 300 episodes of training, the UAV adjusts its strategies and reaches a low level of average delay of data delivery. Note that the performance of the DQN scheme is slightly worse in terms of the UAV average delay. The reason is that delay is only a part of the optimization objective, i.e., to maximize the long-term expected utility, as in Fig. 6(a). The DQN scheme can choose to trade the average delay performance for better overall performance in terms of utility, by adjusting the UAV behaviours on other actions, e.g., energy charging and data collecting.

Figures 7(a) and (b) illustrate the feasibility of applying DRL when the system state set becomes large. In Fig. 7(a), the number of total locations $|\mathbb{L}|$ increases from 2 to 10. As assumed, the UAV has a specific queue for each corresponding location of IoT nodes. When the location number increases by 5 times, the total number of queue state also increases by $5Q$ times. From Fig. 7(a), the proposed DRL still solves the optimized long-term expected utility of the UAV. As the location number increases, the long-term expected utility decreases. The reason is that the UAV has more queues to store data packets, which consequently results higher delay during the operation of the UAV. Figure 7(b) shows the long-term expected utility of the UAV when the energy storage capacity increases to a large amount, i.e., $E = 2, 20, 50$, respectively. As shown from Fig. 7(b), an increase of energy storage capacity results in an increase of the long-term expected utility. However, as the energy storage capacity is too large, the UAV is not able to fully utilize all the capacity. Therefore, the long-term expected utility gained by the UAV does not increase in this case.

VII. Conclusion

In this work, we have proposed a wireless energy transfer supported UAV system to collect data from different geographical locations, and deliver the data to their destinations. We have modeled the mobility, energy storage and data storage patterns by employing Markove decision process to consider the time-variant system states observed by the UAV and incorporate their impacts into the decision strategies. The MDP problem has been solved by using a value iteration algorithm, a Q-learning, and a DRL based schemes respectively to train the UAV for energy and data management schemes. Numerical simulations show that the MDP based data delivery and energy charging scheme outperforms conventional baseline schemes in terms of long-term expected utility of the UAV.

References

[1] L. Gupta, R. Jain, and G. Vaszkun, "Survey of important issues in UAV communication networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1123–1152, 2nd Quart., 2015.

[2] F. Lyu *et al.*, "LEAD: Large-scale edge cache deployment based on spatio-temporal WiFi traffic statistics," *IEEE Trans. Mobile Comput.*, early access, Apr. 2, 2020, doi: 10.1109/TMC.2020.2984261.

[3] Q. Ye, W. Zhuang, S. Zhang, A. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[4] Y. Zeng, R. Zhang, and T. J. Lim, "Throughput maximization for UAV-enabled mobile relaying systems," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 4983–4996, Dec. 2016.

[5] C. Zhan, Y. Zeng, and R. Zhang, "Energy-efficient data collection in UAV enabled wireless sensor network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 328–331, Jun. 2018.

[6] J. Baek, S. I. Han, and Y. Han, "Energy-efficient UAV routing for wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1741–1750, Feb. 2020.

[7] T. D. Ponnimbaduge Perera, D. N. K. Jayakody, S. K. Sharma, S. Chatzinotas, and J. Li, "Simultaneous wireless information and power transfer (SWIPT): Recent advances and future challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 264–302, 1st Quart., 2018.

[8] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. Shen, and X. Li, "End-to-end quality of service in 5G networks: Examining the effectiveness of a network slicing framework," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 65–74, Jun. 2018.

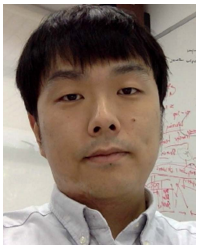
[9] S. Garg, K. Kaur, S. H. Ahmed, A. Bradai, G. Kaddoum, and M. Atiquzzaman, "MobQoS: Mobility-aware and QoS-driven SDN framework for autonomous vehicles," *IEEE Wireless Commun.*, vol. 26, no. 4, pp. 12–20, Aug. 2019.

[10] M. Abu Alsheikh, D. T. Hoang, D. Niyato, H. Tan, and S. Lin, "Markov decision processes with applications in wireless sensor networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1239–1267, 3rd Quart., 2015.

[11] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, Feb. 2015.

[12] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.
[13] 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, no. 4, pp. 434–437, Aug. 2017.
[14] H. He, S. Zhang, Y. Zeng, and R. Zhang, "Joint altitude and beamwidth optimization for UAV-enabled multiuser communications," *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 344–347, Feb. 2018.
[15] S. Zhang, H. Zhang, Q. He, K. Bian, and L. Song, "Joint trajectory and power optimization for UAV relay networks," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 161–164, Jan. 2018.
[16] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.
[17] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.
[18] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.
[19] H. Wang, G. Ding, F. Gao, J. Chen, J. Wang, and L. Wang, "Power control in UAV-supported ultra dense networks: Communications, caching, and energy transfer," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 28–34, Jun. 2018.
[20] L. Xie, J. Xu, and R. Zhang, "Throughput maximization for UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1690–1703, Apr. 2019.
[21] J. Xu, Y. Zeng, and R. Zhang, "UAV-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5092–5106, Aug. 2018.
[22] Y. Lai, Y. L. Che, S. Luo, and K. Wu, "Optimal wireless information and energy transmissions for UAV-enabled cognitive communication systems," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Chengdu, China, 2018, pp. 168–172.
[23] C. Su, F. Ye, L. Wang, L. Wang, Y. Tian, and Z. Han, "UAV-assisted wireless charging for energy-constrained IoT devices using dynamic matching," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 4789–4800, Jun. 2020.
[24] B. Galkin, J. Kibilda, and L. A. DaSilva, "UAVs as mobile infrastructure: Addressing battery lifetime," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 132–137, Jun. 2019.
[25] A. Trotta, M. Di Felice, F. Montori, K. R. Chowdhury, and L. Bononi, "Joint coverage, connectivity, and charging strategies for distributed UAV networks," *IEEE Trans. Robot.*, vol. 34, no. 4, pp. 883–900, Aug. 2018.
[26] S. Yin, L. Li, and F. R. Yu, "Resource allocation and basestation placement in downlink cellular networks assisted by multiple wireless powered UAVs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 2171–2184, Feb. 2020.
[27] S. Chai and V. K. Lau, "Online trajectory and radio resource optimization of cache-enabled UAV wireless networks with content and energy recharging," *IEEE Trans. Signal Process.*, vol. 68, pp. 1286–1299, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8982038
[28] V. Hassija, V. Saxena, and V. Chamola, "Scheduling drone charging for multi-drone network based on consensus time-stamp and game theory," *Comput. Commun.*, vol. 149, pp. 51–61, Jan. 2020.
[29] V. Hassija, V. Chamola, N. G. K. Dara, and M. Guizani, "A distributed framework for energy trading between UAVs and charging stations for critical applications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5391–5402, May 2020.
[30] F. Lyu *et al.*, "Characterizing urban vehicle-to-vehicle communications for reliable safety applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2586–2602, Jun. 2020.
[31] F. Lyu *et al.*, "Towards rear-end collision avoidance: Adaptive beaconing for connected vehicles," *IEEE Trans. Intell. Transp. Syst.*, early access, Jan. 22, 2020, doi: 10.1109/TITS.2020.2966586.
[32] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of UAVs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2124–2136, Mar. 2019.
[33] S. Garg, G. S. Aujla, N. Kumar, and S. Batra, "Tree-based attack–defense model for risk assessment in multi-UAV networks," *IEEE Consum. Electron. Mag.*, vol. 8, no. 6, pp. 35–41, Nov. 2019.
[34] C. Jiang, Y. Chen, Y. Gao, and K. J. R. Liu, "Joint spectrum sensing and access evolutionary game in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2470–2483, May 2013.
[35] J. Wang, C. Jiang, H. Zhang, Y. Ren, K. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.
[36] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, 4th Quart., 2019.
[37] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 44–52, Jun. 2019.
[38] C. Jiang and X. Zhu, "Reinforcement learning based capacity management in multi-layer satellite networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4685–4699, Jul. 2020.
[39] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in LTE-U unmanned aerial vehicle (UAV) networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1504–1517, Mar. 2019.
[40] J. Li, Q. Liu, P. Wu, F. Shu, and S. Jin, "Task offloading for UAV-based mobile edge computing via deep reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Beijing, China, 2018, pp. 798–802.
[41] B. Zhang, C. H. Liu, J. Tang, Z. Xu, J. Ma, and W. Wang, "Learning-based energy-efficient data collection by unmanned vehicles in smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1666–1676, Apr. 2018.
[42] Q. Liu, L. Shi, L. Sun, J. Li, M. Ding, and F. Shu, "Path planning for UAV-mounted mobile edge computing with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5723–5728, May 2020.
[43] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.
[44] N. Gao, Z. Qin, X. Jing, Q. Ni, and S. Jin, "Anti-intelligent UAV jamming strategy via deep Q-networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 569–581, Jan. 2020.
[45] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, "Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1046–1061, May 2017.

Zehui Xiong (Member, IEEE) received the B.Eng. degree (with Highest Hons.) from the Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree from Nanyang Technological University, Singapore, where he is currently a Researcher with the Alibaba-NTU Singapore Joint Research Institute. He is the visiting scholar with Princeton University and the University of Waterloo. He has published more than 80 peer-reviewed research papers in leading journals and flagship conferences, and four of them are ESI Highly Cited Papers. His research interests include network economics, wireless communications, blockchain, and edge intelligence. He has won several Best Paper Awards. He is an Editor for *Computer Networks* (Elsevier) and *Physical Communication* (Elsevier), and an Associate Editor for *IET Communications*. He was a recipient of the Chinese Government Award for Outstanding Students Abroad in 2019, and the NTU SCSE Outstanding Ph.D. Thesis Runner-Up Award in 2020.



Yang Zhang (Member, IEEE) received the B.Eng. and B.Eng. (Minor) degrees from Beihang University in 2008 and 2010, respectively, the M.Eng. degrees from the Beijing University of Aeronautics and Astronautics in 2011, and the Ph.D. degree in computer engineering from Nanyang Technological University, Singapore, in 2015. He is currently an Associate Professor with the Wuhan University of Technology, China. He is an Associate Editor of *EURASIP Journal on Wireless Communications and Networking* and the Technical Committee Member of *Computer Communications*.

Wei Yang Bryan Lim received the graduation degree (with Double First Class Hons.) in economics and business administration (finance) from the National University of Singapore in 2018. He is currently an Alibaba Ph.D. candidate with the Alibaba Group and Alibaba-NTU Joint Research Institute, Nanyang Technological University, Singapore. His research interests include federated learning and edge intelligence.

Jiawen Kang received the M.S. and Ph.D. degrees from the Guangdong University of Technology, China, in 2015 and 2018, respectively. He is currently a Postdoctoral with Nanyang Technological University, Singapore. His research interests mainly focus on blockchain, security and privacy protection in wireless communications, and networking.

Dusit Niyato (Fellow, IEEE) received the B.Eng. degree from the King Mongkuts Institute of Technology Ladkrabang, Thailand, in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests are in the area of energy harvesting for wireless communication, Internet of Things, and sensor networks.

Cyril Leung (Life Member, IEEE) received the B.Sc. (First Class Hons.) degree from Imperial College, University of London, U.K., and the M.S. and Ph.D. degrees in electrical engineering from Stanford University. He has been an Assistant Professor with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, and the Department of Systems Engineering and Computing Science, Carleton University. Since 1980, he has been with the Department of Electrical and Computer Engineering, University of British Columbia (UBC), Vancouver, BC, Canada, where he is a Professor and currently holds the PMC-Sierra Professorship in Networking and Communications. He served as an Associate Dean of Research and Graduate Studies with the Faculty of Applied Science, UBC, from 2008 to 2011. His research interests include wireless communication systems, data security and technologies to support ageless aging for the elderly. He is a Member of the Association of Professional Engineers and Geoscientists of British Columbia, Canada.

Chunyan Miao (Senior Member, IEEE) received the B.S. degree from Shandong University, Jinan, China, in 1988, and the M.S. and Ph.D. degrees from Nanyang Technological University, Singapore, in 1998 and 2003, respectively, where she is currently a Professor with the School of Computer Science and Engineering and the Director of the Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY). Her research focus on infusing intelligent agents into interactive new media (virtual, mixed, mobile, and pervasive media) to create novel experiences and dimensions in game design, interactive narrative, and other real world agent systems.