



CISTER

Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Deep Q-Learning based Resource Management in UAV-assisted Wireless Powered IoT Networks

Kai Li

Wei Ni

Eduardo Tovar

and Abbas Jamalipour

CISTER-TR-200201

Deep Q-Learning based Resource Management in UAV-assisted Wireless Powered IoT Networks

Kai Li, Wei Ni, Eduardo Tovar, and Abbas Jamalipour

CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail:

<https://www.cister-labs.pt>

Abstract

In Unmanned Aerial Vehicle (UAV)-assisted Wireless Powered Internet of Things (IoT), the UAV is employed to charge the IoT nodes remotely via Wireless Power Transfer (WPT) and collect their data. A key challenge of resource management for WPT and data collection is preventing battery drainage and buffer overflow of the ground IoT nodes in the presence of highly dynamic airborne channels. In this paper, we consider the resource management problem in practical scenarios, where the UAV has no a-prior information on battery levels and data queue lengths of the nodes. We formulate the resource management of UAV-assisted WPT and data collection as Markov Decision Process (MDP), where the states consist of battery levels and data queue lengths of the IoT nodes, channel qualities, and positions of the UAV. A deep Q-learning based resource management is proposed to minimize the overall data packet loss of the IoT nodes, by optimally deciding the IoT node for data collection and power transfer, and the associated modulation scheme of the IoT node.

Deep Q-Learning based Resource Management in UAV-assisted Wireless Powered IoT Networks

Kai Li*, Wei Ni[†], Eduardo Tovar*, and Abbas Jamalipour[‡]

*CISTER Research Centre, Porto, Portugal.

Email: {kai_li, emt}@isep.ipp.pt.

[†]Commonwealth Scientific and Industrial Research Organization (CSIRO), Sydney, Australia.

Email: wei.ni@csiro.au.

[‡]School of Electrical and Information Engineering, The University of Sydney, Australia.

Email: a.jamalipour@ieee.org.

Abstract—In Unmanned Aerial Vehicle (UAV)-assisted Wireless Powered Internet of Things (IoT), the UAV is employed to charge the IoT nodes remotely via Wireless Power Transfer (WPT) and collect their data. A key challenge of resource management for WPT and data collection is preventing battery drainage and buffer overflow of the ground IoT nodes in the presence of highly dynamic airborne channels. In this paper, we consider the resource management problem in practical scenarios, where the UAV has no a-prior information on battery levels and data queue lengths of the nodes. We formulate the resource management of UAV-assisted WPT and data collection as Markov Decision Process (MDP), where the states consist of battery levels and data queue lengths of the IoT nodes, channel qualities, and positions of the UAV. A deep Q-learning based resource management is proposed to minimize the overall data packet loss of the IoT nodes, by optimally deciding the IoT node for data collection and power transfer, and the associated modulation scheme of the IoT node.

Index Terms—Unmanned Aerial Vehicle, Internet of Things, wireless power transfer, resource management, deep Q-learning

I. INTRODUCTION

Recent advances in scalable Internet of Things (IoT) and Wireless Power Transfer (WPT) are developed for sustainable sensing of urban weather, environmental pollutions, or traffic and road conditions in smart cities [1]–[3]. A large number of wireless powered IoT nodes are deployed in the network, and the IoT nodes with random data arrivals buffer the data to be collected in the data queue [4]. Moreover, Unmanned Aerial Vehicles (UAVs) are employed to collect data of the IoT nodes in the area of interest, thanks to UAVs' excellent mobility and maneuverability [5], as shown in Figure 1. The UAV moves sufficiently close to each IoT node, exploiting short-distance line-of-sight (LoS) communication links, for recharging batteries remotely while collecting their data [6].

WPT and data transmission could be severely affected by time-varying channels due to movements of the UAV, while the UAV has no up-to-date knowledge about battery levels and data queue lengths of the IoT nodes. Therefore, the resource management of WPT and data collection for preventing battery drainage and buffer overflow is critical,

given highly dynamic airborne channels in UAV-assisted wireless powered IoT networks.

In this paper, we first formulate the resource management of UAV-assisted WPT and data collection as a Markov Decision Process (MDP). Each MDP state contains battery levels and queue lengths of the IoT node, channel qualities, and waypoints of the UAV along the flight trajectory. Then, a new deep Q-learning based resource management (DQL-RM) scheme is developed, which derives the optimal resource management strategy based on network state, actions of the UAV and a corresponding Q value. DQL-RM learns the optimal Q value asymptotically through training a deep Q-network on the UAV, where the selection of the IoT node and the associated modulation scheme of the selected IoT node are jointly optimized based on the Q values.

This paper is structured as follows. Related work on resource management in wireless powered sensor networks is presented in Section II. Network model and communication protocol design are investigated in Section III. DQL-RM is proposed in Section IV to address the resource management problem. In Section V, we show numerical results and performance evaluation. This paper is concluded in Section VI.

II. LITERATURE REVIEW

A UAV-assisted wireless powered communication network is considered in [7], where the ground nodes are charged by a UAV with constant power supply and the UAV collects data from the nodes by time-division multiple access. The UAV position and time allocation are studied to improve the transmission rate for the ground nodes. In [8], UAVs are utilized in mobile edge computing architectures, where the ground nodes offload some of their computation tasks to the UAV. The ground nodes can harvest energy from the UAV by using WPT that is integrated into the mobile edge computing architecture. Transmit power of the ground nodes, task offloading time, and the UAV trajectory are scheduled to enhance the computation capability of the ground nodes. A UAV-assisted WPT system is studied in [9], where the UAV is dispatched to charge two ground nodes. The resource allocation algorithms are developed to

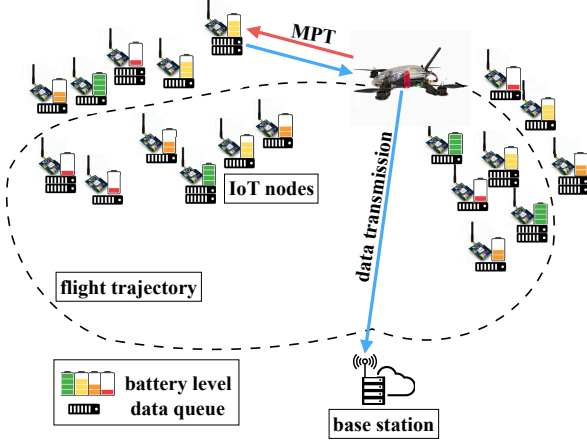


Fig. 1: WPT and data collection in UAV-assisted IoT with wireless powered ground nodes.

extend the WPT efficiency by adjusting the trajectory, flight altitude, and transmit beamwidth of the UAV. In [10], the techniques for extending the UAV's flight range and mission duration are reviewed. The UAV is equipped with a solar cell array capable of receiving energy from an infrared laser via a laser beaming. The ground charging stations are deployed to provide energy to the UAV via WPT. An energy-efficient cooperative UAV relaying scheme is developed in [11], which guarantees the successful transmission rate of the UAV. The schedule of the data relay is formulated to reduce energy consumption of the UAVs under guaranteed bit error rates. In [12], the authors focus on energy efficiency of the legitimate surveillance in UAV communications. The resource allocation of the legitimate UAV is formulated to eavesdrop the communication of the suspicious UAVs with uncertain channel dynamics. However, most of the existing literature only study the energy efficiency of the UAV without considering the data loss of the ground nodes due to buffer overflows and poor channels.

Scheduling strategies in [13]–[15] are studied to reduce packet loss of wireless powered static sensors according to their battery and buffer status. Due to the low-dimensional channel and sensor state spaces, the resource management problem can be solved by reinforcement learning or dynamic programming. However, in the UAV-assisted WPT and data collection, the mobility of the UAV with the varying patrolling velocity causes rapidly changing wireless channels, which leads to the exceedingly large state space and action space. This prevents conventional resource management approaches, such as [13], [14] and [15], from scaling to the high-dimensional input spaces.

III. NETWORK MODEL

This section studies the network model of UAV-assisted wireless powered IoT networks. The communication protocol is also designed for the UAV and ground IoT nodes.

A. Network model

Let I denote the total number of wireless powered IoT nodes on the ground. The UAV that acts as a data collection node flies a predetermined trajectory for Z laps. The flight waypoint of the UAV at t in lap z is denoted by $\zeta_z(t)$. The UAV uses WPT to remotely charge the IoT nodes. The IoT node i ($i \in [1, I]$) harvests energy from the UAV to power its operations, e.g., sensing, computing and communication. The rechargeable battery of the IoT node is finite with the capacity of E Joules, and the battery overflows if overcharged.

We denote $\Gamma(\cdot)$ as the Gamma function [16]. The required BER of IoT node's data transmission is denoted by ε . According to the Nakagami- m channel model [17], we can have

$$\varepsilon \approx \frac{0.2}{\Gamma(m)} \left(\frac{m}{\gamma_i} \right)^m \left[\frac{\Gamma(m, b_{\phi_i^z(t)} \gamma_i (\phi_i^z(t)))}{(b_{\phi_i^z(t)})^m} - \frac{\Gamma(m, b_{\phi_i^z(t)} \gamma_i (\phi_i^z(t) + 1))}{(b_{\phi_i^z(t)})^m} \right], \quad (1)$$

$$b_{\phi_i^z(t)} = \frac{m}{\gamma_i} + \frac{3}{2(2^{\phi_i^z(t)} - 1)}, \quad (2)$$

where $\gamma_i(\phi_i^z(t))$ is the SNR between node i and the UAV using $\phi_i^z(t)$, and the average SNR of the channel is $\bar{\gamma}_i$. $\phi_i^z(t)$ denotes the modulation scheme of node i at t in lap z . Moreover, we have $\gamma_i(\phi_i^z(t)) = \frac{\|\mathbf{h}_i^z(t)\|^2 P_i^z(t)}{\sigma_0^2}$, where σ_0^2 is noise power of the channel, and $P_i^z(t)$ is the transmit power of i [18]. $\mathbf{h}_i^z(t)$ is the complex channel coefficient, which can be known by channel reciprocity.

For illustration convenience, we consider $m = 1$ in (1) as an example in this paper [19]. Note that other Nakagami fading channel model with any m values can also be applied to the proposed DQL-RM scheme. Given a specific ε , the transmit power of the IoT node can be given by [20]

$$P_i^z(t) \approx \frac{\kappa_2^{-1} \ln \frac{\kappa_1}{\varepsilon}}{\|\mathbf{h}_i^z(t)\|^2} (2^{\phi_i^z(t)} - 1), \quad (3)$$

where κ_1 and κ_2 denote the two channel constants.

The distance between the UAV and node i at t in lap z is $d_i^z(t)$. The WPT transceiver alignment between the UAV and the IoT node is $\theta_i^z(t)$. Based on [21], [22], we have the WPT efficiency factor, which is $\omega(d_i^z(t), \theta_i^z(t))$. The power transferred from the UAV to the IoT node via WPT can be given by

$$\tilde{P}_i^z(t) = \omega(d_i^z(t), \theta_i^z(t)) P_{\text{UAV}}^{tx} \|\mathbf{h}_i^z(t)\|^2, \quad (4)$$

where the transmit power of WPT at the UAV is P_{UAV}^{tx} , and $\|\cdot\|$ denotes norm.

B. Data collection protocol

Figure 2 presents the data collection protocol for the UAV-assisted wireless powered IoT network. Specifically, DQL-RM (will be illustrated in Section IV) is carried out on the UAV to take the action of scheduling one of the IoT nodes for WPT and data transmission at each time

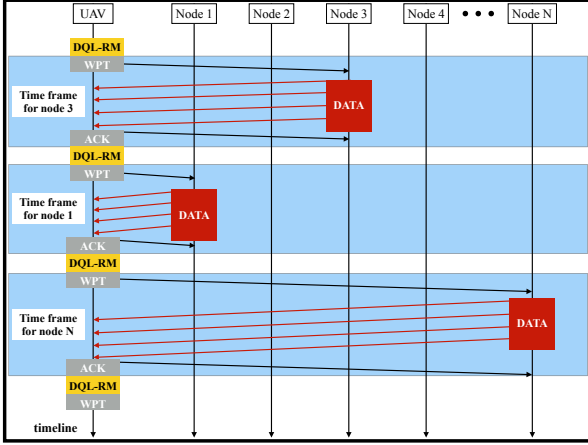


Fig. 2: Data communication protocol for the UAV-assisted wireless powered IoT network.

slot while allocating the modulation scheme (or transmit power) of the IoT node. Then, the UAV transfers power to the selected IoT node via WPT, followed by the data transmission from the IoT node to the UAV. In particular, a control segment of the IoT node's data packet contains $q_{i,z}(t)$ and $e_{i,z}(t)$. Additionally, the updated state information of the selected IoT node is known by the UAV based on the received data.

In each communication frame, only one IoT node can be scheduled to transmit data in case of packet collision. The UAV can process the received data packets online, and respond to the IoT node's requests by using ACK packets. Meanwhile, DQL-RM is conducted with the updated channel state information to schedule the other IoT node to transmit data in the next communication frame.

IV. DQL-RM FOR THE UAV

In this section, we first present MDP formulation of the resource management of UAV-assisted WPT and data collection. Then, DQL-RM is proposed to optimize the node selection for WPT and data collection, while the modulation scheme of the selected IoT node is optimally allocated to maximize the harvested energy.

A. Markov Decision Process of the resource management

We formulate the resource management in UAV-assisted wireless powered IoT networks as a discrete-time MDP. Each MDP state is composed of battery levels and queue lengths of the IoT node, channel qualities, and waypoints of the UAV along the flight trajectory. Thus, we have $\mathcal{S}_\alpha = \{(q_{i,z}(t), e_{i,z}(t), \mathbf{h}_i^z(t), \zeta_z(t)), i = 1, 2, \dots, I; z = 1, 2, \dots, Z\}$, where $q_{i,z}(t)$, $e_{i,z}(t)$, $\mathbf{h}_i^z(t)$, and $\zeta_z(t)$ denote the battery level, queue length, channel quality, and the location of the UAV at time t in lap z , respectively.

In the proposed MDP formulation, action $k \in \mathcal{A}$ to be taken is to schedule one IoT node to harvest energy from the UAV and transmit data, while allocating the

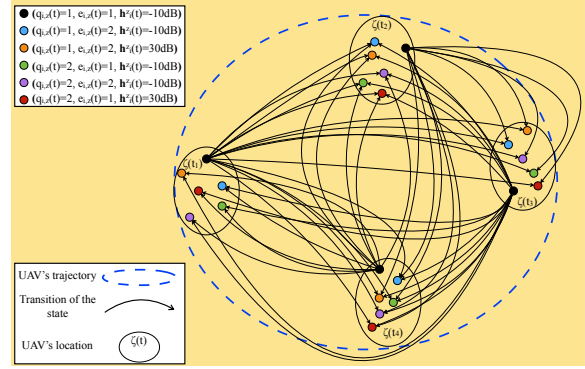


Fig. 3: An example of a transition diagram of 24 MDP states, which has 4 waypoints on the UAV's trajectory.

modulation of the selected IoT node, i.e., $\mathcal{A} \in \{(i, \phi_i^z(t)) :$

$i = 1, 2, \dots, I; z = 1, 2, \dots, Z; \phi_i^z(t) \in \{1, 2, \dots, \Phi\}\}$. The actions of the UAV are optimized to minimize packet loss from queue overflows and unsuccessful data transmissions of the IoT nodes. Note that the optimization needs to be achieved in the long term over the entire stochastic control process (rather than myopically at an individual time slot).

To illustrate the proposed MDP model, Figure 3 presents an example of transition diagram with 24 MDP states in one lap of the UAV's flight. We set $I = 1$, $Z = 1$, $K = 1$, $D = 1$, $H = 2$ (e.g., -10dB , 30dB), and $V = 4$, where H and V denote the total channel states and waypoints of the UAV, respectively. The vertices stand for all possible states in MDP, i.e., $\{(q_{i,z}(t), e_{i,z}(t), \mathbf{h}_i^z(t), \zeta_z(t))\}$. The edges show the transition from each state to other states according to $\Pr\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\}$. The state transition depends on the change of $\{(q_{i,z}(t), e_{i,z}(t), \mathbf{h}_i^z(t))\}$ of the IoT node and $\zeta_z(t)$ along the trajectory of the UAV. In other words, the next state of $\{(q_{i,z}(t_1), e_{i,z}(t_1), \mathbf{h}_i^z(t_1), \zeta_z(t_1))\}$ can be one of the states at $\zeta_z(t_2)$, $\zeta_z(t_3)$, or $\zeta_z(t_4)$. For example, for t_1 , the next state of $\{(q_{i,z}(t_1) = 1, e_{i,z}(t_1) = 1, \mathbf{h}_i^z(t_1) = -10\text{dB}, \zeta_z(t_1))\}$ can be $\{(q_{i,z}(t_2) = 2, e_{i,z}(t_2) = 2, \mathbf{h}_i^z(t_2) = -10\text{dB}, \zeta_z(t_2))\}$, if IoT node i is selected, but the data collection is not successful; or $\{(q_{i,z}(t_2) = 1, e_{i,z}(t_2) = 2, \mathbf{h}_i^z(t_2) = -10\text{dB}, \zeta_z(t_2))\}$, if the data collection is successful. Note that Figure 3 gives a small-scale example of the transition of one of the states, i.e., $\{(q_{i,z}(t) = 1, e_{i,z}(t) = 1, \mathbf{h}_i^z(t) = -10\text{dB}, \zeta_z(t))\}$.

Furthermore, we define the expected cost after \mathcal{S}_α is observed while action k is conducted as $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\}$. We have

$$Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\} = (1 - \varrho)Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\} + \varrho \left[C\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\} + \delta \min_{k' \in \mathcal{A}} Q\{\mathcal{S}_{\beta'} | \mathcal{S}_\beta, k'\} \right]. \quad (5)$$

where $\varrho, \delta \in (0, 1]$ (two small positive fractions) indicate learning rate and discount factor, respectively.

Classical approaches in MDP, e.g., value iteration, can be

applied to solve the optimal policy, by assuming that the UAV has the a-prior knowledge on the transition probability and the cost of all MDP states. According to the Bellman optimality equation with the value iteration method, the estimate of the optimal action-value function is repeatedly updated. When the Bellman optimality equation converges, the cost function is minimized. The MDP model can be stabilized. However, the transition probability and the cost of the states have to be known in prior, hence the action-value function of MDP can only be evaluated offline. In contrast, this paper considers a practical scenario where the UAV has no a-priori knowledge on the transition probability and the cost of the states. The proposed DQL-RM scheme is designed to learn the transition probabilities and the costs, while evaluating the action-value function online. The proposed DQL-RM scheme converges after several episodes when the learning results and action-value function stabilize.

B. Optimizing $\phi_i^z(t)$

To minimize the packet loss stemming from insufficient energy, $\phi_i^z(t)$ of the selected node i is to be chosen to maximize the energy harvested during a contact time with the UAV, with a length of $\hat{T}_i^z(t)$. The optimal modulation of the IoT node, $\phi_i^z(t)^*$, is independent of the battery level and the queue length. This is because $\phi_i^z(t)^*$ is selected to maximize the increase of the battery level at node i , under the bit error rate requirement ϵ_i for the packet transmitted. As a result, $\phi_i^z(t)$ can be decoupled from \mathcal{A} , and optimized in prior by [23]

$$\phi_i^z(t) = \arg \max_{\phi=1, \dots, \Phi} \left\{ (\hat{T}_i^z(t) - \frac{B}{\phi W}) P_i^z(t) - \frac{B}{\phi W} P_i^z(\phi) \right\}, \quad (6)$$

the right-hand side (RHS) of which, by substituting (4) and (3), can be rewritten as

$$\max_{\phi=1, \dots, \Phi} \left\{ (\hat{T}_i^z(t) - \frac{B}{\phi W}) \omega(d_i^z(t), \theta_i^z(t)) P_{\text{UAV}}^{tx} \|\mathbf{h}_i^z(t)\|^2 - \frac{B \kappa_2^{-1} \ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i^z(t)\|^2 \phi W} (2^\phi - 1) \right\}, \quad (7)$$

where W is the bandwidth of the uplink data transmission, $\frac{1}{W}$ is the duration of an uplink symbol, and $\frac{B}{\phi W}$ is the duration of uplink data transmission. $(\hat{T}_i^z(t) - \frac{B}{\phi W})$ is the rest of the time slots used for downlink WPT, and $\hat{T}_i^z(t)$ is the contact time between the IoT node and the UAV in the time slot, which is affected by the patrolling velocity of the UAV $v^z(t)$. Thus, we have

$$\hat{T}_i^z(t) = \frac{2\sqrt{(d_i^z(t))^2 - (b^z)^2}}{v^z(t)}, \quad (8)$$

where b^z is the altitude of the UAV at lap z . We assume that the UAV maintains the same altitude and the same heading in each lap.

By using the first-order necessary condition of the optimal solution, we have

$$\frac{d}{d\phi} \left((\hat{T}_i^z(t) - \frac{B}{\phi W}) \omega(d_i^z(t), \theta_i^z(t)) P_{\text{UAV}}^{tx} \|\mathbf{h}_i^z(t)\|^2 - \frac{B \kappa_2^{-1} \ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i^z(t)\|^2 \phi W} (2^\phi - 1) \right) = 0, \quad (9)$$

$$\phi^{-2} \frac{B}{W} \omega(d_i^z(t), \theta_i^z(t)) P_{\text{UAV}}^{tx} \|\mathbf{h}_i^z(t)\|^2 - \frac{B \kappa_2^{-1} \ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i^z(t)\|^2 W} \cdot (\phi^{-1} 2^\phi \ln 2 - \phi^{-2} 2^\phi) - \frac{B \kappa_2^{-1} \ln(\frac{\kappa_1}{\epsilon})}{\|\mathbf{h}_i^z(t)\|^2 W} \phi^{-2} = 0. \quad (10)$$

The ϕ values are then given as follows:

$$\phi 2^\phi \ln 2 - 2^\phi = \frac{B}{W} \omega(d_i^z(t), \theta_i^z(t)) P_{\text{UAV}}^{tx} \|\mathbf{h}_i^z(t)\|^2 \frac{\|\mathbf{h}_i^z(t)\|^2 W}{B \kappa_2^{-1} \ln(\frac{\kappa_1}{\epsilon})} - 1. \quad (11)$$

Since the left-hand side (LHS) of (11) monotonically increases with ϕ , the optimal value ϕ^* can be obtained by applying a bisection search method, and evaluating the two closest integers about the fixed point of the bisection method [24]. Specifically, $\phi_- = 1$ and $\phi_+ = \Phi$ are initialized. Each iteration of the bisection method contains 4 steps applied over the range of $\phi = [1, \Phi]$, as follows.

- The midpoint of the modulation interval $[\phi_-; \phi_+]$ is calculated, which gives $\phi_{\text{mid}} = \frac{\phi_- + \phi_+}{2}$.
- Substitute ϕ_{mid} into (11) to obtain the function value $f(\phi_{\text{mid}})$.
- If the convergence is attained (that is, the modulation interval or $|f(\phi_{\text{mid}})|$ cannot be further reduced), return ϕ_{mid} and stop the iteration.
- Replace either $(\phi_-, f(\phi_-))$ or $(\phi_+, f(\phi_+))$ with $(\phi_{\text{mid}}, f(\phi_{\text{mid}}))$.

C. Deep Q-network on the UAV

The proposed DQL-RM builds a deep Q-network on the UAV to optimize the resource management by approximating the optimal Q value. Figure 4 depicts the proposed deep Q-network, where the Q value in (5) is derived according to the network state and the action of the UAV. The learning weight values φ_l in the deep Q-network are iteratively adjusted to approximate $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k; \varphi_l\}$, where $l \leq \Omega$ and Ω is the total number of iterations. The approximated $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k; \varphi_l\}$, which is the outputs of the deep Q-network, can be minimized by optimizing the weights φ_l .

The weight φ_l at iteration l is adjusted for training the deep Q-network, while minimizing $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k; \varphi_l\}$. At each iteration of minimizing $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k; \varphi_l\}$, the weight φ_{l-1} from iteration $(l-1)$ is fixed. Thus, the subproblem of learning $Q\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k; \varphi_l\}$ at iteration l ($l \leq \Omega$) defines $C\{\mathcal{S}_\beta | \mathcal{S}_\alpha, k\} + \delta \min_{k' \in \mathcal{A}} Q\{\mathcal{S}_{\beta'} | \mathcal{S}_\beta, k'; \varphi_{l-1}\}$. For deriving the weight φ_l at iteration l , gradient descent

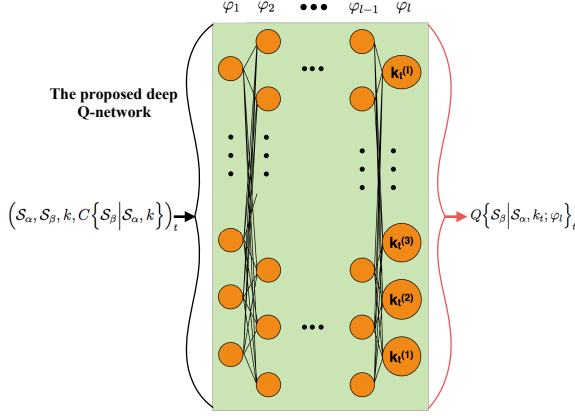


Fig. 4: Schematic illustration of the deep Q-network with DQL-RM on the UAV.

is applied to iteratively compute the gradient value, and update the neural network's weights to reach the global minimum (refer to [25] for details).

The proposed DQL-RM scheme optimizes the actions based on the deep Q-network to solve the resource management problem. The proposed deep Q-network maintains two separate Q-networks $Q\{S_\beta|S_\alpha, k; \varphi_l\}$ and $Q\{S_\beta|S_\alpha, k; \varphi_{l-1}\}$ with the weights φ_l at iteration l and the weights φ_{l-1} at iteration $l-1$, respectively. DQL-RM updates φ_l with multiple times per time-step, and φ_l is copied into φ_{l-1} . DQL-RM trains the deep Q-network to minimize a set of loss functions at every update iteration [26], hence, minimizing the mean-squared Bellman error. Therefore, the optimality can be asymptotically achieved by DQL-RM. For maximizing the harvested energy, DQL-RM also determines the optimal modulation scheme $\phi_i^z(t)$ of the IoT node once the optimal IoT node is selected from the deep Q-network.

V. PERFORMANCE EVALUATION

DQL-RM is implemented in Python 3.5 based on Google TensorFlow, and we assess the performance when the number of IoT nodes enlarges from 50 to 200. Figure 5 shows the network cost (i.e., data packet loss) at each episode, given $I = 180$. In particular, "episodes" are a sequence of training epochs, where the deep Q-network is trained to find the optimal actions. According to Figure 4, the proposed deep Q-network executes actions, obtains the next states, and updates the learning weights at each episode. As observed in Figure 5, the packet loss (i.e., network cost) with DQL-RM drops around 58.3% at the first 50 episodes. From episode 50 to episode 500, the packet loss drops from 2×10^4 to about 140. Moreover, the performance reaches a relatively stable value after episode 400, which confirms the convergence of the proposed deep Q-network.

We compared DQL-RM with two resource management scheme based on either randomized MDP states (RRM),

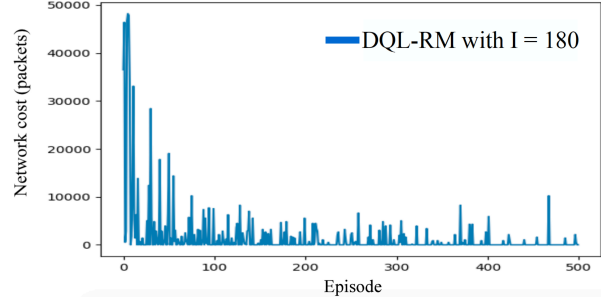


Fig. 5: Network cost at the episode given $I = 180$.

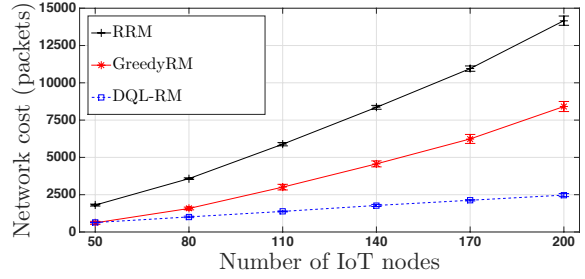


Fig. 6: Network cost with regards to different number of nodes.

or greedy policies (GreedyRM). RRM randomly schedules one IoT node at each time slot to transfer power and collect data, while GreedyRM gives the highest priority of WPT and data collection to the IoT node with the longest buffer.

Figure 6 shows the packet loss of DQL-RM, where the number of IoT nodes increases from 50 to 200. The data queue length of the IoT node is set to 10. Generally, the network cost of the proposed DQL-RM is much lower than RRM and GreedyRM. When the number of IoT nodes is 200, DQL-RM achieves 82.8% and 69.2% lower network cost than RRM and GreedyRM, respectively. The performance gains keep growing with the number of IoT nodes in the network. This is because DQL-RM schedules WPT and data communications to minimize the data packet loss of the entire network, by learning the IoT nodes' battery levels and queue lengths.

We define the packet loss rate as the ratio of the packet loss and the total number of data packets. Figure 7 studies the packet loss rate with regards to different number of IoT nodes. When $I = 50$, the packet loss rate of DQL-RM is similar to GreedyRM. When the number of IoT nodes increases from 80 to 200, DQL-RM achieves lower packet loss rate than RRM and GreedyRM. Moreover, when the number of IoT nodes increases from 50 to 200, the packet loss rate of DQL-RM only slightly grows about 2%. This indicates that the performance of DQL-RM is not effected by the number of IoT nodes in the network. The reason is that DQL-RM efficiently adapts the IoT node selection and WPT duration to minimize the data packet loss in the presence of the channel dynamics.

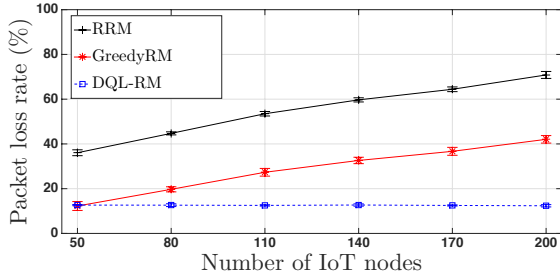


Fig. 7: Packet loss rate with regards to different number of nodes.

VI. CONCLUSION AND FUTURE WORK

This paper investigates the resource management of UAV-assisted WPT and data collection for preventing battery drainage and buffer overflow of the ground IoT nodes in the presence of highly dynamic airborne channels. DQL-RM is proposed to minimize the overall data packet loss of the IoT nodes, by jointly optimizing the IoT node for WPT and data collection, and the associated modulation scheme of the IoT node. DQL-RM builds and trains a deep Q-network to determine the optimal actions of the UAV with the MDP states of battery levels and data queue lengths of the IoT nodes, channel conditions, and the waypoints given the trajectory of the UAV.

For future work, the IoT networks will consider heterogeneous ground nodes with dynamic battery capacity and data queue size. The proposed DQL-RM will be further evaluated in multiple application scenarios, e.g., intelligent transportation and 5G networks.

ACKNOWLEDGEMENTS

This work was partially supported by National Funds through FCT/MCTES (Portuguese Foundation for Science and Technology), within the CISTER Research Unit (CEC/04234); also by the Operational Competitiveness Programme and Internationalization (COMPETE 2020) through the European Regional Development Fund (ERDF) and by national funds through the FCT, within project POCI-01-0145-FEDER-029074 (ARNET).

REFERENCES

- [1] W. Ejaz, M. Naeem, A. Shahid, A. Anpalagan, and M. Jo, "Efficient energy management for the internet of things in smart cities," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 84–91, 2017.
- [2] B. P. L. Lau, T. Chaturvedi, B. K. K. Ng, K. Li, M. S. Hasala, and C. Yuen, "Spatial and temporal analysis of urban space utilization with renewable wireless sensor network," in *IEEE/ACM International Conference on Big Data Computing Applications and Technologies (BDCAT)*. IEEE, 2016, pp. 133–142.
- [3] P. Ramezani and A. Jamalipour, "Toward the evolution of wireless powered communication networks for the future internet of things," *IEEE Network*, vol. 31, no. 6, pp. 62–69, 2017.
- [4] P. Ramezani and A. Jamalipour, "Throughput maximization in dual-hop wireless powered communication networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 10, pp. 9304–9312, 2017.
- [5] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile Internet of Things: Can UAVs provide an energy-efficient mobile architecture?" in *GLOBECOM*. IEEE, 2016, pp. 1–6.
- [6] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2241–2263, 2018.
- [7] M. Jiang, Y. Li, Q. Zhang, and J. Qin, "Joint position and time allocation optimization of UAV enabled time allocation optimization networks," *IEEE Transactions on Communications*, vol. 67, no. 5, pp. 3806–3816, 2019.
- [8] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in UAV-enabled wireless-powered mobile-edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 1927–1941, 2018.
- [9] Y. Wu, L. Qiu, and J. Xu, "UAV-enabled wireless power transfer with directional antenna: A two-user case," in *International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, 2018, pp. 1–6.
- [10] M. Lu, M. Bagheri, A. P. James, and T. Phung, "Wireless charging techniques for UAVs: A review, reconceptualization, and extension," *IEEE Access*, vol. 6, pp. 29 865–29 884, 2018.
- [11] K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "EPLA: Energy-balancing packets scheduling for airborne relaying networks," in *International Conference on Communications (ICC)*. IEEE, 2015, pp. 6246–6251.
- [12] X. Wang, K. Li, S. S. Kanhere, D. Li, X. Zhang, and E. Tovar, "PELE: Power efficient legitimate eavesdropping via jamming in UAV communications," in *International Wireless Communications and Mobile Computing Conference (IWCMC)*. IEEE, 2017, pp. 402–408.
- [13] K. Li, W. Ni, M. Abolhasan, and E. Tovar, "Reinforcement learning for scheduling wireless powered sensor communications," *IEEE Transactions on Green Communications and Networking*, 2018.
- [14] K. Li, W. Ni, L. Duan, M. Abolhasan, and J. Niu, "Wireless power transfer and data collection in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 3, pp. 2686–2697, 2018.
- [15] K. Li, W. Ni, L. Duan, M. Abolhasan, and J. Niu, "SWPT: A joint-scheduling model for wireless powered sensor networks," in *IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2017, pp. 1–6.
- [16] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.
- [17] M.-S. Alouini and A. J. Goldsmith, "Adaptive modulation over nakagami fading channels," *Wireless Personal Communications*, vol. 13, no. 1-2, pp. 119–143, 2000.
- [18] K. Li, W. Ni, E. Tovar, and M. Guizani, "Optimal rate-adaptive data dissemination in vehicular platoons," *IEEE Transactions on Intelligent Transportation Systems*, 2019.
- [19] K. Li, N. Ahmed, S. S. Kanhere, and S. Jha, "Reliable transmissions in AWSNs by using O-BESPAR hybrid antenna," *Pervasive and Mobile Computing*, vol. 30, pp. 151–165, 2016.
- [20] K. Li, W. Ni, X. Wang, R. P. Liu, S. S. Kanhere, and S. Jha, "Energy-efficient cooperative relaying for unmanned aerial vehicles," *IEEE Transactions on Mobile Computing*, no. 6, pp. 1377–1386, 2016.
- [21] K. Li, C. Yuen, and S. Jha, "Fair scheduling for energy harvesting WSN in smart city," in *SenSys*. ACM, 2015, pp. 419–420.
- [22] S. He, J. Chen, F. Jiang, D. K. Yau, G. Xing, and Y. Sun, "Energy provisioning in wireless rechargeable sensor networks," *IEEE Transactions on Mobile Computing*, vol. 12, no. 10, pp. 1931–1942, 2012.
- [23] K. Li, R. C. Voicu, S. S. Kanhere, W. Ni, and E. Tovar, "Energy efficient legitimate wireless surveillance of UAV communications," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2283–2293, 2019.
- [24] D. Estep, "The bisection algorithm," *Practical Analysis in One Variable*, pp. 165–177, 2002.
- [25] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep Q-network for UAV-assisted online power transfer and data collection," *IEEE Transactions on Vehicular Technology*, 2019.
- [26] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications," in *International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.