Reconfigurable intelligent surface and UAV-assisted communications
A deep reinforcement learning approach

Nguyen, Khoi Khac

Award date:
2022

Awarding institution:
Queen's University Belfast

Link to publication

Reconfigurable Intelligent Surface and UAV-assisted Communications: A Deep Reinforcement Learning Approach

Khoi Khac Nguyen

School of Electronics, Electrical Engineering and Computer Science

Queen's University Belfast

A thesis submitted for the degree of

Doctor of Philosophy

April 4, 2022

Abstract

Unmanned aerial vehicles (UAVs) and reconfigurable intelligent surface (RIS) have been considered as promising techniques for enhancing network performance and coverage in wireless communication. The UAV-assisted wireless networks are reliable, low-cost, and on-demand by using the agility and mobile features of the UAVs. The UAVs can provide the maximum coverage and capacity for the targeted ground users by adjusting their altitude. Their nimble mobility feature helps them avoid signal blockages and have better connections with the ground users. However, due to the limitation of their on-board power and flight time, it is challenging to obtain an optimal resource allocation scheme for the UAV-assisted Internet of Things (IoT). The RISs reflect the signal from the transmitters to the receivers by controlling the phase-shift value of a massive amount of scattering reflectors. The reflected signals can be combined coherently to improve the received signal or destructively to suppress the interference. In addition, the reliability and zero-delay are also notable advantages of the RIS in supporting reliable and low-cost wireless communications.

Many of the devices used in IoT applications are energy-limited, and thus supplying energy while maintaining seamless connectivity for IoT devices is of considerable importance. In this context, we propose

a simultaneous wireless power transfer and information transmission scheme for IoT devices with the support from RIS-aided UAV communications. In particular, IoT devices harvest energy from the UAV through wireless power transfer; and then, the UAV collects data from the IoT devices through information transmission. To characterise the agility of the UAV, we consider two scenarios: a hovering UAV and a mobile UAV. Aiming at maximising the total network sum-rate, we jointly optimise the trajectory of the UAV, the energy harvesting scheduling of IoT devices, and the phase-shift matrix of the RIS.

We also investigate RIS-assisted multi-UAV networks that can utilise both advantages of UAVs' agility and RIS's reflection for enhancing the network's performance. Aiming at maximising the energy efficiency (EE) of the considered networks, we jointly optimise the power allocation of the UAVs and the phase-shift matrix of the RIS.

This thesis presents three major contributions. Firstly, we design a new UAV-assisted IoT system relying on the shortest flight path of the UAVs while maximising the amount of data collected from IoT devices. Then, a deep reinforcement learning (DRL)-based technique is conceived for finding the optimal trajectory and throughput in a specific coverage area. After training, the UAV has the ability to autonomously collect all the data from user nodes at a significant total sum-rate improvement while minimising the associated resources used. Our proposed techniques strike a balance between the achieved throughput, trajectory, and the time spent. Secondly, we formulate a Markov decision process and propose two DRL algorithms to solve the optimisation problem of maximising the total network sum-rate in the RIS-assisted UAV communications. Given the strict requirements of the RIS and UAV, the significant improvement in processing time and throughput performance demonstrates that our proposed scheme is well applicable for practical IoT applications. Thirdly, a DRL approach is proposed for solving the UAV's power allocation and the RIS's phase shift optimisation problem in the RIS-assisted multi-UAVs communications. The centralised fashion and the parallel learning approach are also proposed for maximising the EE performance. Our proposed DRL methods for RIS-assisted UAV networks can be used for real-time applications thanks to their capability of instant decision-making and handling the time-varying channel with the dynamic environmental setting.

As a result, this thesis proposes novel methods based on the DRL algorithms for maximising the EE, sum-rate in RIS and UAV-aided communications. We transform a real-life problem into a digital form to formulate the environment and define the agents that interact with the environment to improve the network's performance.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Trung Q. Duong, for giving me the opportunity to pursue my Ph.D. research at Queen's University Belfast. I would not have been able to get my work done without his continuous support, patience, motivation and immense guidance throughout my Ph.D study at Queen's University Belfast. I am thankful for your priceless advice, supervision and giving me freedom to pursue diverse research directions which has made my Ph.D journey unforgettable.

I am immensely thankful to my friends and the Vietnamese folks at Belfast for their help and support, especially in a tough time during the Covid-19 crisis.

Last but not least, words cannot express how grateful I am to have my family and my girlfriend there for their unconditional love and encouragement.

Khoi Khac Nguyen
Belfast, January 2022

To my family and my girlfriend!

Author's publications

A. Accepted and Published:

1. **K. K. Nguyen**, S. Khosravirad, D. B. da Costa L. D. Nguyen, and T. Q. Duong, "Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning," *IEEE J. Selected Topics in Signal Process.*, 2021 (accepted).
2. **K. K. Nguyen**, T. Q. Duong, T. Do-Duy, H. Claussen, and L. Hanzo, "3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning," *IEEE Trans. Commun.*, 2021 (accepted).
3. **K. K. Nguyen**, N. A. Vien, L. D. Nguyen, M.-T. Le, L. Hanzo, and T. Q. Duong, "Real-time Energy Harvesting Aided Scheduling in UAV-assisted D2D Networks Relying on Deep Reinforcement Learning," *IEEE Access*, vol. 9, pp. 3638–3648, 2021.
4. **K. K. Nguyen**, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and L. D. Nguyen, "Distributed Deep Deterministic Policy Gradient for Power Allocation Control in D2D-based V2V Communications," *IEEE Access*, vol. 7, pp. 164 533–164 543, Nov. 2019.

5. **K. K. Nguyen**, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and N. M. Nguyen, “Non-cooperative Energy-Efficient Power Allocation Game in D2D Communication: A Multi-Agent Deep Reinforcement Learning Approach,” *IEEE Access*, vol. 7, pp. 100 480–100 490, Jul. 2019.

B. Under Review:

1. **K. K. Nguyen**, A. Masaracchia, C. Yin, L. D. Nguyen, O. A. Dobre, and T. Q. Duong, “Deep Reinforcement Learning for Intelligent Reflecting Surface-assisted D2D Communications,” *IEEE Trans. Veh. Technol.*, 2021 (major revision).
2. **K. K. Nguyen**, A. Masaracchia, Vishal Sharma, H. V. Poor, and T. Q. Duong, “RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning,” *IEEE J. Selected Topics in Signal Process.*, 2021 (under review).

C. Co-author:

1. A. Masaracchia, Y. Li, **K. K. Nguyen**, C. Yin, S. R. Khosravirad, D. B. Da Costa, and T. Q. Duong,“UAV-Enabled Ultra-Reliable Low-Latency Communications for 6G: A Comprehensive Survey,” *IEEE Access*, vol. 9, pp. 137338-137352, Oct. 2021.
2. L. D. Nguyen, **K. K. Nguyen**, A. Kortun and T. Q. Duong, “Real-Time Deployment and Resource Allocation for Distributed UAV Systems in Disaster Relief,” in *Proc. IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Cannes, France, Jul. 2019, pp. 1-5.

Table of Contents

Author's publications **vi**

Table of Contents **viii**

List of Tables **xiii**

List of Figures **xiv**

List of Notations **xvii**

1 Introduction and Overview **1**

1.1 UAV-enabled wireless communications 2

1.1.1 Applications of UAV Communications 3

1.1.2 Challenges in UAV-assisted Communications 4

1.2 Reconfigurable Intelligent Surfaces-aided Wireless Networks . . . 5

1.2.1 Benefits and Applications of RIS-assisted Communications 5

1.2.2 Challenges in RIS-assisted Communications 7

1.3 Deep Reinforcement Learning in Wireless Networks 8

1.3.1 An Overview of Markov Decision Processes 9

1.3.2 A Brief Overview of Deep Reinforcement Learning 9

1.3.2.1 Value Function 10

1.3.2.2 Policy Search 11

1.4 Motivation, Contributions and Organisation 12

1.4.1 Research Motivation 13

1.4.2 Summary of Contributions 14

1.4.3 Outlines of the thesis 18

2 Literature Review 20

2.1 Trajectory Design and Resource Management in UAV-assisted Communications 20

2.1.1 UAV Trajectory Design and Deployment 21

2.1.2 Wireless Power Transfer Technique 22

2.1.3 Resource Management in UAV-assisted Networks 23

2.2 Optimisation in Reconfigurable Intelligent Surface-aided Communications 24

2.2.1 RIS for Energy-Efficient Communications 25

2.2.2 RIS-assisted UAV Communications 26

2.2.3 Others Trends for RIS-assisted Networks 27

2.3 Deep Reinforcement Learning in UAV-assisted Wireless Networks 28

2.4 Deep Reinforcement Learning for the Autonomous RIS-aided Communications 29

2.5 Contributions and Novelty of the Thesis 31

3 3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning 34

3.1 Introduction 34

3.2 System Model and Problem Formulation 36
3.2.1 Observation model 37
3.2.2 Game formulation 38
3.3 Q-learning algorithm for UAV-assisted IoT Networks 43
3.4 An Effective Deep Reinforcement Learning Approach for UAV-assisted IoT Networks 45
3.5 Deep Reinforcement Learning Approach for UAV-assisted IoT networks: A Dueling Deep Q-learning Approach 48
3.6 Simulation Results 51
3.6.1 Expected reward 52
3.6.2 Throughput comparison 60
3.6.3 Parametric Study 63
3.7 Conclusion 67

4 RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning 69
4.1 Introduction 69
4.2 System Model and Problem Formulation 73
4.2.1 Channel model 73
4.2.2 Power transfer phase 76
4.2.3 Information transmission phase 76
4.3 Hovering UAV for downlink power transfer and uplink information transmission in RIS-assisted UAV communications 78
4.3.1 The DDPG method 78
4.3.2 Game solving 80

4.4 Joint trajectory, EH time scheduling and the RIS phase shift optimisation using deep reinforcement learning 82

4.5 Proximal policy optimisation technique for joint trajectory, EH time and the phase shift optimisation 84

4.6 Simulation Results 87

4.7 Conclusion 93

5 Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning 96

5.1 Introduction 96

5.2 System Model and Problem Formulation 99

5.2.1 System Model 100

5.2.2 Problem Formulation 102

5.3 Centralised Optimisation for Power Allocation and Phase-shift Matrix 104

5.4 Parallel DRL for Joint Power Allocation and Phase-shift Matrix Optimisation 108

5.5 Proximal Policy Optimisation for Centralised and Decentralised Problem. 111

5.6 Simulation Results 113

5.7 Conclusions 120

6 Conclusions and future work 121

6.1 Summary of the thesis 121

6.1.1 UAVs and RIS-assisted Wireless Communications 121

6.1.2 Literature Review 121
6.1.3 3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning 122
6.1.4 RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning 122
6.1.5 Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning 122
6.2 Open Problems and Future Works 123
6.2.1 Multi-UAV Deployment and Trajectory Optimisation . . . 123
6.2.2 Resource Managements in Multi-RIS-aided Wireless Networks 123
6.2.3 Performance Analysis 124
6.2.4 Transfer Learning and Meta Learning in Wireless Communications 124

References **126**

List of Tables

2.1 A comparison with existing literature in Chapter 3 31

2.2 A comparison with existing literature in Chapter 4 31

3.1 Simulation parameters in Chapter 3. 52

4.1 Simulation parameters in Chapter 4 87

5.1 Simulation parameter in Chapter 5 115

List of Figures

1.1 Reinforcement learning model in wireless networks. 14

3.1 System model of UAV-aided IoT communications. 37

3.2 Trajectory obtained by using our dueling DQL algorithm. 53

3.3 The performance when using the DQL and dueling DQL algorithms with 3 clusters while considering different β/ζ values. . . . 54

3.4 The expected reward when using the DQL and dueling DQL algorithms with 5-cluster scenario. 55

3.5 The performance when using the DQL and dueling DQL algorithms with 5 clusters and different β/ζ values. 56

3.6 The expected reward when using the DQL algorithm with 5 clusters and different reward function settings. 57

3.7 The performance when using the dueling DQL with 5 clusters, and different β/ζ values 58

3.8 The network's sum-rate when using the DQL and dueling DQL algorithms with 3 clusters. 59

3.9 The obtained total throughput when using the DQL algorithm with 5 clusters. 61

3.10 The obtained throughput when using the DQL and dueling DQL algorithms in 5-cluster scenario. 62

3.11 The expected throughput when using the DQL and dueling DQL algorithms with 5 clusters. 64

3.12 The performance when using the DQL algorithm with different discount factors, γ, and exploration factors, ϵ. 65

3.13 The performance when using the DQL algorithm in 5-cluster scenario and different batch sizes, K. 65

3.14 The performance when using DQL algorithm with different learning rate, lr. 66

4.1 An illustration of the system model of UAV-assisted IoT wireless communications with the support of a RIS. 74

4.2 The sum-rate performance in the hovering UAV scenario with different numbers of IoT devices, N. 89

4.3 The sum-rate performance in the hovering UAV scenario with varying number of RIS elements, K. 90

4.4 The performance while using the PPO algorithm in the hovering UAV scenario with different number of RIS elements, K 91

4.5 The sum-rate performance with different number of IoT devices, N. 91

4.6 The sum-rate performance with different numbers of RIS elements, K. 92

4.7 The performance with different values of the discounting factor, ζ. 93

5.1 System setup. 100

5.2 A multi-agent learning for the RIS-assisted wireless networks . . . 109

5.3 The EE with $M = 10$ and $K = 20$. 116

5.4 The EE versus the number of UEs in each cluster, M. 117

5.5 The EE versus the number of the RIS elements, K. 117

5.6 The EE of the C-DDPG algorithm with different number of the RIS elements, K. 118

5.7 The EE of the P-DDPG algorithm with different number of the RIS elements, K. 119

5.8 The EE of the C-PPO algorithm with different number of the RIS elements, K. 119

5.9 The EE of the P-PPO algorithm with different number of the RIS elements, K. 120

List of Notations

$\boldsymbol{X}$, $\boldsymbol{x}$	Matrix $\boldsymbol{X}$ and vector $\boldsymbol{x}$
$\mathbb{E}[\cdot]$	The expectation function
$\mathsf{diag}(\boldsymbol{X})$	The vector of diagonal elements of $\boldsymbol{X}$
$x \sim \mathcal{CN}(0, \sigma_x^2)$	Complex-valued Gaussian random scalar with zero-mean and covariance σ_x^2

Chapter 1

Introduction and Overview

Recently, the use of unmanned aerial vehicles (UAVs) has received tremendous attention for applications such as surveillance [1], rescue missions, telecommunications [2–9]. UAVs can deliver low-cost, mobile and reliable wireless communication solutions for solving various real-life applications thanks to their advantages of agility and mobility. Reconfigurable intelligent surface (RIS) or intelligent reflecting surface (IRS), refer to the technology of intelligently using massive scattering reflectors. They can be deployed in a high location for cost-effective and reliable solutions to extend the network coverage and performance.

Despite the numerous advantages, there are several technical challenges in the UAV and RIS-assisted wireless network such as 3D trajectory design, air-to-ground modelling, channel estimation, flight time optimisation, large-scale optimisation, information transmission [10–12]. In this chapter, we first present the advantages, practical applications and challenges of RIS and UAV-assisted wireless communications. Then, we provide an overview of deep reinforcement learning (DRL) and our research motivation, followed by our contributions.

1.1 UAV-enabled wireless communications

Thanks to the agility of UAVs, they are capable of supporting compelling applications and are beginning to be deployed more broadly. The high altitude of UAVs can overcome some bottlenecks of the existing scenarios, such as shadowing, blockages, remote areas, and emergency services. Some real-life applications of the UAVs are surveillance [1], geography exploration [13], disaster rescue mission [14–16], and wireless communications [17, 18]. Recently, the UK and Chile authorities deployed UAVs to deliver medical support and other essential supplies to vulnerable people in response to Covid-19 [19, 20]. In addition, UAVs were used for image collection and high-resolution topography exploration [21].

Recently, UAVs have received colossal attention for supporting wireless communications to extend network coverage [2–9]. The UAVs can serve as aerial base stations (BSs) that can deliver low-cost, mobile and on-demand networks with ubiquitous coverage and robust handover [10, 22–26]. The high altitude of the UAVs helps them to effectively connect with the ground users and BSs by line-of-sight (LoS) communication links. The UAVs can provide the maximum coverage and capacity for the targeted ground users by adjusting their altitude. Their nimble mobility feature helps them avoid signal blockages and have better connections with the ground users. Moreover, the UAV can be flexibly deployed in the poor terrestrial coverage areas, such as remote villages and disaster areas, to provide on-demand services.

1.1.1 Applications of UAV Communications

UAVs have been widely used for enhancing wireless networks' performance as a benefit of their high altitude and mobility features [1–10, 14–16, 18, 22–28]. The applications of UAVs in wireless networks span across diverse research fields, such as wireless sensor networks (WSNs) [29], caching [30], heterogeneous cellular networks [31], massive multiple-input multiple-output (MIMO) [32], disaster communications [14, 33] and device-to-device communications (D2D) [34]. For example, the UAV can be deployed for enhancing the network coverage and capacity in sports events in which the existing network infrastructure cannot meet the demand and needs to be boosted rapidly. Moreover, the UAV can provide ubiquitous wireless coverage in rural areas where terrestrial infrastructure (e.g., cables) is costly. In such scenarios, the UAV-enabled wireless network is an ideal solution to provide low-cost and on-demand internet to ground users.

The UAVs are also used for public safety communications, such as during natural disasters [16, 35, 36]. In such critical scenarios, the terrestrial networks can be damaged and destroyed while there is a need for communications between the victims and rescue teams. Thus, the aerial network based on UAVs is a promising solution to provide a robust, fast and on-demand communication system. The UAVs can quickly fly to the positions of ground users to provide connections. In addition, the UAVs can function as data collection machines to collect data in the Internet-of-Things (IoT) networks in which the ground users are limited in terms of the transmit power and communication range [3, 37–42]. For example, in environments with no terrestrial infrastructure, such as remote areas and mountains, the UAVs can be deployed with the emerging technologies (e.g., wireless power transfer) to provide energy-efficient and reliable communications [43–51]. Moreover, the UAVs can be used in smart cities where deploying a base station is expensive and the transmit signal is blocked due to high buildings and obstacles [52, 53]. Clearly, the UAV-assisted wireless communications can effectively establish high speed, on-demand and cost-effective services in crowded locations or in areas poorly covered by terrestrial networks.

1.1.2 Challenges in UAV-assisted Communications

Despite the numerous advantages of UAVs for supporting wireless communications, there are still some challenges to UAVs adoption [10, 11]. Firstly, the high variance and sensitive vibration of UAVs can affect the channel characteristics. The air-to-ground channel is generally used to formulate the links between the UAV and the users. However, it depends on the altitude of the UAV and the propagation environment. Thus, there is a need for comprehensive measurements to formulate a generic channel model. Secondly, the maintenance of the networks and deployment of the UAVs are still challenging due to the limited propulsion power level, the flight time of the UAVs and several stringent communications constraints. Given the several limitations of on-board power level and the ability to adapt to changes in the environment, UAVs may not be fully autonomous and can only operate for short flight durations unless remote laser-charging is used [54].

Moreover, solving a continuous trajectory design for the UAVs is a unique challenge with the high variance of 3D position at the UAVs. Unlike terrestrial base stations, UAVs can move in a continuous 3D space. Thus, when optimising the UAV's trajectory or deployment, the channel variation due to the UAV's attitude and energy consumption needs to be explicitly taken into account. In addition, due to some challenging tasks such as topographic surveying, data collection or obstacle avoidance, the existing UAV technologies cannot operate in an optimal manner.

1.2 Reconfigurable Intelligent Surfaces-aided Wireless Networks

RIS has recently received significant attention for enhancing the network quality and coverage. The signal arrived at the RISs is reflected toward the receivers by adjusting the phase-shift matrix and active elements. Thus, the received signals can be improved, and the interference from unexpected sources can be mitigated for better network services. In addition, the reliability and zero-delay are also notable advantages of the RIS in supporting reliable and low-cost wireless communications.

1.2.1 Benefits and Applications of RIS-assisted Communications

RISs have become an emerging technique owing to their capability of modifying wireless communications. There are several advantages of RISs as: *easy deployment and sustainable operations*, *capacity, spectral efficiency and energy efficiency enhancement*, *flexible reconfiguration and compatibility* [12,55]. With the low-cost passive scattering elements, the RISs can be easily deployed and replaced. They can be attached to the high buildings, ceilings, vehicles, etc., and on UAVs to provide better cellular services and extend the network coverage at a low cost. The high locations of the RISs can help to enhance the received signal at the receivers and suppress the interference. As such, the capacity, energy efficiency and max-min fairness among users are significantly improved in the RIS-assisted wireless networks. Additionally, the phase-shift matrix at the RIS can be optimised to reflect the signal toward specific directions. The efficient phase-shift optimisation can significantly improve the network performance. Furthermore, the RISs are also compatible with many other emerging technologies such as UAV [2, 8, 38, 56, 57] and mobile edge computing [58] to bring more reliable and high-speed services to the users.

The applications of the RISs are diverse in the wireless networks. The RISs can be used in the cellular network to bypass the obstacles and improve the links from the BS toward users [59–61]. Moreover, the RISs can also be deployed for strengthening the signal and mitigating interference in device-to-device communications [62–66]. On the other hand, the RISs can act as reliable middle layers to cancel the undesired signal in the physical layer security [67–69]. In addition, the RIS is attached to the UAV to provide an on-demand and high-speed aerial network to the ground users [70, 71]. The RIS can also be exploited for assisting the smart city [59, 61], autonomous vehicles [72] and intelligent wireless sensor networks [8, 73].

1.2.2 Challenges in RIS-assisted Communications

The RISs effectively enhance the received signal, suppress the interference, and minimise the transmit power. However, there are still some challenges: there is a need for energy-efficient channel estimation, practical protocols for information exchange, real-time and distributed optimisation, and light-weight phase reconfiguration [12].

Given the massive array of the scattering elements, the RIS, controlled by a processing unit, reconfigures the phase shift to reflect the signal toward the specific directions. To achieve a better accuracy of channel estimation, the BSs or UAVs need to use more power for computational purposes and information exchange. Moreover, most existing studies still consider full knowledge of channel state information (CSI) in the RIS-assisted networks. Unlike the traditional systems, besides the estimation of the direct channel, the reflected channel, which is cascaded by the BS-RIS link, the RIS phase-shift matrix and the RIS-users link, also needs to be estimated. The direct channel estimation can be made by the traditional methods. However, it is challenging to estimate the channel of the BS-RIS and of the RIS-users link due to the limited power level and processing capability at the RIS. Thus, there is a need for energy-efficient channel estimation in RIS-assisted wireless networks.

As aforementioned, the RIS is comprised of a large array of elements. Thus, the optimisation of the RIS phase shift leads to large-scale optimisation problems. The joint optimisation of the phase shift with the resource management at the BS or users also makes the problem more challenging, especially when the number of RIS's elements increases or when multi-RISs are deployed. Moreover, RIS-assisted wireless communications have more sensitive and vulnerable channel links than conventional wireless networks. The stringent constraints in terms of communications, delay and power, bit error rate, etc., also need to be satisfied. Hence, there is a need to investigate more accurate and efficient methods to deal with the high-dynamic environment of RIS-aided wireless communications.

1.3 Deep Reinforcement Learning in Wireless Networks

Machine learning has recently been proposed for the intelligent support of UAVs and other devices in the network [18,27,28,32,46,74–78]. Reinforcement learning (RL) is capable of searching for an optimal policy by trial-and-error learning. However, it is challenging for model-free RL algorithms like the Q-learning algorithm to obtain an optimal strategy while considering a large state and action space. Fortunately, with the emerging neural networks, the sophisticated combination of RL and deep learning, namely deep reinforcement learning (DRL), is eminently suitable for solving high-dimensional problems. Hence, DRL algorithms have been widely applied in fields such as robotics [79], business management [80], and gaming [81]. Recently, DRL has also become popular in solving diverse problems in wireless networks thanks to their decision-making ability and flexible interaction with the environment [18, 27, 28, 30, 32, 46, 76–78, 82–84]. For example, DRL was used for solving problems in the areas of resource allocation [27,28,83], navigation [6,32], and interference management [76]. This section introduces the fundamental concept of the Markov decision process (MDP) [85] and the RL algorithm.

1.3.1 An Overview of Markov Decision Processes

The MDP [85] is defined by a 4-tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} >$ where $\mathcal{S}$ and $\mathcal{A}$ is the finite set of states and actions, respectively; $\mathcal{P}$ is the transition probability function with $\mathcal{P}_{ss'}(a)$ is the probability from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ after the action $a \in \mathcal{A}$ is taken; $\mathcal{R}$ is the reward obtained after the action a is executed. We define π to be the policy which is mapping the state into the action. The objective of an MDP is to find an optimal policy π^* to maximise the defined reward that is represented as follows:

$$\max \mathbb{E}_{\pi}\Big[\sum_{t=1}^{T} \gamma \mathcal{R}(s', s, a)\Big] \tag{1.1}$$

where $\mathbb{E}[\cdot]$ is the expectation function, γ is the discounting factor.

1.3.2 A Brief Overview of Deep Reinforcement Learning

In the RL algorithm, one or many agents interact with the environment and learn through interaction. The essence of RL is trial-and-error learning, in which the agent observes the state and executes the action toward the environment to adjust its behaviour in response to obtained rewards. Deep learning enables RL to work in more complicated problems with a high-dimensional state and action space.

There are two main approaches for solving the RL problems: value functions (deep Q-learning, SARSA, double deep Q-learning, etc.) and policy search (vanilla policy gradient, proximal policy optimisation, etc.). There is also a hybrid model, namely the actor-critic approach (deep deterministic policy gradient algorithm, etc.), based on both value function and policy search. In DRL, we use deep neural networks for approximating the value function V, the action-value function Q, the advantage function A and the policy π. In the value search approach, we consider the gap between the received rewards in two samples to adjust the value function. In the policy search algorithm, we directly find the policy for the problems.

1.3.2.1 Value Function

The idea of the value function methods relies on the estimation of the value in a given state. The state-value function $V^{\pi}(s)$ is obtained following the policy π starting at the state s as

$$V^{\pi} = \mathbb{E}\Big[\mathcal{R}|s, \pi\Big], \tag{1.2}$$

where the expectation operation $\mathbb{E}[\cdot]$ depends on the transition function $\mathcal{P}_{ss'(a)} = p(s'|s, a)$ and the stochastic property of the policy π.

Our goal is to find the optimal policy π^*, which has a corresponding optimal state-value function $V^*(s)$ as

$$V^*(s) = \max_{\pi} V^{\pi}(s), s \in \mathcal{S}. \tag{1.3}$$

To maximise the expected cumulative reward, the agent chooses the action $a \in \mathcal{A}$ following the optimal policy π^* that satisfies the Bellman equation [86]

$$V^*(s) = V^{\pi^*} = \max_{a \in \mathcal{A}} \left\{ \mathbb{E}\Big(r(s, a)\Big) + \zeta \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) V^*(s') \right\}. \tag{1.4}$$

The action-value function is defined as the obtained reward when the agent takes action a at the state s under the policy π as

$$Q^{\pi}(s,a) = \mathbb{E}\Big(r(s,a)\Big) + \zeta \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}(a) V(s'). \tag{1.5}$$

The optimal policy $Q^{*}(s,a) = Q^{\pi^{*}}$, we have

$$V^{*}(s) = \max_{a \in \mathcal{A}} Q^{*}(s,a) \tag{1.6}$$

1.3.2.2 Policy Search

Instead of considering the value function model, the agent can directly find an optimal policy π^{*}. Among policy search methods, the policy gradient is most popular due to its efficient sampling with a large number of parameters. The reward function is defined by the performance under the policy π as

$$J(\theta) = \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \pi_{\theta}(a|s) r^{\pi}(s,a), \tag{1.7}$$

where θ_{π} is the vector of the policy parameters and $d^{\pi}(s)$ is the stationary distribution of Markov chain with the policy π_{θ}. The optimal policy π^{*} can be obtained by using gradient ascent for adjusting the parameters θ_{π} relying on the $\nabla_{\theta} J(\theta_{\pi})$. For any MDP, we have [87]

$$\begin{aligned} \nabla_{\theta} J &= \sum_{s \in \mathcal{S}} d^{\pi}(s) \sum_{a \in \mathcal{A}} \nabla_{\theta} \pi(a|s) Q^{\pi}(s,a) \\ &= \mathbb{E}_{\pi_{\theta}} \Big[\nabla_{\theta} \ln \pi_{\theta}(s,a) Q^{\pi}(s,a) \Big] \end{aligned} \tag{1.8}$$

The REINFORCE algorithm, a Monte-Carlo policy gradient learning, adjusts the parameters θ_π by estimating the return using Monte-Carlo methods and episode samples. The optimal policy parameter θ_π^* can be obtained by

$$\theta_\pi^* = \underset{\theta_\pi}{\operatorname{argmax}} \, \mathbb{E}\left[\sum_{a\in\mathcal{A}} \pi(a|s;\theta_\pi) r(s,a)\right], \tag{1.9}$$

The gradient is defined as

$$\nabla\theta_\pi = \mathbb{E}_\pi\Big[\nabla_{\theta_\pi} \ln \pi(a|s;\theta_\pi) r(s,a)|_{s=s^t,a=a^t}\Big]. \tag{1.10}$$

We use the gradient ascent to update the parameters θ_π as

$$\theta_\pi \leftarrow \theta_\pi + \varepsilon\nabla\theta_\pi, \tag{1.11}$$

where $0 \leq \varepsilon \leq 1$ is the step-size parameter. The optimal action a^* can be obtained with the maximum probability as follows:

$$a^* = \underset{a\in\mathcal{A}}{\operatorname{argmax}} \, \pi(a|s;\theta_\pi). \tag{1.12}$$

1.4 Motivation, Contributions and Organisation

Machine learning is an effective tool for optimising the performance of large-scale networks under dynamic environments. One of the approaches is the DRL algorithm, a combination of RL and neural networks. In this thesis, we propose methods based on the DRL algorithms for the UAV's trajectory design, resource allocation, and the phase-shift matrix for dealing with the aforementioned emerging challenges in the UAV and RIS-assisted wireless communications.

1.4.1 Research Motivation

UAV and RIS bring several advantages to extend the coverage, enhance the network quality and improve the energy-efficient performance. There are still some challenges for adopting UAV and RIS in real-life applications due to stringent constraints of flying time and power consumption. Some traditional approaches are proposed for solving the resource allocations, trajectory design and RIS's phase shift optimisation. However, the existing works mostly ignore the difficulty in channel estimation, mobile users, power consumption and delay in information transmission. In addition, some challenging tasks of joint optimisation are often considered as sub-problems. They are not applicable in real-time optimisation and large scale optimisation.

RL algorithm is specially designed for agents to make a sequence of decisions. In Fig. 1.1, the agent can be any component that has a processing unit in the wireless networks, such as the UAVs, servers, ground users, etc. The agents interact with the environment to achieve the optimal reward through trial-and-error learning. Thus, DRL techniques have been used for lending each node some degree of autonomy to make the wireless network more intelligent [18, 27, 28, 30, 46, 82, 83, 88]. Firstly, the wireless environments are transferred into a digital form of IoT locations, UAV flying velocity, angle, position, and channel model between each element in the network. Then, the DRL algorithms are used to train the agents. In this thesis, we use several algorithms including deep Q-learning (DQL), double Q-learning (DDQL), dueling deep Q-learning (dueling DQL) for the discrete problem and deep deterministic policy gradient (DDPG), proximal policy optimisation (PPO) for the continuous design. Moreover, to reduce the

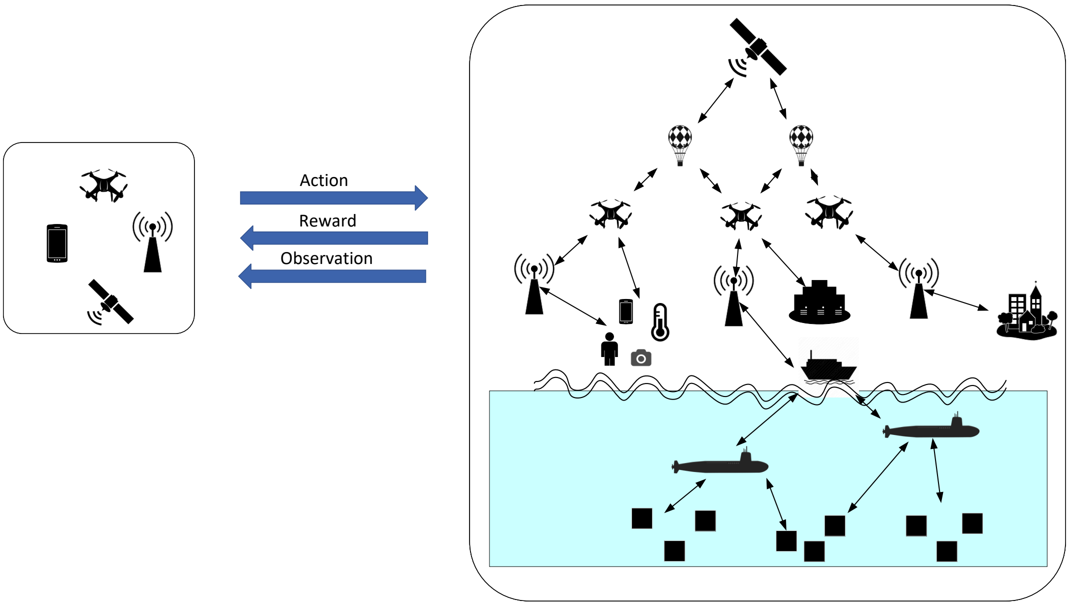


Figure 1.1: Reinforcement learning model in wireless networks.

delay in the information transmission between the centralised processing unit and users, we also use multi-agent learning.

1.4.2 Summary of Contributions

The main contributions of this thesis are summarised as follows:

In **Chapter 3**, we consider a UAV-assisted IoT system for finding the shortest flight path of the UAVs while maximising the joint reward function based on the shortest flight distance and the uplink transmission rate from IoT devices.

- In our UAV-aided system, the maximum amount of data is collected from the users with the shortest distance travelled.

- Our UAV-aided system is specifically designed for tackling the stringent constraints owing to the position of the destination, the UAV's limited flight time and the communication link's realistic constraints. The UAV's objective is to find the optimal trajectory for maximising the total network throughput, while minimising its distance travelled.

- Explicitly, these challenges are tackled by conceiving bespoke DRL techniques for solving the above problem. To elaborate, the area is divided into a grid to enable fast convergence. Following its training, the UAV can have the autonomy to make a decision concerning its next action at each position in the area, hence eliminating the need for human navigation. This makes our UAV-aided system more reliable, practical and optimises the resource requirements.
- A pair of scenarios are considered relying either on three or five clusters for quantifying the efficiency of our novel DRL techniques in terms of both the sum-rate, the trajectory and the associated time. A convincing 3D trajectory visualisation is also provided.
- Finally, but most importantly, it is demonstrated that our DRL techniques approach the performance of the optimal "genie-solution" associated with the perfect knowledge of the environment.

Although the existing DRL algorithms have been well exploited in wireless networks, it is challenging to apply to current scenarios due to stringent constraints of the considered system, such as UAV's flying time, transmission distance, and mobile users. As with the DQL and dueling DQL algorithm, we discretise the flying path into grid and the UAV only needs to decide the action in a finite action space. With the finite state and action space, the neural networks can be easily trained and deployed for online phase. With other existing RL algorithm, we have tried and found out that some of them are not effective in solving our proposed problem. Meanwhile, the continuous solver RL algorithms are not suitable and so challenging for the trade-off problem. Therefore, in this chapter, we propose the DQL and dueling DQL algorithm to obtain the optimal trade-off in total achievable sum-rate and trajectory. As such, we can transfer a real-life application into a digital environment for optimisation and solve it efficiently.

In **Chapter 4**, we introduce a new system model for RIS-assisted UAV communications with the downlink power transfer and uplink information transmission protocol for maximising the network sum-rate.

- We conceive a system model of UAV-assisted IoT wireless power transfer with the support of a RIS. The IoT devices harvest energy in the downlink and transmit information in the uplink to the UAVs.

- To characterise the agility of UAVs in supporting the energy harvesting (EH) and information transmission of IoT devices, we consider two scenarios of UAVs. Firstly, the UAV is hovering at the centre of the cluster and provides energy to the IoT devices. The RIS helps alleviate the uplink interference when the IoT devices transmit their information to the UAV. Secondly, the UAV is deployed in an initial location and required to find a better location for communication. In each location of the UAV's flying trajectory, the EH time scheduling and the RIS's phase shift matrix are optimised for maximising the network throughput performance.

- For the defined problem, we formulate a Markov decision process (MDP) [86] with the definition of the state space, action space and the reward function. Then, we propose a method based on deep deterministic policy gradient (DDPG) and proximal policy optimisation algorithm (PPO) for solving the maximisation game.

- Our results suggest that with the support of the RIS, a better connection is established and the overall performance is significantly improved.

However, when deploying the optimisation algorithm with DRL into RIS-assisted UAV communications, previous works assumed the perfect condition of the environment, flat fading channels, static users, and perfect CSI, which are unrealistic and infeasible for real-life applications. Furthermore, the delay when using a mathematical model and in the centralised learning is huge for real-time use cases. To overcome these aforementioned shortcomings, in this chapter, we also propose a parallel learning for reducing the information transmission requirement of the centralised approach.

In **Chapter 5**, we consider multi-UAV networks supported by a RIS panel to enhance the network performance.

- We conceive a wireless network of multi-UAVs supported by an RIS. Each UAV is deployed for serving a specific cluster of UEs. Due to the severe shadowing effect, the RIS is used to enhance the received signal's quality at the UEs from the associated UAV and to mitigate the interference from others.

- The EE problem is formulated for the downlink channel with the power restrictions and the RIS's requirement. To optimise the EE network performance, we propose a centralised DRL technique for jointly solving the power allocation at the UAVs and phase-shift matrix of the RIS. Then, a parallel learning is used for training each element in our model to be intelligent and to reduce the delay when transmitting the action between UAV and the RIS.

- To improve the network performance, we introduce the proximal policy optimisation (PPO) algorithm with a better sampling technique.
- Through the numerical results, we demonstrate that our proposed methods efficiently solve the joint optimisation problem with the dynamic environmental setting and time-varying CSI and outperform the other benchmarks.

1.4.3 Outlines of the thesis

The organisation of the thesis are as follows. In Chapter 2, we introduce some existing works related to RIS and UAV-assisted wireless communications. In Chapter 3, we propose a novel DRL-aided UAV-assisted system for finding the optimal UAV path to maximise the joint reward function based on the shortest flight distance and the uplink transmission rate. In Chapter 4, we consider the IoT wireless networks with the support of a UAV and one RIS and employ the downlink power transfer and uplink information transmission protocol for maximising the network's sum-rate. In particular, we adopt the harvest-then-transmit protocol, which means the IoT devices use all the harvested energy in the first phase for transmitting during the remaining time. Then, the methods based on the DQL algorithm and dueling DQL algorithm are deployed for solving the problem in RIS-assisted UAV communications. In Chapter 5, we exploit the efficiency of DRL techniques in multi-UAV-assisted wireless communications with the support of one RIS. We propose efficient DRL algorithms by jointly optimising the power allocation of the UAVs and the RIS's phase-shift matrix for maximising the EE performance. The DRL approaches bring a flexible and autonomous ability to the UAVs and the RIS. With trained neural networks, the UAVs can choose a proper flying direction and velocity while the RIS can adjust the phase shift without delay. Furthermore, continuous learning with up-to-date data by interaction with the environment helps the UAVs and RIS to adapt to the dynamic environment. Moreover, we also improve the model with multi-agent learning to reduce the information transmission delay. The summary of the thesis and the potential future works are presented in Chapter 6.

Chapter 2

Literature Review

2.1 Trajectory Design and Resource Management in UAV-assisted Communications

Given the mobility, agility, and flexibility, using the UAVs is a promising technique for enhancing the network performance. In particular, the UAV can act as an aerial base station to provide ubiquitous coverage and on-demand services to ground users in different scenarios such as natural disaster relief and temporary hotspots. However, numerous challenges such as 3D trajectory design, resource management, and energy limitations must be solved for enabling effective UAV-assisted networks. Several techniques have been proposed to overcome these challenges to enhance the energy efficiency and the network performance.

2.1.1 UAV Trajectory Design and Deployment

The high-flying altitude of the UAV helps the wireless networks improve the coverage and transmit signal [7, 15, 17, 18]. In [15], multiple UAVs were deployed in a disaster area to efficiently support the users. The K-means algorithm was proposed for the deployment mission, while the Block Coordinate Descent (BCD) procedure was used for maximising the worst end-to-end sum-rate. In [17], the authors used the UAV as a mobile data collector. The optimised UAV's flying path and the wake-up scheduling at the sensor nodes helped to reduce the energy consumption in both the UAV and the sensors. The authors in [18] considered the UAV as an energy supplier for the non-fixed power source devices to assist communications in D2D networks. In [7], the UAV's trajectory was optimised to maximise the energy efficiency (EE) in an unconstrained condition and circular trajectory.

The issues of data collection, energy minimisation, and path planning have been also considered in [5, 17, 77, 89–96]. In [17], the authors minimised the energy consumption of the data collection task considered by jointly optimising the sensor nodes' wake-up schedule and the UAV trajectory. The authors of [89] proposed an efficient algorithm for joint trajectory and power allocation optimisation in UAV-assisted networks to maximise the sum-rate during a specific length of time. A pair of near-optimal approaches for optimal trajectory was proposed for a given UAV power allocation and power allocation optimisation for a given trajectory. In [90], the authors introduced a communication framework for UAV-to-UAV communication under the constraints of the UAV's flight speed, location uncertainty and communication throughput. Then, a path planning algorithm was proposed for minimising the associated completion time task while balancing the performance by computational complexity trade-off.

2.1.2 Wireless Power Transfer Technique

Along with the development of the IoT devices is the increased power supply for each device. However, not all the nodes are equipped with fixed power providers and have solar batteries. Moreover, in some environments such as human bodies, toxic locations, or underwater, replacing and recharging the batteries are expensive, complicated and even impossible. Thus, wireless power transfer (WPT) is a promising technique that enables the IoT nodes to have enough energy to maintain the connections with the BSs. There are two major applications of WPT in wireless communications: wireless powered communication network (WPCN) and simultaneous wireless information and power transfer (SWIPT) [97, 98]. In WPCN, the devices harvest energy from the source in the downlink in order to send information in the uplink, whereas in SWIPT, the energy and information are transferred from the source to devices [99, 100].

The downlink power transfer and uplink information transmission protocol is one of the solutions to enable the IoT devices to harvest energy from source providers and switch to information transmission in the uplink phase on demand [73, 101–105]. That helps reduce the power consumption as well as the cables and wires for providing power. In [106], the author considered joint optimisation of the time scheduling, the transmit signal and the transmit power for maximising the network throughput in a general multi-user wireless powered interference channel. In [107], the authors present an energy trading game in the WPCN where two scenarios of the multiantenna power station (PS) and the user nodes belonging to the same service operator and in different service operators were considered.

UAVs also act as an energy provider source for powering the sensor nodes. Although the UAVs have limited power onboard, the UAV-enable WPT system can improve the energy efficiency due to the LoS links between the UAV and the IoT nodes. Very recently, UAVs have been used as an energy supplier for the energy constrained IoT devices due to the fact that the UAV can easily be recharged at the docking station and the energy of these IoT devices is comparably smaller than UAV's capacity. Moreover, UAVs can adjust their locations and altitude to approach the ground users and transfer energy thanks to their flexibility feature [43–51]. For example, in [44], the UAV flies to charge the device on the ground and then return to the landing docks. The authors used an iterative algorithm and a transition-based design to optimise the UAV trajectory, lengths of working period and charging phase. In [45], the UAV-mounted mobile energy transmitter was deployed to provide energy to the receivers. In [48], the authors considered four sub-problems to jointly optimise the 3D trajectory and time allocation to maximise the energy harvested at the receivers.

2.1.3 Resource Management in UAV-assisted Networks

With the aforementioned benefits, the use of UAVs in wireless networks efficiently enhances the network performance. However, the associated resource allocation problems remain challenging in real-life applications. Several techniques have been developed for solving resource allocation problems [6,15,27,28,49,108–111]. In [108], the authors have conceived a multi-beam UAV communications and a cooperative interference cancellation scheme for maximising the uplink sum-rate received from multiple UAVs by the base stations (BS) on the ground. The UAVs were deployed as access points to serve several ground users in [49]. Then, the authors proposed successive convex programming for maximising the minimum uplink rate gleaned from all the ground users. In [6], the authors characterised the trade-off between the ground terminal transmission power and the specific UAV trajectory both in a straight and in a circular trajectory. In [110], the authors jointly optimised the UAV trajectory, backscatter devices, and carrier emitters on the ground to maximise the EE in the UAV-assisted backscatter communication network. UAV-aided wireless networks have also been used for machine-to-machine communications [90], and D2D scenarios in 5G [111–113]. In [111], the UAVs were working as relays to help the D2D communications. The authors optimised the UAV's power, D2D users' power, the available bandwidth and the UAV trajectory to maximise the network throughput. In addition, the UAVs were also used to assist the ultra-reliable low-latency computation offloading in [37]. The UAVs' position, resource allocation and the offloading decisions were divided into two sub-problems, and a two-stage approximate algorithm was proposed for maximising the rate of served requests.

2.2 Optimisation in Reconfigurable Intelligent Surface-aided Communications

The RISs reflect the signal from the transmitters to the receivers by controlling the phase-shift value of a massive amount of scattering reflectors. The reflected signals can be combined coherently to improve the received signal or destructively to suppress the interference.

2.2.1 RIS for Energy-Efficient Communications

Recently, RIS technology has been introduced as a low-cost, and easily installed technology to mitigate interference and direct transmitted signals toward their receivers [59–61, 114]. In [60], the authors considered two-way communications assisted by a RIS. The reciprocal channel to maximise the signal-to-interference-plus-noise ratios (SINR) and the non-reciprocal channel with the target of maximisation of the minimum SINR were considered. The gamma approximation was used for the reciprocal channel, while the semi-definite programming relaxation and a greedy-iterative method were used for the non-reciprocal channel. In [114], an iterative algorithm with low computation complexity was proposed to solve the joint optimisation of transmit beamforming vector and the phase shift of a RIS under proper and improper Gaussian signalling. In [61], the authors optimised the beamforming matrices at the BS and the reflective vector at the RIS to minimise the total transmit power at a multiple-input single-output (MISO) non-orthogonal multiple access (NOMA) network. An algorithm based on the second-order cone programming-alternating direction method of multipliers was proposed to reach an optimal local problem.

As aforementioned, RIS has been recently attracting enormous attention as an emerging technology for enabling beyond 5G due to its unique characteristics, which include the low-cost production and less energy consumption [59, 60, 62, 101, 114–119]. In [115], an algorithm was proposed for maximising the weighted sum-rate of all users via beamforming vector and RIS phase-shift optimisation under the perfect CSI and imperfect CSI scenarios. In [116], the power allocation and the phase-shift optimisation algorithm was proposed for maximising the EE performance. In [62], the RIS was used for enhancing communication and reducing interference in the D2D networks. Two sub-problems with the fixed power transmission and the discrete RIS's phase-shift matrix were considered and solved efficiently. The authors in [117] optimised the beamforming vector at secondary users transmitter and the RIS phase-shift in a downlink multiple-input single-output (MISO) cognitive radio system with multiple RISs. The perfect CSI and imperfect CSI scenarios were considered; then, the block coordinate descent procedure was used to maximise the achievable sum-rate.

2.2.2 RIS-assisted UAV Communications

By utilising both advantages of the UAV and the RIS, the received signal at the ground users is strengthened while the power consumption is reduced and the flying time of the UAV can be extended [2, 8, 38, 56, 57]. In [57], the UAV's trajectory and the RIS's passive beamforming vector were optimised to maximise the average rate in RIS-assisted UAV communications. The problem was derived into two subproblems; then, a closed-form phase shift algorithm was introduced to find the local optimal reflective matrix and the successive convex approximation was used to find the suboptimal trajectory solution. In [38], the UAV acts as a mobile relay, and the RIS was used to provide short packets communications ultra-reliable and low-latency between ground transmitter and ground IoT devices. The UAV's position, the RIS phase shift and the blocklength were optimised to minimise the total decoding error rate by using a polytope-based method, namely Nelder-Mead simplex. In [56], the joint beamforming vector, trajectory and phase-shift optimisation algorithm was proposed for maximising the received signal at the ground users in the UAV-assisted wireless communications.

2.2.3 Others Trends for RIS-assisted Networks

The efficiency of RIS is also investigated in wireless power transfer [101, 102] and mobile edge computing [58]. In [101], the authors designed a time-switching protocol for a RIS with the energy harvesting phase to charge the RIS capacitor and the signal reflecting phase to assist the transmission from the access point (AP) to the receivers. The AP's transmit beamforming, the RIS's phase scheduling and the passive beamforming were optimised to maximise the information rate. Two sub-problems were solved following the conventional semi-definite relaxation method and monotonic optimisation. In [102], the transmit precoding matrices of the BS and the RIS's passive phase shift matrix were optimised for maximising the weight sum-rate of all information receivers in power transfer scenarios. The RIS is also used for supporting the D2D communications [62–66]. In [62], the authors considered two sub-problems with fixed passive beamforming vector at transmitters and fixed phase shift matrix at the RIS. The gradient method and a local search algorithm were proposed to solve these sub-problems.

2.3 Deep Reinforcement Learning in UAV-assisted Wireless Networks

By relying on their decision-making ability, DRL algorithms have been used for lending each node some degree of autonomy [18, 27, 28, 30, 46, 82, 83, 88]. In [82], an optimal DRL-based channel access strategy to maximise the sum-rate and α-fairness was considered. In [27, 28], DRL techniques were used for enhancing the energy efficiency of D2D communications. As a further advance, caching problems were considered in [30] to maximise the cache success hit rate and to minimise the transmission delay. The authors designed both a centralised and a decentralised system model and used an actor-critic algorithm to find the optimal policy.

In UAV-assisted wireless networks, the DRL algorithms have shown impressive results for solving the resource management problems [120–123]. In [120], the multiple cooperative UAVs was deployed for assisting the cellular network. The iterative algorithm with two steps of using the deep Q-learning algorithm and a difference of convex algorithm was proposed to optimise the UAV's positions, transmit beamforming and the UAV-users association to maximise the network's sum-rate. In [121], the UAV was deployed with the WPT technique to charge the ground devices and collect data. The multi-objective deep deterministic policy gradient method was proposed to solve the three objectives: sum-rate maximisation, harvested energy maximisation, and energy consumption minimisation.

DRL algorithms have also been applied for path planning in UAV-assisted wireless communications [32, 76–78, 84, 124, 125]. In [76], the authors proposed a DRL algorithm based on the echo state network of [126] for finding the flight path, transmission power and associated cell in UAV-powered wireless networks. The so-called deterministic policy gradient algorithm of [127] was invoked for UAV-assisted cellular networks in [84]. The UAV's trajectory was designed for maximising the uplink sum-rate attained without the knowledge of the user location and the transmit power. Moreover, in [32], the authors used the DQL algorithm for the UAV's navigation based on the received signal strengths estimated by a massive MIMO scheme. In [77], Q-learning was used for controlling the movement of multiple UAVs in a pair of scenarios, namely for static user locations and for dynamic user locations under a random walk model. In [46], the authors characterised the DQL algorithm for minimising the data packet loss of UAV-assisted power transfer and data collection systems. The multi-agent DRL was used for trajectory design and model selection in a cellular internet of UAVs in [125]. However, the aforementioned contributions have not addressed the joint trajectory and data collection optimisation of UAV-assisted networks, which is a difficult research challenge. Furthermore, these existing works mostly neglected interference, 3D trajectory and dynamic environment.

2.4 Deep Reinforcement Learning for the Autonomous RIS-aided Communications

The demand for a technique that is flexible and adaptive to changes in the environment while satisfying real-life constraints is rising, and DRL algorithms are among the most potential methods to deal with these problems in wireless networks [18,27,28,128]. Recently, DRL algorithms are also used for the RIS-assisted wireless networks with promising results [68, 129–131]. The power allocation and the phase shift optimisation were optimised for maximising the sum rate in [129]. In [130], a RIS-assisted UAV was deployed for serving ground users. The trajectory and phase shift optimisation relying on DRL for maximising the sum rate and fairness of all users was proposed. In [68], the authors used a RIS to assist the secure communications against eavesdroppers. The DRL algorithms were used to optimise the BS beamforming and the RIS's reflecting beamforming were shown to improve the secrecy rate and the quality-of-service satisfaction probability. In [131], a deep deterministic policy gradient was proposed to obtain the optimal phase shift matrix at the RIS to maximise the received signal-to-noise ratio (SNR) in a MISO system. In [2], the joint optimisation of the power and the RIS's phase shift in a multi-UAV-assisted network is considered.

Since DRL is an effective solution for solving the dynamic environment with continuous moving [18, 27, 28, 82], some recent works have explored the efficiency of the DRL techniques for RIS-assisted wireless networks [71, 102, 131, 132]. The author in [102] optimised the transmit beamforming vector and the RIS phase-shift model by using the DRL algorithm to maximise the total sum-rate. A deep Q-learning and deep deterministic policy gradient were proposed and showed impressive results in the MISO communications. To minimise the sum age-of-information, the authors in [71] proposed a DRL algorithm to adjust the UAV's altitude and the RIS phase-shift.

Table 2.1: A comparison with existing literature in Chapter 3

	[5]	[29]	[46]	[77]	[91]	[32]	[92]	[124]	[93]	[94]	Our work
3D trajectory				✓			✓			✓	✓
Sum-rate maximisation	✓		✓	✓				✓	✓	✓	✓
Time minimisation		✓	✓				✓		✓		✓
Dynamic environment				✓	✓			✓			✓
Unknown users											✓
Reinforcement learning			✓	✓		✓		✓			✓
Deep neural networks			✓			✓		✓			✓

Table 2.2: A comparison with existing literature in Chapter 4

	[57]	[38]	[101]	[129]	[84]	[68]	[131]	[133]	[134]	[70]	**Our work**
UAV's trajectory design	✓				✓				✓		✓
Sum-rate maximisation	✓		✓	✓	✓					✓	✓
Energy harvesting time optimisation			✓								✓
RIS phase shift optimisation	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Random users				✓	✓	✓		✓	✓	✓	✓
Reinforcement learning					✓	✓	✓		✓		✓
Deep neural networks						✓	✓		✓	✓	✓

2.5 Contributions and Novelty of the Thesis

As mentioned earlier, it is crucial to improve the decision-making time for meeting the stringent requirements of UAV-assisted wireless networks. In Chapter 3, we conceive emerging methods based on the DRL algorithm to design the UAV trajectory and resource allocation for maximising the EE and the network's sum-rate. We design an approach based on the DQL and the dueling DQL algorithm for a trade-off problem. Our proposed methods show the efficient trajectory and sum-rate while satisfying all the strict constraints of distance, flying time and communications. We boldly and explicitly contrast our proposed solution to the state-of-the-art in Table 2.1. Our proposed solution can work in a dynamic environment with randomly distributed and mobile users. The designated function is flexible for adjustment to adapt to several missions of fast deployment and sum-rate maximisation. Moreover, the aforementioned works mostly assume the perfect conditions of CSI and static users. In addition, the high computational complexity of these methods cause delays and make the system unrealistic for deploying in real-life applications. In Chapter 3, the UAV needs only the local information of its location to formulate the state space. In Chapter 4, we jointly optimise the UAV trajectory and the phase shift to maximise the network's sum-rate while we optimise the power allocation at the UAVs and phase-shift at the RIS to maximise the EE performance in Chap 5. In both chapters, due to the acquisition delay and the feedback overhead incurred during the mobility of UAV and users, obtaining a perfect CSI of the links between UAV and RIS as well as grounds devices is a formidable challenge. Thus, we assume that the perfect CSI can be achieved by the perfect channel estimation model at UAV and RIS. Our proposed solution is one of the first research works that investigates the efficiency of DRL for optimisation in RIS-assisted UAV communications with mobile users and the dynamic environment. Particularly, in Chapter 4, we conceive a method based on the DDPG algorithm and the PPO algorithm for optimising the UAV's trajectory, the EH time scheduling and the RIS's phase-shift matrix. We compare our proposed methods to the existing literature in Table 2.2. In Chapter 5, we propose a method based on the DDPG algorithm for single-agent learning and multi-agent learning to maximise the EE in RIS-assisted multi-UAV communications. Then, the PPO algorithm with a better sampling technique is conceived for improving the performance and convergence.

Chapter 3 is published as

K. K. Nguyen, T. Q. Duong, T. Do-Duy, H. Claussen, and L. Hanzo, "3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning," *IEEE Trans. Commun.*, 2021 (accepted).

Chapter 3

3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning

3.1 Introduction

As discussed in Chapter 1, wireless networks supported by UAVs constitute a promising technology for enhancing the network performance [135]. For example, in [16], UAVs were deployed to provide network coverage for people in remote areas and disaster zones. UAVs were also used for collecting data in a WSN [29]. Nevertheless, the benefits of UAV-aided wireless communication are critically dependent on the UAV's limited onboard power level. Thus, the resource allocation of UAV-aided wireless networks plays a pivotal role in approaching optimal performance. Yet, the existing works typically assume static environments [14,33,112] and often ignore the stringent flight-time constraints in real-life applications [29, 31, 136].

Recently, thanks to the real-time decision-making ability, the DRL algorithms have become a promising solution to solve the complicated problems in wireless networks [18, 27, 28, 30, 32, 46, 76–78, 82–84]. For example, the DRL was used for solving the resource management problems [120–123], path planning and deployment in the UAV-assisted wireless communications [32, 76–78, 84, 124, 125], caching problems [30], and interference management [76]. In the DRL algorithm, the neural networks are used to calculate the value function and policy approximation. Thus, after being trained, the agents can instantly choose the most appropriate action for maximising the designed reward.

In this chapter, we propose a novel DRL-aided UAV-assisted system for finding the optimal UAV path for maximising the joint reward function based on the shortest flight distance and the uplink transmission rate. Firstly, we define the objective of our UAV-aided system as to maximise the amount of data collected from the users with the shortest distance travelled. Our UAV-aided system is specifically designed for tackling the stringent constraints owing to the position of the destination, the UAV's limited flight time and the communication link's realistic constraints. Next, these challenges are tackled by conceiving bespoke DRL techniques for solving the above problem. To elaborate, the area is divided into a grid to enable fast convergence. Following its training, the UAV can have the autonomy to make a decision concerning its next action at each position in the area, hence eliminating the need for human navigation. This makes our UAV-aided system more reliable, practical and optimises the resource requirements. In the simulation results, a pair of scenarios are considered relying either on three or five clusters for quantifying the efficiency of our novel DRL techniques in terms of both the sum-rate, the trajectory and the associated time. A convincing 3D trajectory visualisation is also provided. Finally, but most importantly, it is demonstrated that our DRL techniques approach the performance of the optimal "genie-solution" associated with the perfect knowledge of the environment.

The rest of this chapter is organised as follows. In Section 3.2, we describe our data collection system model and the problem formulation of IoT networks relying on UAVs. Then, the Q-learning algorithm is presented in Section 3.3. Deep Q-learning (DQL) is employed for finding the best trajectory and for solving our data collection problem in Section 3.4. Furthermore, we use the dueling DQL algorithm of [137] for improving the system performance and convergence speed in Section 3.5. Next, we characterise the efficiency of the DRL techniques in Section 3.6. Finally, in Section 3.7, we summarise our findings and discuss our future research.

3.2 System Model and Problem Formulation

Consider a system consisting of a single UAV and M groups of users, as shown in Fig. 3.1, where the UAV relying on a single antenna visits all the clusters to cover all the users. The 3D coordinate of the UAV at time step t is defined as $X^t = (x_0^t, y_0^t, H_0^t)$. Each cluster consists of K users, which are unknown and distributed randomly within the coverage radius of C. The users are moving following the random walk model with the maximum velocity v. The position of the kth user in the mth cluster at time step t is defined as $X_{m,k}^t = (x_{m,k}^t, y_{m,k}^t)$. The UAV's objective is to find the best trajectory while covering all the users and to reach the dock upon completing its mission.

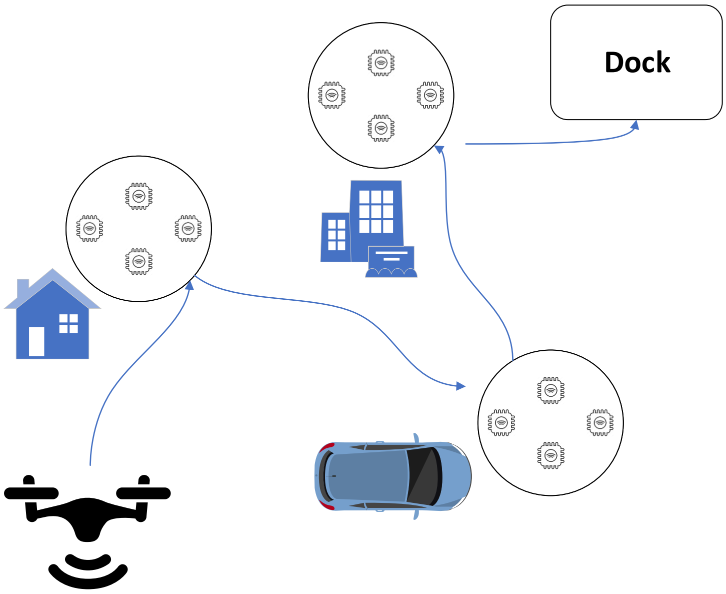


Figure 3.1: System model of UAV-aided IoT communications.

3.2.1 Observation model

The distance from the UAV to user k in cluster m at time step t is given by:

$$d_{m,k}^{t} = \sqrt{(x_0^t - x_{m,k}^t)^2 + (y_0^t - y_{m,k}^t)^2 + {H_0^t}^2}. \tag{3.1}$$

We assume that the communication channels between the UAV and users are dominated by line-of-sight (LoS) links; thus the channel between the UAV and the kth user in the mth cluster at time step t follows the free-space path loss model, which is represented as

$$\begin{aligned} h_{m,k}^{t} &= \beta_0 {d_{m,k}^{t}}^{-2} \\ &= \frac{\beta_0}{(x_0^t - x_{m,k}^t)^2 + (y_0 - y_{m,k}^t)^2 + {H_0^t}^2}, \end{aligned} \tag{3.2}$$

where the channel's power gain at a reference distance of $d = 1m$ is denoted by β_0.

The achievable throughput from the kth user in the mth cluster to the UAV at time t if the user belongs to the coverage of the UAV is defined as follows:

$$R^t_{m,k} = W \log_2 \left(1 + \frac{p^t_{m,k} h^t_{m,k}}{\sum_{i \neq m}^{M} \sum_{j}^{K} p^t_{i,j} h^t_{i,j} + \sum_{u \neq k}^{K} p^t_{m,u} h^t_{m,u} + \alpha^2} \right), \forall m, k, \tag{3.3}$$

where W and α^2 are the bandwidth and the noise power, respectively; $p_{m,k}$ is the transmit power at the kth user in the mth cluster. Then the total sum-rate over the T time step from the kth user in cluster m to the UAV is given by:

$$R_{m,k} = \int_0^T R^t_{m,k} dt, \forall m, k. \tag{3.4}$$

3.2.2 Game formulation

Both the current location and the action taken jointly influence the rewards obtained by the UAV; thus the trial-and-error based learning task of the UAV satisfies the Markov property. We formulate the associated Markov decision process (MDP) [85] as a 4 tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}_{ss'}, \mathcal{R} >$, where $\mathcal{S}$ is the state space of the UAV, $\mathcal{A}$ is the action space; $\mathcal{R}$ is the expected reward of the UAV and $\mathcal{P}_{ss'}$ is the probability of transition from state s to state s', where we have $s' = s^{t+1} | s = s^t$. Through learning, the UAV can find the optimal policy $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$ for maximising the reward $\mathcal{R}$. As the definition of RL, the UAV does not have any knowledge about the environment. We transfer a real-life application of the data collection in the UAV-assisted IoT networks into a digital form. Thus, the UAV only has local information of its location and the state is defined by the position of UAV.

We have also discretised the state and action space for learning. More particularly, we formulate the trajectory and data collection game of UAV-aided IoT networks as follows:

- *Agent*: The UAV acts like an agent interacting with the environment to find the peak of the reward.
- *State space*: We define the state space by the position of UAV as

  $$\mathcal{S} = \{x, y, H\}. \tag{3.5}$$

  At time step t, the state of the UAV is defined as $s^t = (x^t, y^t, H^t)$.
- *Action space*: The UAV at state s^t can choose an action a^t of the action space by following the policy at time-step t. By dividing the area into a grid, we can define the action space as follows:

  $$\mathcal{A} = \{\text{left}, \text{right}, \text{forward}, \text{backward}, \text{upward}, \text{downward}, \text{hover}\}. \tag{3.6}$$

  The UAV moves in the environment and begins collecting information when the users are in the coverage area of the UAV. When the UAV has sufficient information $R_{m,k} \geq r_{min}$ from the kth user in the mth cluster, that user will be marked as collected in this mission and may not be visited by the UAV again.
- *Reward function*: In joint trajectory and data collection optimisation, we design the reward function to be dependent on both the total sum-rate of ground users associated with the UAV plus the reward gleaned when the

UAV completes one route, which is formulated as follows:

$$R = \frac{\beta}{MK}\left(\sum_{m}^{M}\sum_{k}^{K} P(m,k)R_{m,k}\right) + \zeta R_{plus}, \tag{3.7}$$

where β and ζ are positive variables that represent the trade-off between the network's sum-rate and UAV's movement, which will be described in the sequel. Here, $P(m,k) \in \{0,1\}$ indicates whether or not user k of cluster m is associated with the UAV; R_{plus} is the acquired reward when the UAV completes its mission by reaching the final destination. On the other hand, the term $\frac{\sum_{m}^{M}\sum_{k}^{K} P(m,k)R_{m,k}}{MK}$ defines the average throughput of all users.

- *Probability*: We define $\mathcal{P}_{s^t s^{t+1}}(a^t, \pi)$ as the probability of transition from state s^t to state s^{t+1} by taking the action a^t under the policy π.

At each time step t, the UAV chooses the action a^t based on its local information to obtain the reward r^t under the policy π. Then the UAV moves to the next state s^{t+1} by taking the action a^t and starts collecting information from the users if any available node in the network satisfies the distance constraint. Meanwhile, the users in clusters also move to new positions following the random walk model with velocity v. Again, we use the DRL techniques to find the optimal policy π^* for the UAV to maximise the reward attained (3.7). Following the policy π, the UAV forms a chain of actions $(a^0, a^1, \ldots, a^t, \ldots, a^{final})$ to reach the landing dock.

Our target is to maximise the reward expected by the UAV upon completing a single mission during which the UAV flies from the initial position over the clusters and lands at the destination. Thus, we design the trajectory reward R_{plus} when the UAV reaches the destination in two different ways. Firstly, the binary reward function is defined as follows:

$$R_{plus} = \begin{cases} 1 & , \quad X_{final} \in X_{target} \\ 0 & , \quad \text{otherwise.} \end{cases}, \tag{3.8}$$

where X_{final} and X_{target} are the final position of UAV and the destination, respectively. The landing dock X_{target} is set by a zone of multiple grids. However, the UAV has to move a long distance to reach the final destination. It may also be trapped in a zone and cannot complete the mission. These situations lead to increased energy consumption and reduced convergence. Thus, we consider the value of R^t_{plus} in a different form by calculating the horizontal distance between the UAV and the final destination at time step t, yielding:

$$R^t_{plus} = \begin{cases} 1 & , \quad X_{final} \in X_{target} \\ \exp\left(\sqrt{(x_{target} - x^t_0)^2 + (y_{target} - y^t_0)^2}\right)^{-1} & , \quad \text{otherwise.} \end{cases} \tag{3.9}$$

When we design the reward function as in (3.9), the UAV is motivated to move ahead to reach the final destination. However, one of the disadvantages is that the UAV only moves forward. Thus, the UAV is unable to attain the best performance in terms of its total sum-rate in some environmental settings. We compare the performance of the two trajectory reward function definitions in Section 3.6 to evaluate the pros and cons of each approach.

In our work, we optimise the 3D trajectory of the UAV and data collection for the IoT network. Particularly, we have design the reward function by a trade-off game between the achievable sum-rate and the trajectory. Denote the flying path of the UAV from the initial point to final position by $X = (X_0, X_1, \ldots, X_{final})$, the agent needs to learn by iterating with the environment to find an optimal X. We have defined a trade-off value β and ζ to make our approach more adaptive and flexible. By modifying the value of β/ζ , the UAV adapts to several scenarios: a) fast deployment for emergency services with lower value of β/ζ, b) maximising the total sum-rate with higher value of β/ζ, and c) maximising the number of connections between the UAV and users. Depending on the specific problems, we can adjust the value of the trade-off parameters β, ζ to achieve the best performance. Thus, the game formulation is defined as follows:

$$\begin{aligned} \max R = \quad & \frac{\beta}{MK}\left(\sum_{m}^{M}\sum_{k}^{K} P(m,k)R_{m,k}\right) + \zeta R_{plus}, \\ s.t. \quad & X_{final} = X_{target}, \\ & d_{m,k} \leq d_{cons}, \\ & R_{m,k} \geq r_{min}, \\ & P(m,k) \in \{0,1\}, \\ & T \leq T_{cons} \\ & \beta \geq 0,\ \zeta \geq 0, \end{aligned} \tag{3.10}$$

where the term $X_{final} = X_{target}$ denotes the completed flying route when the final position of the UAV belongs to the destination zone. We have designed the reward function following this constraint with two functions: binary reward function in (3.8) and exponential reward function in (3.9). The terms $d_{m,k} \leq d_{cons}, R_{m,k} \geq r_{min}, P(m,k) \in \{0,1\}$ denote the communication constraints. Particularly, the distance constraint $d_{m,k} \leq d_{cons}$ indicates that the served (m,k)-user has a satisfying distance to the UAV. $P(m,k) \in \{0,1\}$ indicates whether or not user k of cluster m is associated with the UAV. $R_{m,k} \geq r_{min}$ denotes the minimum information collected during the flying path. All the constraints are integrated into the reward functions in the RL algorithm. The term $T \leq T_{cons}$ denotes the constraint about the flying time where T is the flying time of the UAV in a single mission and T_{cons} is the maximum flying time. The UAV needs to complete a route by reaching the destination zone before T_{cons}. If the UAV can not complete a route before T_{cons}, the $R_{plus} = 0$ as we defined in (3.8) and (3.9). We have the trade-off value in reward function $\beta \geq 0$, $\zeta \geq 0$. Those stringent constraints, such as the transmission distance, position and flight time make the optimisation problem more challenging. Thus, we propose DRL techniques for the UAV in order to attain optimal performance.

3.3 Q-learning algorithm for UAV-assisted IoT Networks

In this section, we introduce the fundamental concept of Q-learning, where the so-called value function is defined by a reward of the UAV at state s^t as follows:

$$V(s, \pi) = \mathbb{E}\Big[\sum_{t}^{T} \gamma \mathcal{R}^t(s^t, \pi)|s_0 = s\Big], \tag{3.11}$$

where $\mathbb{E}[.]$ represents an average of the number of samples and $0 \leq \gamma \leq 1$ denotes the discount factor.

In a finite game, there is always an optimal policy π^* that satisfies the Bellman optimality equation [86]

$$V^*(s,\pi) = V(s,\pi^*) = \max_{a\in\mathcal{A}} \Bigg[\mathbb{E}\Big[\mathcal{R}^t(s^t,\pi^*) \Big] + \gamma \sum_{s'\in S} \mathcal{P}_{ss'}(a,\pi^*) V(s',\pi^*) \Bigg]. \quad (3.12)$$

The action-value function is obtained, when the agent at state s^t takes action a^t and receives the reward r^t under the agent policy π. The optimal Q-value can be formulated as:

$$Q^*(s,a,\pi) = \mathbb{E}\Big[\mathcal{R}^t(s^t,\pi^*) \Big] + \gamma \sum_{s'\in S} \mathcal{P}_{ss'}(a,\pi^*) V(s',\pi^*). \quad (3.13)$$

The optimal policy π^* can be obtained from $Q^*(s,a,\pi)$ as follows:

$$V^*(s,\pi) = \max_{a\in\mathcal{A}} Q(s,a,\pi). \quad (3.14)$$

From (3.13) and (3.14), we have

$$\begin{aligned} Q^*(s,a,\pi) &= \mathbb{E}\Big[\mathcal{R}^t(s^t,\pi^*) \Big] + \gamma \sum_{s'\in S} \mathcal{P}_{ss'}(a,\pi^*) \max_{a'\in\mathcal{A}} Q(s',a',\pi), \\ &= \mathbb{E}\Big[\mathcal{R}^t(s^t,\pi^*) + \gamma \max_{a'\in\mathcal{A}} Q(s',a',\pi) \Big], \end{aligned} \quad (3.15)$$

where the agent takes the action $a' = a^{t+1}$ at state s^{t+1}.

Through learning, the Q-value is updated based on the available information as follows:

$$Q(s,a,\pi) = \ Q(s,a,\pi) + \alpha \Big[\mathcal{R}^t(s^t,\pi^*) + \gamma \max_{a'\in\mathcal{A}} Q(s',a',\pi) - Q(s,a,\pi) \Big], \quad (3.16)$$

where α denotes the updated parameter of the Q-value function.

In RL algorithms, it is challenging to balance the *exploration* and *exploitation* for appropriately selecting the action. The most common approach relies on the ϵ-greedy policy for the action selection mechanism as follows:

$$a = \begin{cases} \operatorname{argmax} Q(s, a, \pi) & \text{with} \quad \epsilon \\ \text{randomly} & \text{if} \quad 1 - \epsilon. \end{cases} \tag{3.17}$$

Upon assuming that each episode lasts T steps, the action at time step t is a^t that is selected by following the ϵ-greedy policy as in (3.17). The UAV at state s^t communicates with the user nodes from the ground if the distance constraint of $d_{m,k} \leq d_{cons}$ is satisfied. Following the information transmission phase, the user nodes are marked as collected users and may not be revisited later during that mission. Then, after obtaining the immediate reward $r(s^t, a^t)$ the agent at state s^t takes action a^t to move to state s^{t+1} as well as to update the Q-value function in (3.16). Each episode ends when the UAV reaches the final destination and the flight duration constraint is satisfied.

3.4 An Effective Deep Reinforcement Learning Approach for UAV-assisted IoT Networks

In this section, we conceive the DQL algorithm for trajectory and data collection optimisation in UAV-aided IoT networks. Q-learning technique typically falters for large state and action spaces due to its excessive Q-table size. Thus, instead of applying the Q-table in Q-learning, we use deep neural networks to represent the relationship between the action and state space. Furthermore, we employ a pair of techniques for stabilising the neural network's performance in our DQL algorithm as follows:

- *Experience replay buffer*: Instead of using current experience, we use a so-called replay buffer $\mathcal{B}$ to store the transitions (s, a, r, s') for supporting the neural network in overcoming any potential instability. When the buffer $\mathcal{B}$ is filled with the transitions, we randomly select a mini-batch of K samples for training the networks. The finite buffer size of $\mathcal{B}$ allows it to be always up-to-date, and the neural networks learn from the new samples.

- *Target networks*: If we use the same network to calculate the state-action value Q and the target network, the network can be shifted dramatically in the training phase. Thus, we employ a target network Q' for the target value estimator. After a number of iterations, the parameters of the target network Q' will be updated by the network Q.

The UAV starts from the initial position and interact with the environment to find the proper action in each state. The agent chooses the action a^t following current policy at state s^t. By executing the action a^t, the agent receives the response from the environment with reward r^t and new state s^{t+1}. After each time step, the UAV has a new position and the environment is changed with moving users. The obtained transitions are stored into a finite memory buffer and used for training the neural networks.

The neural network parameters are updated by minimising the loss function defined as follows:

$$\mathbb{L}(\theta) = \mathbb{E}_{s,a,r,s'}\left[\left(y^{DQL} - Q(s, a; \theta)\right)^2\right], \tag{3.20}$$

Algorithm 1 The deep Q-learning algorithm for trajectory and data collection optimisation in UAV-aided IoT networks.

```
Initialise the network $Q$ and the target network $Q'$ with the random parameters $\theta$ and $\theta'$, respectively
Initialise the replay memory pool $\mathcal{B}$
for episode $= 1, \dots, L$ do
    Receive initial observation state $s^0$
    while $X_{final} \notin X_{target}$ or $T \leq T_{cons}$ do
        Obtain the action $a^t$ of the UAV according to the $\epsilon$-greedy mechanism (3.17)
        Execute the action $a^t$ and estimate the reward $r^t$ according to (3.7)
        Observe the next state $s^{t+1}$
        Store the transition $(s^t, a^t, r^t, s^{t+1})$ in the replay buffer $\mathcal{B}$
        Randomly select a mini-batch of $K$ transitions $(s^k, a^k, r^k, s^{k+1})$ from $\mathcal{B}$
        Update the network parameters using gradient descent to minimise the loss
            $\mathbb{L}(\theta) = \mathbb{E}_{s,a,r,s'}\left[\left(y^{DQL} - Q(s,a;\theta)\right)^2\right], \quad (3.18)$
        The gradient update is
            $\nabla_\theta \mathbb{L}(\theta) = \mathbb{E}_{s,a,r,s'}\left[\left(y^{DQL} - Q(s,a;\theta)\right)\nabla_\theta Q(s,a;\theta)\right], \quad (3.19)$
        Update the state $s^t = s^{t+1}$
        Update the target network parameters after a number of iterations as $\theta' = \theta$
    end while
end for
```

where θ is a parameter of the network Q and we have

$$y = \begin{cases} r^t & \text{if terminated at } s^{t+1} \\ r^t + \gamma \max_{a' \in \mathcal{A}} Q'(s', a'; \theta') & \text{otherwise.} \end{cases} \tag{3.21}$$

The details of the DQL approach in our joint trajectory and data collection trade-off game designed for UAV-aided IoT networks are presented in Algorithm 1 where L denotes the number of episodes. Moreover, in this chapter, we design the reward obtained in each step to assume one of two different forms and compare them in our simulation results. Firstly, we calculate the difference between the current and the previous reward of the UAV as follows:

$$r_1^t(s^t, a^t) = r^t(s^t, a^t) - r^{t-1}(s^{t-1}, a^{t-1}). \tag{3.22}$$

Secondly, we design the total episode reward as the accumulation of all immediate rewards of each step within one episode as

$$r_2^t(s^t, a^t) = \sum_{i=0}^{t} r_1^t(s^t, a^t). \tag{3.23}$$

3.5 Deep Reinforcement Learning Approach for UAV-assisted IoT networks: A Dueling Deep Q-learning Approach

The standard Q-learning algorithm often falters due to the over-supervision of all the state-action pairs [137]. On the other hand, it is unnecessary to estimate the value of each action choice in a particular state. For example, in our environment setting, the UAV has to consider moving either to the left or to the right when it hits the boundaries. Thus, we can improve the convergence speed by avoiding visiting all state-action pairs. Instead of using Q-value function of the conventional DQL algorithm, the dueling neural network of [137] is introduced for

Algorithm 2 The dueling deep Q-learning algorithm for trajectory and data collection optimisation in UAV-aided IoT networks.

```
Initialise the network $Q$ and the target network $Q'$ with the random parameters, $\theta$ and $\theta'$, respectively
Initialise the replay memory pool $\mathcal{B}$
for episode $= 1, \dots, L$ do
    Receive the initial observation state $s^0$
    while $X_{final} \notin X_{target}$ or $T \leq T_{cons}$ do
        Obtain the action $a^t$ of the UAV according to the $\epsilon$-greedy mechanism (3.17)
        Execute the action $a^t$ and estimate the reward $r^t$ according to (3.7)
        Observe the next state $s^{t+1}$
        Store the transition $(s^t, a^t, r^t, s^{t+1})$ in the replay buffer $\mathcal{B}$
        Randomly select a mini-batch of $K$ transitions $(s^k, a^k, r^k, s^{k+1})$ from $\mathcal{B}$
        Estimate the Q-value function by combining the two streams as follows:
            $$Q(s, a;\ \theta, \theta_A, \theta_V) = V(s; \theta_V) + \left( A(s, a; \theta_A) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a'; \theta_A) \right). \quad (3.24)$$
        Update the network parameters using gradient descent to minimise the loss
            $$\mathbb{L}(\theta) = \mathbb{E}_{s,a,r,s'} \left[ \left( y^{DuelingDQL} - Q(s, a; \theta, \theta_A, \theta_V) \right)^2 \right], \quad (3.25)$$
        where
            $$y^{DuelingDQL} = r^t + \gamma \max_{a' \in \mathcal{A}} Q'(s', a'; \theta', \theta_A, \theta_V). \quad (3.26)$$
        Update the state $s^t = s^{t+1}$
        Update the target network parameters after a number of iterations as $\theta' = \theta$
    end while
end for
```

improving the convergence rate and stability. The so-called advantage function $A(s, a) = Q(s, a) - V(s)$ related both to the value function and to the Q-value function describes the importance of each action related to each state.

The idea of a dueling deep network is based on a combination of two streams of the value function and the advantage function used for estimating the single output Q-function. One of the streams of a fully-connected layer estimates the value function $V(s;\theta_V)$, while the other stream outputs a vector $A(s,a;\theta_A)$, where θ_A and θ_V represent the parameters of the two networks. The Q-function can be obtained by combining the two streams' outputs as follows:

$$Q(s,a;\theta,\theta_A,\theta_V) = V(s;\theta_V) + A(s,a;\theta_A). \tag{3.27}$$

Equation (3.27) applies to all (s,a) instances; thus, we have to replicate the scalar $V(s;\theta_V)$, $|\mathcal{A}|$ times to form a matrix. However, $Q(s,a;\theta,\theta_A,\theta_V)$ is a parameterised estimator of the true Q-function; thus, we cannot uniquely recover the value function V and the advantage function A. Therefore, (3.27) results in poor practical performances when used directly. To address this problem, we can map the advantage function estimator to have no advantage at the chosen action by combining the two streams as follows:

$$Q(s,a;\theta,\theta_A,\theta_V) = V(s;\theta_V) + \Bigg(A(s,a;\theta_A) - \max_{a' \in |\mathcal{A}|} A(s,a';\theta_A)\Bigg). \tag{3.28}$$

Intuitively, for $a^* = \operatorname{argmax}_{a' \in \mathcal{A}} Q(s,a';\theta,\theta_A,\theta_V) = \operatorname{argmax}_{a' \in \mathcal{A}} A(s,a';\theta_A)$, we have $Q(s,a^*;\theta,\theta_A,\theta_V) = V(s;\theta_V)$. Hence, the stream $V(s;\theta_V)$ estimates the value function and the other streams is the advantage function estimator. We can transform (3.28) using an average formulation instead of the max operator as follows:

$$Q(s,a;\theta,\theta_A,\theta_V) = V(s;\theta_V) + \left(A(s,a;\theta_A) - \frac{1}{|\mathcal{A}|}\sum_{a'} A(s,a';\theta_A) \right). \tag{3.29}$$

Now, we can solve the problem of identifiability by subtracting the mean as in (3.29). Based on (3.29), we propose a dueling DQL algorithm for our joint trajectory and data collection problem in UAV-assisted IoT networks relying on Algorithm 2. Note that estimating $V(s;\theta_V)$ and $A(s,a;\theta_A)$ does not require any extra supervision and they will be computed automatically.

3.6 Simulation Results

In this section, we present our simulation results characterising the joint optimisation problem of UAV-assisted IoT networks. To highlight the efficiency of our proposed model and the DRL methods, we consider a pair of scenarios: a simple having three clusters, and a more complex one with five clusters in the coverage area. We use Tensorflow 1.13.1 [138] and the Adam optimiser [139] for training the neural networks. In this paper, we set the maximum value of β/ζ not too large because we prefer the completion of a mission. The maximum value is set to $\beta/\zeta = 4/1$. All the other parameters are provided in Table 3.1.

In Fig. 3.2, we present the trajectory obtained after training using the DQL algorithm in the 5-cluster scenario. The green circle and blue dots represent the clusters' coverage and the user nodes, respectively. The red line and black line in the figure represent the UAV's trajectory based on our method in (3.8) and (3.9), respectively. The UAV starts at $(0,0)$, visits about 40 users, and lands

Table 3.1: Simulation parameters in Chapter 3.

Parameters	Value
Bandwidth (W)	1 MHz
UE transmission power	0.1 W
The start position of UAV	$(0, 0, 200)$
Discounting factor	$\gamma = 0.9$
Max number of users per cluster	10
Rate constraint $r_{\min} = 1$bits/s/Hz	
Noise power	$\alpha^2 = -110$dBm
The reference channel power gain	$\beta_0 = -50$dB
Path-loss exponent	2

at the destination that is denoted by the black star. In a complex environment setting, it is challenging to expect the UAV to visit all users, while satisfying the flight-duration and power level constraints.

3.6.1 Expected reward

We compare our proposed algorithm with opitimal performance and the Q-learning algorithm. The optimal performance scenario is based on the assumptions of knowing the IoT devices' position and unlimited power level of the UAV. For comparison, we run the algorithm five times in five different environmental settings and take the average to draw the figures. Firstly, we compare the reward obtained following (3.7). Let us consider the 3-cluster scenario and $\beta/\zeta = 2 : 1$ in Fig. 3.3a, where the DQL and the dueling DQL algorithms using the exponential function (3.9) reach the best performance. When using the exponential trajectory design function (3.9), the performance converges faster than that of the DQL and of the dueling DQL methods using the binary trajectory function

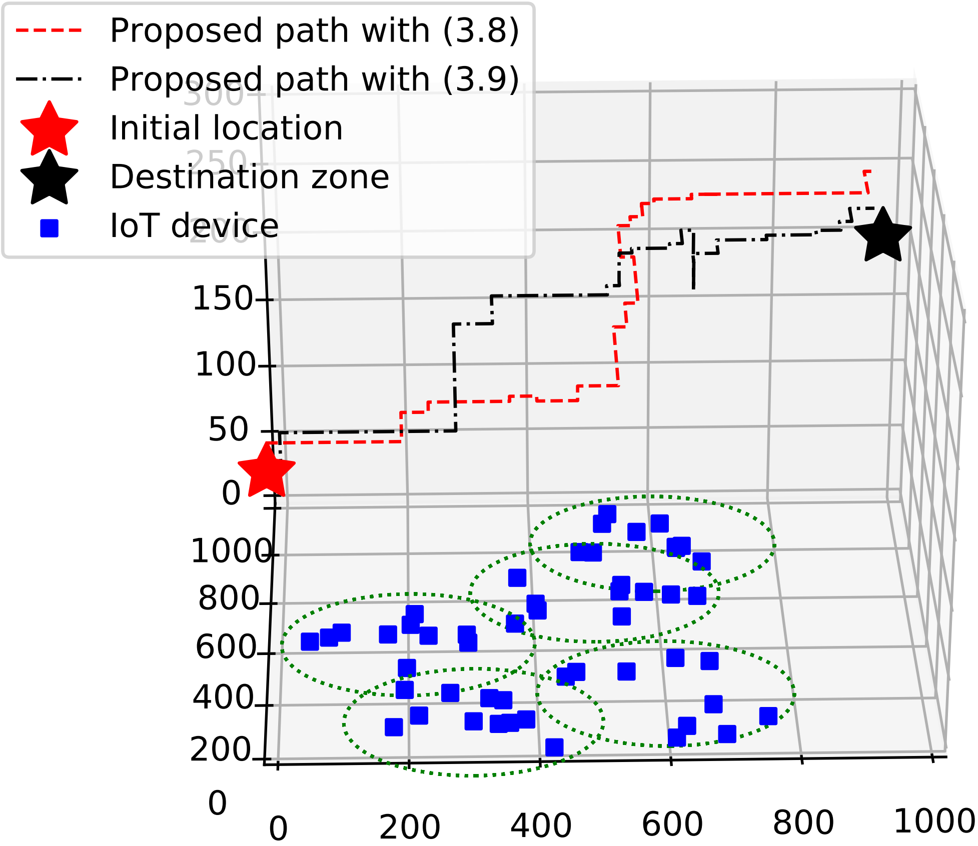
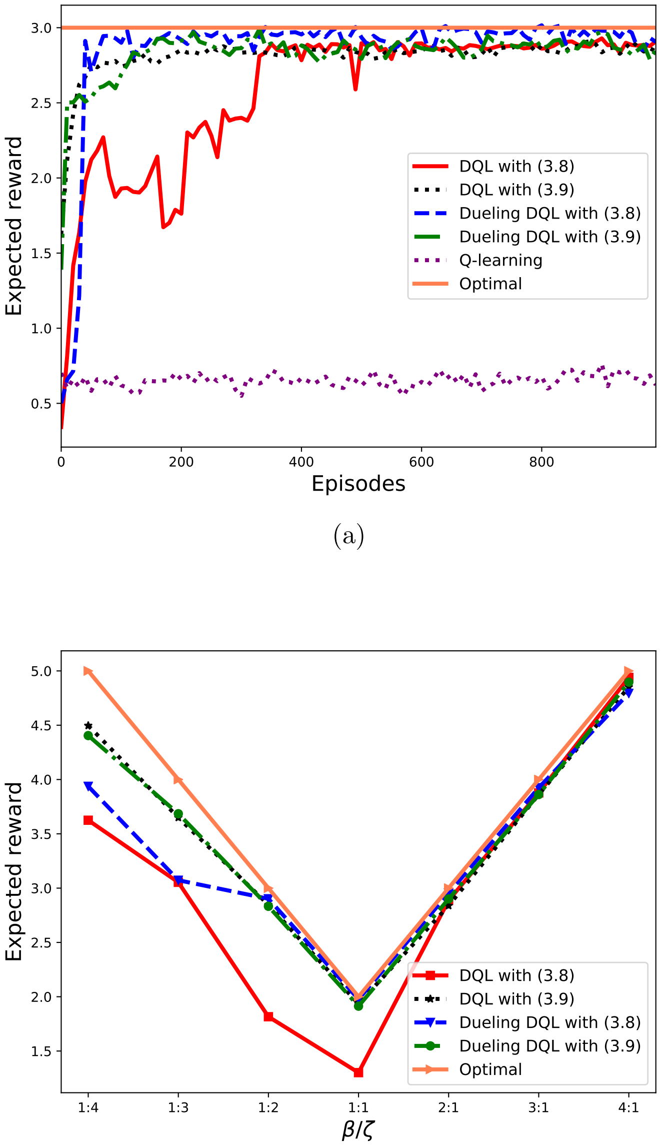


Figure 3.2: Trajectory obtained by using our dueling DQL algorithm.

(3.8). The performance of using the Q-learning algorithm is worst. In addition, in Fig. 3.3b, we compare the performance of the DQL and dueling DQL techniques using different β/ζ values. The average performance of the dueling DQL algorithm is better than that of the DQL algorithm. Furthermore, the results of using the exponential function (3.9) are better than that of the ones using the binary function (3.8). When the value of $\beta/\zeta \geq 1 : 2$, the performance achieved by the DQL and dueling DQL algorithm close to the optimal performance.

Furthermore, we compare the rewards obtained by the DQL and dueling DQL

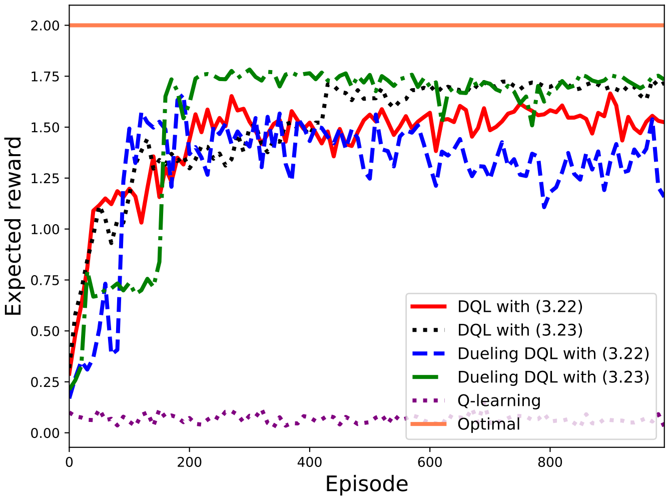


(a)

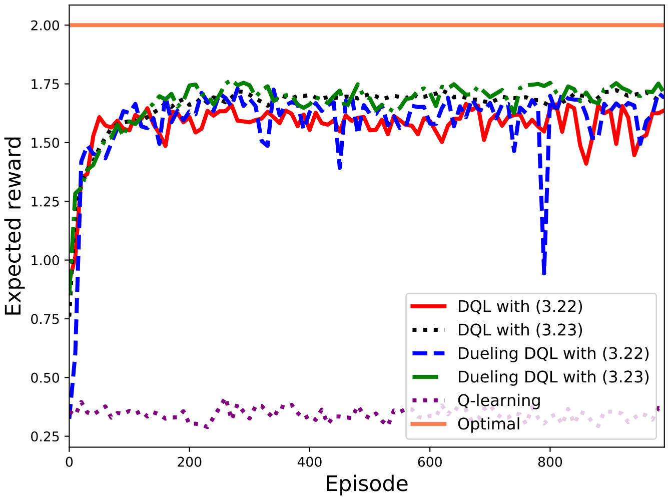
(b)

Figure 3.3: The performance when using the DQL and dueling DQL algorithms with 3 clusters while considering different β/ζ values.

algorithms in complex scenarios with 5 clusters and 50 user nodes in Fig. 3.4. The performance of using the episode reward (3.23) is better than that using the



(a) With (3.8)



(b) With (3.9)

Figure 3.4: The expected reward when using the DQL and dueling DQL algorithms with 5-cluster scenario.

immediate reward (3.22) in both trajectory designs relying on the DQL and dueling DQL algorithms. In Fig. 3.4a, we compare the performance in conjunction with the binary trajectory design while in Fig. 3.4b the exponential trajectory design is considered. For $\beta/\zeta = 1 : 1$, the rewards obtained by the DQL and dueling DQL are similar and stable after about 400 episodes. When using the exponential function (3.9), the dueling DQL algorithm reaches the best performance and close to the optimal performance. Moreover, the convergence of the dueling DQL technique is faster than that of the DQL algorithm. In both reward definitions, the Q-learning with (3.22) shows the worst performance.

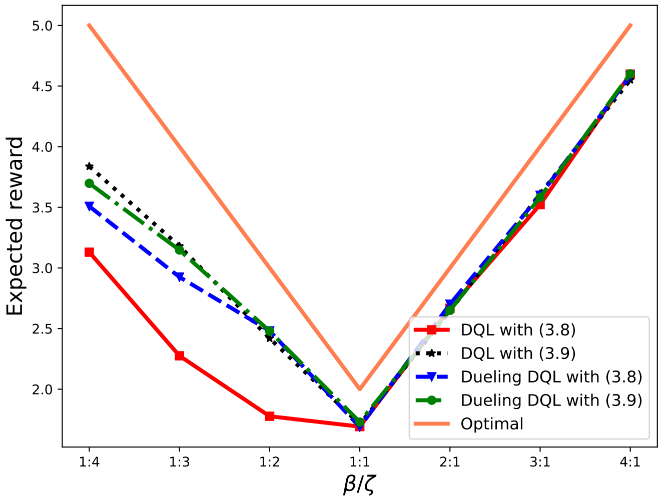
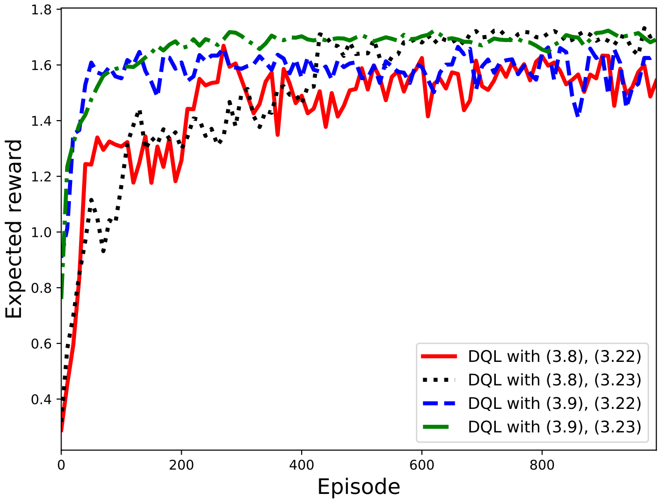
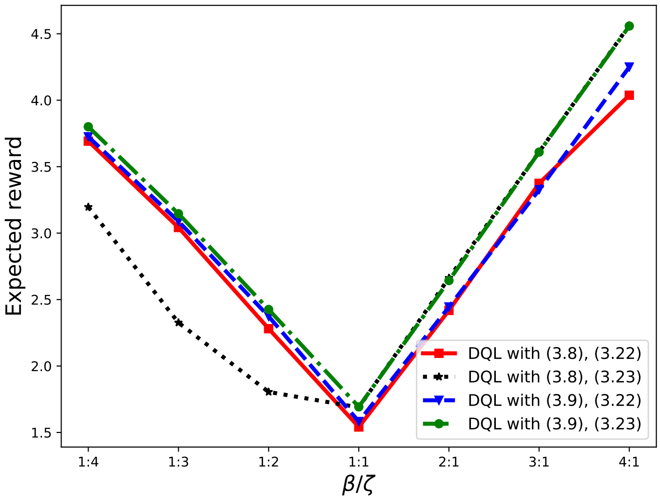


Figure 3.5: The performance when using the DQL and dueling DQL algorithms with 5 clusters and different β/ζ values.

In Fig. 3.5, we compare the performance of the DQL and of the dueling DQL algorithms while considering different β/ζ parameter values. The dueling DQL algorithm shows better performance for all the β/ζ pair values, exhibiting better rewards. Additionally, when using the exponential function (3.9), both proposed algorithms show better performance than the ones using the binary function (3.8) if $\beta/\zeta \leq 1 : 1$, but it becomes less effective when β/ζ is set higher. Again, we achieve a near-optimal solution while we consider a complex environment without knowing the IoT nodes' position and mobile users. It is challenging to expect the UAV to visit all IoT nodes with limited flying power and duration.



(a)



(b)

Figure 3.6: The expected reward when using the DQL algorithm with 5 clusters and different reward function settings.

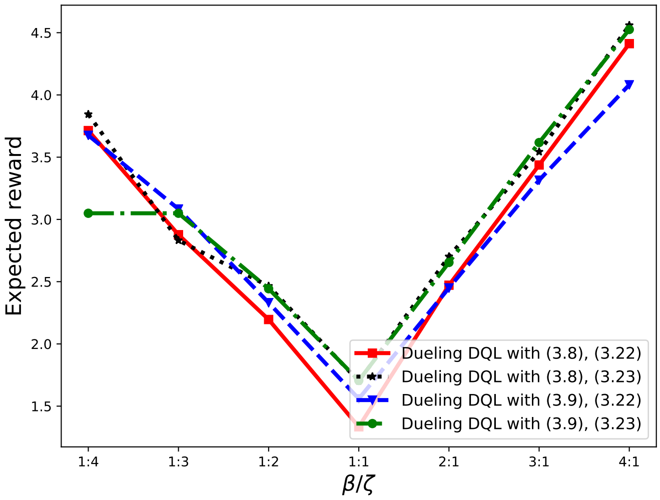
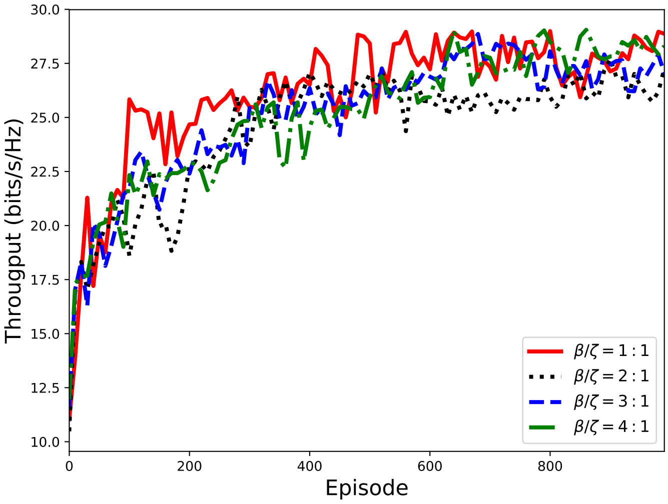
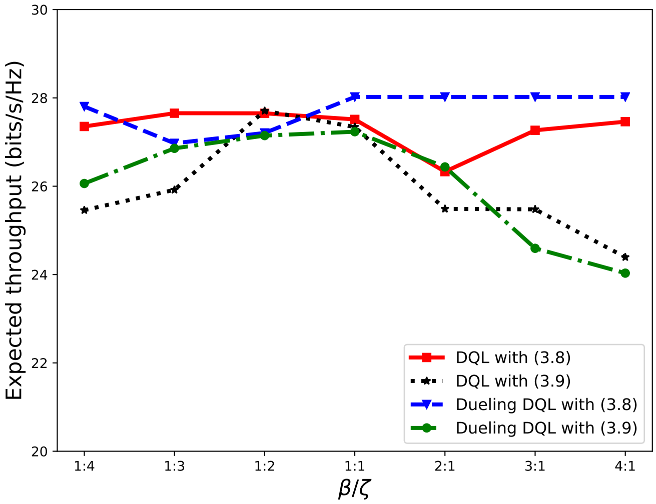


Figure 3.7: The performance when using the dueling DQL with 5 clusters, and different β/ζ values

We compare the performance of the DQL and of the dueling DQL algorithm using different reward function setting in Fig. 3.6 and in Fig. 3.7, respectively. The DQL algorithm reaches the best performance when using the episode reward (3.23) in Fig. 3.6a while the fastest convergence speed can be achieved by using the exponential function (3.9). When $\beta/\zeta \geq 1:1$, the DQL algorithm relying on the episode function (3.23) outperforms the ones using the immediate reward function (3.22) in Fig. 3.6b. The reward (3.7) using the exponential trajectory design (3.9) has a better performance than that using the binary trajectory design (3.8) for all the β/ζ values. The similar results are shown when using the dueling DQL algorithm in Fig. 3.7. The immediate reward function (3.22) is less effective than the episode reward function (3.23).



(a) With (3.8)



(b)

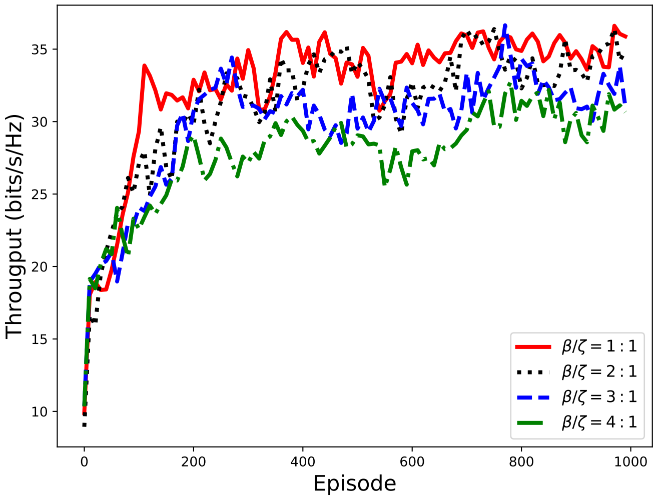
Figure 3.8: The network's sum-rate when using the DQL and dueling DQL algorithms with 3 clusters.

3.6.2 Throughput comparison

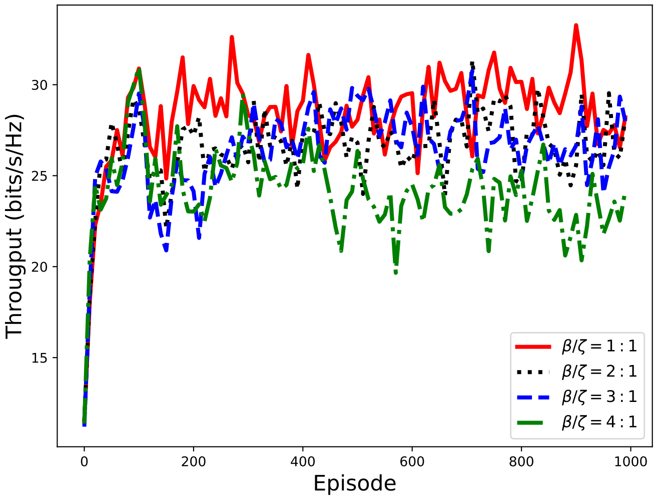
In (3.7), we consider two elements: the trajectory cost and the average throughput. In order to quantify the communication efficiency, we compare the total throughput in different scenarios. In Fig. 3.8, the performances of the DQL algorithm associated with several β/ζ values are considered while using the binary trajectory function (3.8), the episode reward (3.23) and 3 clusters. The throughput obtained for $\beta/\zeta = 1:1$ is higher than that of the others and when β increases, the performance degrades. However, when comparing with the Fig. 3.3b, we realise that in some scenarios the UAV was stuck and could not find the way to the destination. That leads to increased flight time spent and distance travelled. More details are shown in Fig. 3.8b, where we compare the expected throughput of both the DQL and dueling DQL algorithms. The best throughput is achieved when using the dueling DQL algorithm with $\beta/\zeta = 1:1$ in conjunction with (3.8), which is higher than the peak of the DQL method with $\beta/\zeta = 1:2$.

In Fig. 3.9, we compare the throughput of different techniques in the 5-cluster scenario. Let us now consider the binary trajectory design function (3.8) in Fig. 3.9a, where the DQL algorithm achieves the best performance using $\beta/\zeta = 1:1$ and $\beta/\zeta = 2:1$. There is a slight difference between the DQL method having different settings, when using exponential the trajectory design function (3.9), as shown in Fig. 3.9b.

In Fig. 3.10 and Fig. 3.11, we compare the throughput of different β/ζ pairs. The DQL algorithm reaches the optimal throughput with the aid of trial-and-learn methods, hence it is important to carefully design the reward function to



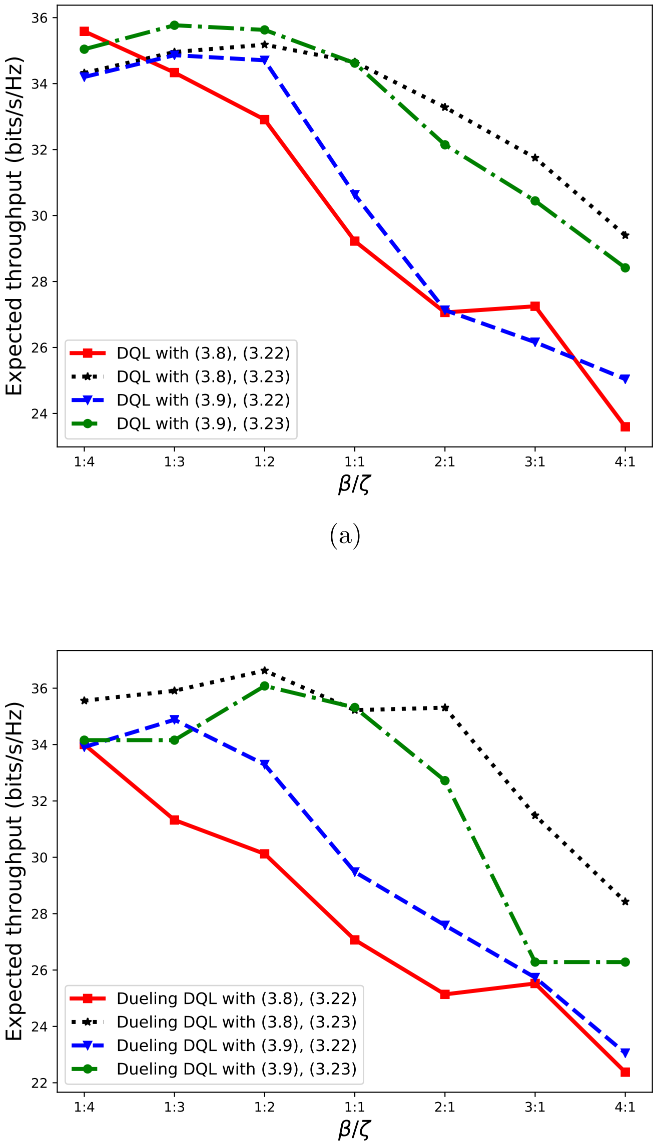
(a) With (3.8), (3.23)



(b) With (3.9), (3.23)

Figure 3.9: The obtained total throughput when using the DQL algorithm with 5 clusters.

avoid excessive offline training. As shown in Fig. 3.10, the DQL and dueling DQL algorithm exhibit reasonable stability for several $\beta/\zeta \leq 1:1$ pairs as well



(a)

(b)

Figure 3.10: The obtained throughput when using the DQL and dueling DQL algorithms in 5-cluster scenario.

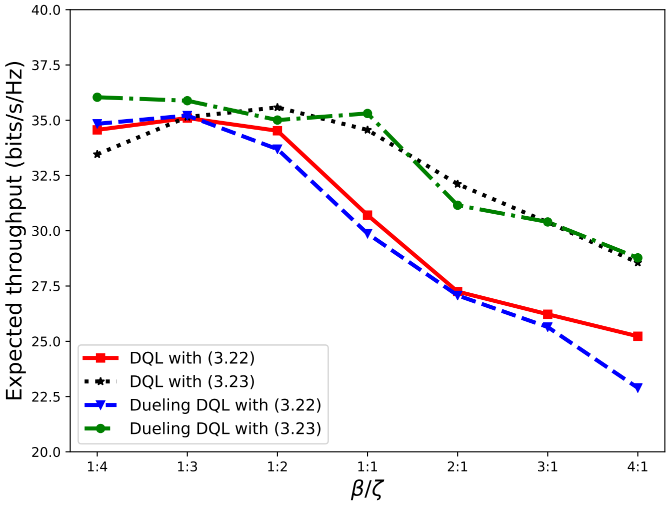
as reward functions. While we can achieve the similar expected reward with different reward setting in Fig. 3.6, the throughput is degraded when the β/ζ

increases. In contrast, with higher β values, the UAV can finish the mission faster. It is a trade-off game when we can choose an approximate β/ζ value for our specific purposes. When we employ the DQL and the dueling DQL algorithms with the episode reward (3.23), the throughput attained is higher than that using the immediate reward (3.22) with different β/ζ values.

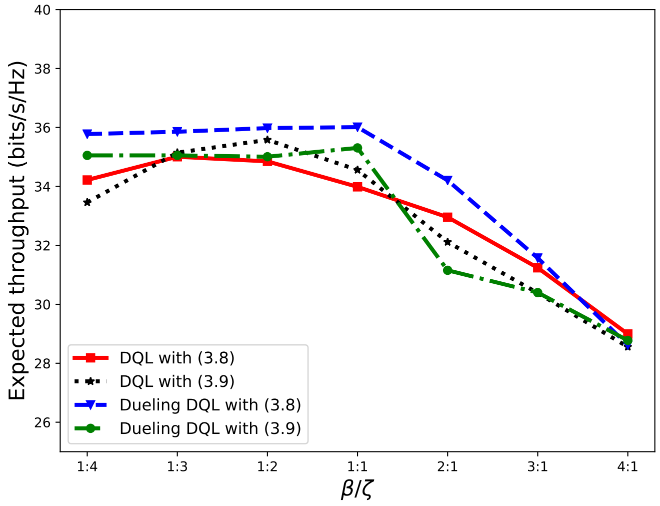
Furthermore, we compare the expected throughput of the DQL and of the dueling DQL algorithm when using the exponential trajectory design (3.9) in Fig. 3.11a and the episode reward (3.23) in Fig. 3.11b. In Fig. 3.11a, the dueling DQL method outperforms the DQL algorithm for almost all β/ζ values in both function (3.22) and (3.23). When we use the episode reward (3.23), the obtained throughput is stable with different β/ζ values. The throughput attained by using the exponential function (3.9) is lower than that using the binary trajectory (3.8) and by using the episode reward (3.23) is higher than that using the immediate reward (3.22). We can achieve the best performance when using the dueling DQL algorithm with (3.9) and (3.23). However, in some scenarios, we can achieve the better performance with different algorithmic setting as we can see in Fig. 3.8b and Fig. 3.10a. Thus, there is a trade-off governing the choice of the algorithm and function design.

3.6.3 Parametric Study

In Fig. 3.12, we compare the performance of our DQL technique using different *exploration* parameters γ and ϵ values in our ϵ-greedy method. The DQL algorithm achieves the best performance with the discounting factor of $\gamma = 0.9$ and $\epsilon = 0.9$ in the 5-cluster scenario of Fig. (3.12). Balancing the *exploration* and



(a) With (3.9)



(b) With (3.23)

Figure 3.11: The expected throughput when using the DQL and dueling DQL algorithms with 5 clusters.

exploitation as well as the action chosen is quite challenging, in order to maintain a steady performance of the DQL algorithm. Based on the results of Fig. 3.12,

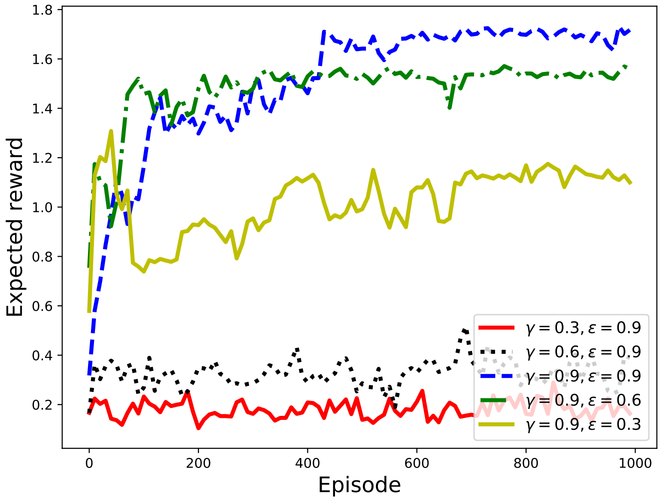


Figure 3.12: The performance when using the DQL algorithm with different discount factors, γ, and exploration factors, ϵ.

we opted for $\gamma = 0.9$ and $\epsilon = 0.9$ for our algorithmic setting.

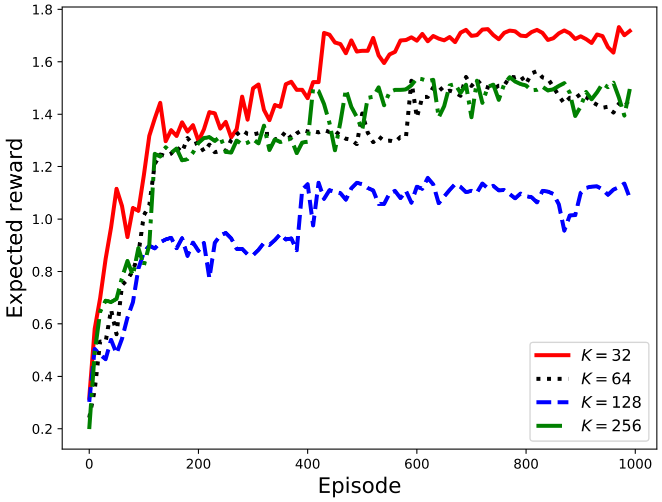


Figure 3.13: The performance when using the DQL algorithm in 5-cluster scenario and different batch sizes, K.

Next, we compare the expected reward of different mini-batch sizes, K. In the 5-cluster scenario of Fig. 3.13, the DQL achieves the optimal performance with a batch size of $K = 32$. There is a slight difference in terms of convergence speed where batch size $K = 32$ is the fastest. Overall, we set the mini-batch size to $K = 32$ for our DQL algorithm.

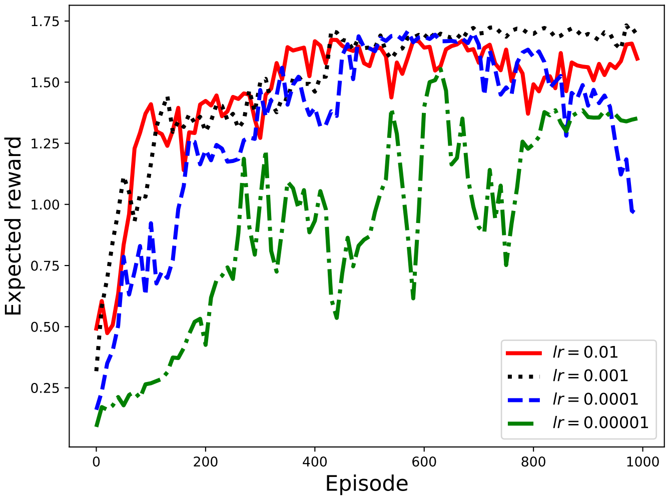


Figure 3.14: The performance when using DQL algorithm with different learning rate, lr.

Fig. 3.14 shows the performance of the DQL algorithm with different learning rates in updating the neural networks parameters while considering the scenarios of 5 clusters. When the learning rate is as high as $\alpha = 0.01$, the pace of updating the network may result the fluctuating performance. Moreover, when $\alpha = 0.0001$ or $\alpha = 0.00001$ the convergence speed is slower and may be stuck in a local optimum instead of reaching the global optimum. Thus, based on our experiments, we opted for the learning rate of $\alpha = 0.001$ for the algorithms.

3.7 Conclusion

In this chapter, the DRL technique has been proposed to jointly optimise the flight trajectory and data collection performance of UAV-assisted IoT networks. The optimisation game has been formulated to balance the flight time and total throughput while guaranteeing the quality-of-service constraints. Bearing in mind the limited UAV power level and the associated communication constraints, we proposed a DRL technique for maximising the throughput while the UAV has to move along the shortest path to reach the destination. Both the DQL and dueling DQL techniques having a low computational complexity have been conceived. Our simulation results showed the efficiency of our techniques both in simple and complex environmental settings.

Chapter 4 is published as

K. K. Nguyen, A. Masaracchia, Vishal Sharma, H. V. Poor, and T. Q. Duong, "RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning," *IEEE J. Selected Topics in Signal Process.*, 2021 (under review).

Chapter 4

RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning

4.1 Introduction

UAVs have recently drawn considerable attention thanks to their agile mobility and cost-effectiveness. UAVs have been used for geometry monitoring, disaster relief [16], emergency services, and wireless networks [18]. UAVs can be deployed at sporting events or in rescue missions to provide and enhance connectivity to the users. UAVs are also used as data collectors that fly to remote areas to collect sensor data [128]. However, restrictions regarding flying time and on-board processing ability are bottlenecks that must be dealt with in unexpected environments and complicated missions.

Reconfigurable intelligent surface (RIS) has emerged as a promising technology for future wireless networks. The arrival signal at a RIS is reflected toward the receiver by the RIS's passive elements operated by a module controller. The received signal at the users is composed of elements from the direct channel and the reflective link. It helps to increase the signal quality and reduce interference. The RIS is usually deployed in high locations such as buildings to reduce the cost of establishing a new station. However, the optimisation of RIS performance is still challenging due to a large number of elements and the processing ability of the controller.

One area in which UAVs can be useful is Internet-of-Things (IoT) applications. Not only can they provide communication coverages, but, since many IoT devices are energy-limited, they can also be sources of power for such devices through downlink power transfer. A downlink power transfer and uplink information transmission protocol can implemented in two phases: wireless power transfer (WPT) and wireless information transmission (WIT). In the first phase, the IoT devices harvest energy from a base station (BS) or from the UAV. The harvested energy is then used for transmitting local information to receivers or back to the UAV and the BS. By using such a downlink power transfer and uplink information transmission protocol, the IoT devices can obtain the energy to establish and maintain communication with the BS and the UAV.

Machine learning is an effective tool for optimising the performance of large-scale networks under dynamic environments. One of the approaches is deep reinforcement learning (DRL), which is a combination of reinforcement learning and neural networks. In wireless networks, DRL algorithms are used for maximising the network performance, reducing power consumption and improving the processing time for real-time applications [18, 27, 28]. DRL algorithms are powerful in wireless networks because the agents do not need pre-collected data for training. Rather, DRL agents interact with their environment and establish training samples for the responses in those interactions. The neural networks are trained by up-to-date state transitions to adjust their parameters for maximising a designated reward. Then, the trained networks are deployed for real-time prediction. However, when deploying the optimisation algorithm with DRL into RIS-assisted UAV communications, previous works assumed the perfect condition of the environment, flat fading channels, static users, and perfect CSI, which are unrealistic and infeasible for real-life applications.

In this chapter, we consider the IoT wireless networks with the support of a UAV, and one RIS, and employ the downlink power transfer and uplink information transmission protocol for maximising the total network's sum-rate. In particular, we adopt the harvest-then-transmit protocol, which means the IoT devices use all the harvested energy in the first phase for transmitting during the remaining time. Then, two DRL algorithms are deployed for solving the problem in RIS-assisted UAV communications. Firstly, we conceive a system model of UAV-assisted IoT wireless power transfer with the support of a RIS. The IoT devices harvest energy in the downlink and transmit information in the uplink to the UAVs. To characterise the agility of UAVs in supporting the energy harvesting (EH) and information transmission of IoT devices, we consider two scenarios of UAVs. In the first scenario, the UAV is hovering at the centre of the cluster and provides energy to the IoT devices. The RIS helps alleviate the uplink interference when the IoT devices transmit their information to the UAV. In the second scenario, the UAV is deployed in an initial location and required to find a better location for communication. In each location of the UAV's flying trajectory, the EH time scheduling and the RIS's phase shift matrix are optimised for maximising the network throughput performance. Next, for the defined problem, we formulate a Markov decision process (MDP) [86] with the definition of the state space, action space and the reward function. Then, we propose a method based on deep deterministic policy gradient (DDPG) and proximal policy optimisation algorithm (PPO) for solving the maximisation game. Furthermore, the delay when using a mathematical model and in the centralised learning is huge for real-time use cases. To overcome these aforementioned shortcomings, in this chapter, we also propose a parallel learning for reducing the information transmission requirement of the centralised approach. Finally, our results suggest that with the support of the RIS, a better connection is established and the overall performance is significantly improved.

In the remaining of this chapter, we present the system model of the UAV-assisted wireless communication with the support of the RIS and problem formulation in Section 4.2. The MDP for maximising the network throughput problem is introduced for the hovering UAV scenario in Section 4.3. In Section 4.4, the DDPG algorithm is deployed for continuous control of the UAV's trajectory, EH time scheduling and the phase shift matrix of the RIS. The PPO technique with the clipping method is introduced in Section 4.5 for improving the network throughput. The simulation results are presented in Section 4.6 to illustrate the efficiency of our proposed methods compared with other baseline schemes. Some existing problems and potential future research topics for the RIS and the UAV in real-life applications are discussed in Section 4.7.

4.2 System Model and Problem Formulation

We consider that the system includes one single-antenna UAV and N ground IoT devices distributed randomly. However, there are some practical scenarios where IoT devices are located in a crowded area with surrounding obstacles and objects. In such a complex environment, IoT devices suffer high attenuation and severe path loss. In this case, the RIS is also installed at the wall of a tall building to enhance the communication quality by reflecting signals from the UAV to the IoT devices. Here, we deploy a RIS composed of K elements to enhance the network performance. The 3D coordinate of the UAV at the time step t is $X^t_{UAV} = (x^t_{UAV}, y^t_{UAV}, z^t_{UAV})$. In this paper, we consider the fixed attitude of the UAV at H_{UAV}. The location of the nth IoT devices at time step t is $X^t_n = (x^t_n, y^t_n)$ with $n = 1, \ldots, N$. The position of the RIS component $k \in K$ at time step t is (x^t_k, y^t_k, z^t_k). In this paper, we use the wireless downlink power transfer and uplink information transmission protocol for deploying the UAV and collecting data. Particularly, we have two phases: wireless power transfer (WPT) and wireless information transmission (WIT). In the first phase, the downlink is activated to transfer energy to the IoT devices from the UAV during time span $\tau\mathfrak{T}$. Then, the WIT phase takes place when the IoT devices transmit information to the UAV in the uplink during $(1-\tau)\mathfrak{T}$. We normalise the length of time step to $\mathfrak{T} = 1$ for convenience.

4.2.1 Channel model

We denote the channel gain between the UAV and the RIS, between the RIS and the nth IoT device, and the direct link from the UAV to nth IoT node at time

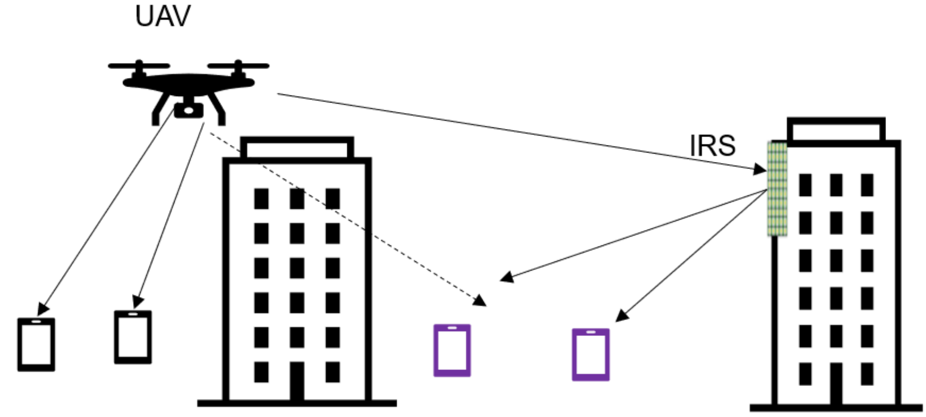


Figure 4.1: An illustration of the system model of UAV-assisted IoT wireless communications with the support of a RIS.

step t by $H^t \in \mathbb{C}^{1\times K}$, $h^t_{RIS,n} \in \mathbb{C}^{1\times K}$, and h^t_n, respectively. The small-scale fading of the direct link from the UAV to the IoT devices is assumed to be Rayleigh fading due to the extensive scatters. The air-to-air channel is considered for the UAV and the RIS link, while the link from the RIS to the IoT devices can be modelled by the Rician fading channel.

The distance between UAV and the kth element of RIS in time step t is given by

$$d^t_k = \sqrt{(x^t_{UAV} - x^t_k)^2 + (y^t_{UAV} - y^t_{UAV})^2 + (z^t_{UAV} - z^t_k)^2}. \tag{4.1}$$

Similarly, we denote the distance between the UAV and the nth IoT device and between the kth RIS element and the nth IoT node by d^t_n and $d^t_{k,n}$, respectively.

The channel gain between the UAV and the nth IoT device is given by

$$h^t_n = \sqrt{\beta_0 (d^t_n)^{-\kappa_1}}\hat{h}, \tag{4.2}$$

where β and κ_1 are the path loss at reference distance $1m$ and the path loss exponent for the UAV and the IoT devices link, respectively; $\hat{h}$ represents the small-scale fading modelled by complex Gaussian distribution with zero-mean and unit-variance $\mathcal{CN}(0,1)$.

Similarly, the channel gain between the UAV and the RIS is an air-to-air channel dominated by the line-of-sight (LoS) links. Thus, the channel of the UAV-RIS link in time step t is denoted as follows:

$$H^t = \sqrt{\beta_0 (d_k^t)^{-\kappa_2}} \big[1, e^{-j\frac{2\pi}{\lambda} d \cos(\phi_{AoA}^t)}, \ldots, e^{-j\frac{2\pi}{\lambda}(K-1) d \cos(\phi_{AoA}^t)}\big]^T, \tag{4.3}$$

where the right term is the array signal from the UAV to the RIS, $\cos(\phi_{AoA}^t)$ is the cosine of the angle of arrival (AoA) from the UAV to RIS; κ_2, d and λ are the path loss exponent for the UAV and the RIS link, the element spacing and the carrier wavelength, respectively.

The channel gain between the RIS and the nth IoT device following the Rician fading is expressed as

$$h_{RIS,n}^t = \sqrt{\beta_0 (d_{k,n}^t)^{-\kappa_3}} \big(\sqrt{\frac{\beta_1}{1+\beta_1}} h_{RIS,n}^{LoS} + \sqrt{\frac{1}{\beta_1+1}} h_{RIS,n}^{NLoS}\big), \tag{4.4}$$

where the deterministic LoS component is denoted by $h_{RIS,n}^{LoS} = [1, e^{-j\frac{2\pi}{\lambda} d \cos(\phi_{AoD}^t)}, \ldots,$ $e^{-j\frac{2\pi}{\lambda}(K-1) d \cos(\phi_{AoD}^t)}]$ and the non-line-of-sight (NLoS) component is the Rayleigh fading that follows the complex Gaussian distribution with zero mean and unit variance; $\cos\phi_{AoD}$ is the angle of departure (AoD) from the RIS to IoT devices; β_1 is the Rician factor, and κ_3 is the path loss exponent for the RIS and IoT devices link.

4.2.2 Power transfer phase

The achievable signal at the nth IoT device is composed of the direct signal from the UAV and the reflected signal from the RIS at time step t as

$$y_{1n}^t = (h_n^t + H^t \Phi^t h_{RIS,n}^t)\sqrt{P_0}x + \varrho^2, \tag{4.5}$$

where ϱ^2 is the noise signal following the complex Gaussian distribution $\mathcal{CN}(0, \alpha^2)$, x is the symbol signal from the UAV and P_0 is the transmission power at the UAV; $\Phi^t = \text{diag}[\phi_1^t, \phi_2^t, \ldots, \phi_K^t]$ is the diagonal matrix at the RIS, where $\phi_k^t = e^{j\theta_k^t}, \forall k = 1, 2, \ldots, K$ and $\theta_k^t \in [0, 2\pi]$ denotes the phase shift of the kth element in the RIS at time step t.

In the WPT phase, the UAV transfers energy to the IoT devices during time span τ^t at time step t. Thus, the received power at the nth IoT devices at time step t is given by

$$p_n^t = \tau^t \eta P_0 |h_n^t + H^t \Phi^t g_n^t|^2, \tag{4.6}$$

where η is the power transfer efficiency. It is important to note that although we employ the linear EH model in this paper the non-linear counterpart should have been adopted in realistic scenarios.

4.2.3 Information transmission phase

We assume that the IoT devices do not have fixed energy sources and use all the harvested energy for the WIT phase. The signal received at the UAV from the

nth IoT device is given by

$$y^t_{2n} = (h^t_n + H^t \Phi^t h^t_{RIS,n})\sqrt{p_n} u_n + \varrho^2, \tag{4.7}$$

where u_n is the symbol signal from the nth IoT device to the UAV. The received SINR at the UAV from transmission of the nth IoT device at time step t can be formulated as follows:

$$\gamma^t_n = \frac{p^t_n |h^t_n + H^t \Phi^t g^t_n|^2}{\sum_{m \neq n}^N p^t_m |h^t_m + H^t \Phi^t g^t_m|^2 + \alpha^2}, \tag{4.8}$$

The sum-rate from the IoT devices at time step t is formulated as follows:

$$R^t_{total} = \sum_{n=1}^N (1-\tau^t) W \log_2(1+\gamma^t_n), \tag{4.9}$$

where W is the bandwidth.

Our objective is to maximise the achieved sum-rate performance by optimising the phase shift matrix Φ at the RIS, the UAV's trajectory $\mathit{\Gamma}$ and the EH time τ as

$$\begin{aligned} &\max_{\tau,\Phi,\mathit{\Gamma}} \sum_{n=1}^N (1-\tau^t) B \log_2(1+\gamma^t_n), \\ &s.t. \quad 0 < \tau^t < 1, \\ &\qquad \theta^t_k \in [0, 2\pi], \forall k \in K, \\ &\qquad v^t \leq v_{max}, \\ &\qquad X^t_{UAV} \in Z, \end{aligned} \tag{4.10}$$

where Z represents the flying restricted area in the vertical and horizontal dimensions; v^t and v_{max} are the UAV's velocity at time step t and the UAV's maximum flying velocity, respectively.

4.3 Hovering UAV for downlink power transfer and uplink information transmission in RIS-assisted UAV communications

Besides WPT, the UAV uses most of the energy for its movement. Thus, to extend the operating time, the UAV is considered to hover at a fixed position at the centre of the cluster. Firstly, we present the mathematical background of the DRL algorithm and then we apply the DRL algorithm for solving the sum-rate maximisation problem in RIS-assisted UAV communications.

4.3.1 The DDPG method

The DDPG algorithm is a hybrid model composed of the value function and policy search methods. Thus, the DDPG algorithm is suitable for large-scale action and state spaces. Based on the current policy, the actor function $\mu(s;\theta_\mu)$ maps the states to a specific action with θ_μ being the actor network parameters, while the critic function $Q(s,a)$ evaluates the quality of the action taken. In the DDPG algorithm, we use *experience replay buffer* and *target network* technique to improve the convergence speed and avoid excessive calculations.

The agent iteratively interacts with the environment by executing the action a^t and receiving the response with instant reward r^t and the next state s^{t+1}. The tuple of (s^t, a^t, r^t, s^{t+1}) is then stored in a replay buffer D for training the actor and critic network. The buffer D is updated by adding new samples and discarding the oldest ones due to its finite size setting. After achieving enough samples, the agent takes a batch G of transitions for training the network. Particularly, we train the actor and critic network using stochastic gradient descent (SGD) over a mini-batch G samples.

Let us denote the parameters of the critic network and the target critic network by θ_q and $\theta_{q'}$, respectively. The critic network is updated by minimising

$$L = \frac{1}{G}\sum_{i}^{G}\left(y^i - Q(s^i, a^i; \theta_q)\right)^2, \tag{4.11}$$

with

$$y^i = r^i(s^i, a^i) + \zeta Q'(s^{i+1}, a^{i+1}; \theta_{q'})|_{a^{i+1}=\mu'(s^{i+1};\theta_{\mu'})}, \tag{4.12}$$

where the action at time step $(i+1)$ can be obtained by running the target actor network μ' with the state s^{i+1}; $\theta_{\mu'}$ denotes the parameters of the target actor network and ζ is the discounting factor.

The actor network parameters are updated by

$$\nabla_{\theta_\mu} J \approx \frac{1}{G}\sum_{i}^{G} \nabla_{a^i} Q(s^i, a^i; \theta_q)|_{a^i=\mu(s^i)} \nabla_{\theta_\mu}\mu(s^i; \theta_\mu). \tag{4.13}$$

Moreover, we duplicate the actor network and the critic network after a number of episodes to create a target actor and a target critic network. It helps reduce the excessive calculations by using only one network to estimate the target value. The target actor network parameters θ_q and the target critic network parameter

$\theta_{\mu'}$ are updated by using soft target updates associated with $\varkappa \ll 1$, as

$$\theta_{q'} \leftarrow \varkappa\theta_q + (1-\varkappa)\theta_{q'}, \tag{4.14}$$

$$\theta_{\mu'} \leftarrow \varkappa\theta_\mu + (1-\varkappa)\theta_{\mu'}. \tag{4.15}$$

For *explorations* and *exploitations* purpose, we add a noise process of $\mathcal{N}(0,1)$ as follows [127]:

$$\mu'(s^t) = \mu(s^t;\theta_\mu^t) + \psi\mathcal{N}(0,1), \tag{4.16}$$

where ψ is a hyper-parameter.

4.3.2 Game solving

For the hovering scenario, we formulate the MDP [86] by a 4-tuple $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} >$. Then, we formulate a game to solve the problem in (4.10).

- *Agent*: The UAV's centralised processor will act as an agent. The agent interacts with the environment to find an optimal policy π^* for maximising the total sum-rate. After training, the action-making schemes will be deployed to the UAV to predict the proper EH time scheduling τ and the RIS can choose the phase shift matrix Φ.

- *State space*: The channel is composed of both direct link and the reflective channel. Thus, we define the state space as

$$\mathcal{S} = \{h_1 + H\Phi g_1, h_2 + H\Phi g_2, \ldots, h_N + H\Phi g_N\}, \tag{4.17}$$

In time step t, the UAV has the state $s^t = \{h_1^t + H^t \Phi^t g_1^t, h_2^t + H^t \Phi^t g_2^t, \dots, h_N^t + H^t \Phi^t g_N^t\}$

- *Action space*: The UAV hovers at a fixed position; thus, we optimise the EH time τ and the RIS's phase shift Φ. The action space is defined as

$$\mathcal{A} = \{\tau, \theta_1, \theta_2, \dots, \theta_K\} \tag{4.18}$$

At the state s^t, the UAV takes the action $a^t = \{\tau^t, \theta_1^t, \theta_2^t, \dots, \theta_K^t\}$ and move to the next state $s' = s^{t+1}$.

- *Reward function*: The UAV interacts with the environment to find the maximum obtained reward. In our work, we formulate the reward function to obtain the maximum total sum-rate performance as

$$\mathcal{R} = \sum_{n=1}^{N} (1 - \tau^t) B \log_2(1 + \gamma_n^t) \tag{4.19}$$

The UAV is hovering at X_{UAV} and chooses the action a^t based on the achieved channel state information (CSI). Then, the UAV transfers the energy during τ to the IoT devices and the RIS controller adjusts the phase shift for each element. During the remaining time $(1 - \tau)$, the RIS will not change the phase shift while the IoT devices transmit information in the uplink to the UAV. It is challenging since the RIS plays a crucial role in mitigating the interferences. Thus, we need to find an intelligent scheme for the RIS to maximise the network performance in the downlink power transfer and uplink information transmission protocol. We propose a DRL, namely DDPG algorithm, to find an optimal policy for the UAV and the RIS.

At time step t, the agent has the information of the channel (4.17) and uses the actor network to choose the action a^t (4.18). By executing the action a^t, the agent receives the response following the reward function (4.19) from the environment. The critic action takes part to justify the efficiency of the chosen action a^t. After storing enough samples in buffer D, the agent trains the networks over a mini-batch G by using the SGD with Adam optimiser [139].

In this section, we assume that the UAV is hovering at a fixed position to reduce the flying energy consumption. It is a trade-off game with the energy and total achievable sum-rate. In the next section, we propose a joint optimisation of trajectory, EH time and the phase shift to maximise the network throughput in a short operation time.

4.4 Joint trajectory, EH time scheduling and the RIS phase shift optimisation using deep reinforcement learning

Given a short flying time of the UAV, to maximise total achievable sum-rate, we propose a joint optimisation scheme between the UAV's trajectory, EH time scheduling of IoT, and the RIS's phase shift. We define the state space and the reward function as in Section 4.3. We modify the action space as follows:

$$\mathcal{A} = \{\upsilon, \varsigma, \tau, \theta_1, \theta_2, \dots, \theta_K\} \tag{4.20}$$

At the state s^t, the UAV takes the action $a^t = \{v^t, \varsigma^t, \tau^t, \theta_1^t, \theta_2^t, \ldots, \theta_K^t\}$ and moves to the next state $s' = s^{t+1}$. Particularly, the position of the UAV at time step $(t+1)$ is represented as follows:

$$X_{UAV}^{t+1} = \begin{cases} x_{UAV}^{t+1} = & x_{UAV}^{t} + v^t \cos \varsigma^t + \Delta x^{t+1} \\ y_{UAV}^{t+1} = & y_{UAV}^{t} + v^t \sin \varsigma^t + \Delta y^{t+1} \\ H_{UAV}^{t+1} = & H_{UAV}^{t} + \Delta H^{t+1}, \end{cases} \tag{4.21}$$

where $\Delta x^{t+1}, \Delta y^{t+1}$, and ΔH^{t+1} are the environmental noise on the UAV at time step $(t+1)$. The UAV is flying from the position X_{UAV}^t to X_{UAV}^{t+1} but still needs to satisfy the flying zone constraint $X_{UAV} \in Z$. Moreover, the velocity of the UAV is set to satisfy the requirement $v \leq v_{max}$ and the flying angle is set to satisfy a constraint, $\varsigma \in [0, 2\pi]$.

Our objective is to find the optimal policy π^* for maximising the expected reward $\mathcal{R}$. The agent has the local knowledge and interacts with the environment to receive the reward. The local information is used to formulate the state following (4.17) and then the action is chosen based on (4.18). Base on the received reward, the agent adjusts the policy π and executes a new action at a new state. The agent can find a better policy with a better reward by the iterative interactions. After each execution of the action, the UAV will move to a new position and receive responses from the environment. By interacting iteratively with the environment, the agent can choose the proper velocity and the flying direction for the UAV in each time step based on the achieved CSI. Simultaneously, the EH scheduling τ and the phase shift matrix are also optimised for maximising network performance. Here, M and T are the number of the maximum episodes

Algorithm 3 Deep deterministic policy gradient algorithm for joint trajectory design, EH time and phase shift optimisation in RIS-assisted UAV communications.

```
Initialise the actor network $\mu(s;\theta_\mu)$, target actor network $\mu'$ and the critic
network $Q(s,a;\theta_q)$, the target critic networks $Q'$.
Initialise replay memory pool $\mathcal{D}$
for episode $= 1,\dots,M$ do
    Initialise an action exploration process $\mathcal{N}$
    Receive initial observation state $s^0$
    for iteration $= 1,\dots,T$ do
        Find the action $a^t$ for the state $s^t$
        Execute the action $a^t$
        Update the reward $r^t$ according to (4.19)
        Observe the new state $s^{t+1}$
        Store transition $(s^t,a^t,r^t,s^{t+1})$ into replay buffer $\mathcal{D}$
        Sample randomly a mini-batch of $G$ transitions $(s^i,a^i,r^i,s^{i+1})$ from $\mathcal{D}$
        Update critic parameter by SGD using the loss (4.11)
        Update the actor policy parameter (4.13)
        Update the target networks as in (4.14) and (4.15)
        Update the state $s_i^t = s_i^{t+1}$
    end for
end for
```

and time steps, respectively. The details of our DDPG algorithm-based technique for joint trajectory design, EH time and phase shift matrix optimisation in RIS-assisted UAV communications are presented in Algorithm 3.

4.5 Proximal policy optimisation technique for joint trajectory, EH time and the phase shift optimisation

For the continuous state and action space as in our problem, we propose an on-policy algorithm, namely the PPO algorithm, for the joint optimisation of trajectory, EH time and the phase shift of the RIS. Instead of training the critic and actor network, in PPO algorithm, we use a policy network for directly searching for an optimal performance with efficient sampling technique. We define the policy by π with the parameter θ_π. Here, we train the policy and adjust the parameter to find an optimal policy π^* by running the SGD over a mini-batch of G transitions (s^i, a^i, r^i, s^{i+1}). The policy parameters are updated for optimising the objective function as follows:

$$\theta_\pi^{i+1} = \underset{\theta_\pi}{\text{argmax}} \frac{1}{G} \sum_{i}^{G} \nabla_{a^i} \mathcal{L}(s^i, a^i; \theta_\pi). \tag{4.22}$$

In the PPO algorithm, the agent interacts with the environment to find the optimal policy π^* with the parameter θ_{π^*} that maximises the reward as

$$\mathcal{L}(s, a; \theta_\pi) = \mathbb{E}\Bigg[p_\theta^t A^\pi(s, a)\Bigg], \tag{4.23}$$

where $p_\theta^t = \frac{\pi(s,a;\theta_\pi)}{\pi(s,a;\theta_{old})}$ is the probability ratio of the current policy and previous policy; $A^\pi(s, a)$ is the advantage function [140].

Here, if we use only one network for the policy, the excessive modification occurs during the training stage. Thus, we use the clipping surrogate method as follows [141]:

$$\mathcal{L}^{\text{clip}}(s, a; \theta_\pi) = \mathbb{E}\Bigg[\min\Big(p_\theta^t A^\pi(s, a), \text{clip}(p_\theta^t, 1 - \epsilon, 1 + \epsilon) A^\pi(s, a)\Big)\Bigg], \tag{4.24}$$

where ϵ is a small constant. In this paper, the advantage function $A^\pi(s, a)$ [142]

Algorithm 4 Our proposed approach based on the PPO algorithm for the RIS-assisted UAV communications.

```
Initialise the policy $\pi$ with the parameter $\theta_\pi$
Initialise the penalty method parameters $\epsilon$
for episode $= 1, \dots, M$ do
    Receive initial observation state $s^0$
    for iteration $= 1, \dots, T$ do
        Find the action $a^t$ based on the current state $s^t$ by following the current policy
        Execute the action $a^t$
        Update the reward $r^t$ according to (4.19)
        Observe the new state $s^{t+1}$
        Update the state $s_i^t = s_i^{t+1}$
        Collect a set of partial trajectories with $G$ transitions
        Estimate the advantage function as (4.25)
    end for
    Update policy parameters using SGD with a mini-batch $B$ of the collected samples
        $\theta^{i+1} = \underset{\theta_\pi}{\text{argmax}} \frac{1}{G} \sum^{G} \mathcal{L}^{\text{clip}}(s, a; \theta_\pi)$    (4.27)
end for
```

is formulated as follows:

$$A^\pi(s, a) = r^t + \zeta V^\pi(s^{t+1}) - V^\pi(s^t). \tag{4.25}$$

The policy is then trained by a mini-batch B and the parameters are updated by

$$\theta^{i+1} = \underset{\theta_\pi}{\text{argmax}}\, \mathbb{E}\Big[\mathcal{L}^{\text{clip}}(s, a; \theta_\pi)\Big]. \tag{4.26}$$

The details of our PPO algorithm-based technique for joint trajectory design, EH time and phase shift matrix optimisation in RIS-assisted UAV communications are presented in Algorithm 4.

4.6 Simulation Results

Table 4.1: Simulation parameters in Chapter 4

Parameters	Value
Bandwidth (W)	1 MHz
UAV transmission power	5 W
UAV maximum speed per time step	20 m
UAV's coverage	500 m
The initial UAV's position	$(0, 0, 200)$
The RIS's position	$(200, 0, 50)$
Path-loss parameter	$\kappa_1 = 4, \kappa_2 = 2, \kappa_3 = 2.2$
Channel power gain	$\beta_0 = -30$ dB
EH efficiency	$\eta = 0.5$
Rician factor	$\beta_1 = 4$
Noise power	$\alpha^2 = -134$ dBm
Max number of episodes	$M = 5000$
Max number of time step	$T = 200$
Clipping parameter	$\epsilon = 0.2$
Discounting factor	$\zeta = 0.9$
Max number of IoT devices	20
Initial batch size	$K = 32$

In our work, we use the Tensorflow 1.13.1 [138] for implementing our algorithms. We deploy the UAV at $(0, 0, 200)$, the RIS at $(200, 0, 50)$ and assume $d/\lambda = 1/2$ for convenience. All other parameters are provided in Table 4.1. In order to compare our proposed model with other baseline schemes, in this paper, we consider the techniques as follows:

- **Optimisation with the hovering UAV**: the UAV is maintained at a fixed position at the centre of the cluster $(0, 0, H_{UAV})$. We optimise the EH time τ and the phase shift matrix at the RIS. We use the DDPG algorithm (H-

DDPG) and the PPO algorithm (H-PPO) for the problem in the hovering UAV scenario.

- **Our proposed model with mobile UAV**: For the game formulated as in Section 4.4, we use the DDPG algorithm (F-DDPG) and the PPO algorithm (F-PPO) for solving the problem of joint optimisation of trajectory, EH time scheduling and the phase shift matrix at the RIS.

- **Random selection scheme (RSS)**: The value of Φ is selected randomly and we use the DDPG algorithm (RSS-HDDPG) for optimising the EH time τ in the hovering UAV scenario.

- **Random EH time (REH)**: The EH time τ is selected randomly and the flying path and the phase shift Φ are optimised to maximise the performance. We use the DDPG algorithm (REH-DDPG) and the PPO algorithm (REH-PPO) for optimisation.

- **Without RIS**: We do not deploy the RIS in this scenario and optimise the EH time τ in the hovering UAV scenario using the DDPG algorithm (WithoutRIS-HDDPG), PPO algorithm (WithoutRIS-HPPO).

Firstly, we consider the hovering UAV scenario and compare the performance versus the different number of IoT devices, N, with the number of RIS, $K = 20$ in Fig. 4.2. We take the average of over 1000 episodes for each scheme to draw the figures. When using the H-PPO algorithm, the total expected throughput is higher than in other schemes, including the ones using the H-DDPG algorithm, the RSS-HDDPG, WithoutRIS-HDDPG and WithoutRIS-HPPO technique. The results suggest that with the EH time and the RIS's reflecting coefficient opti-

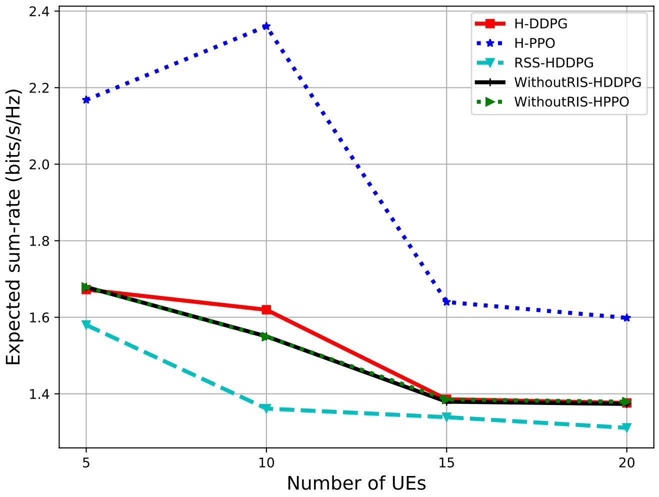


Figure 4.2: The sum-rate performance in the hovering UAV scenario with different numbers of IoT devices, N.

misation, the PPO algorithm is adequate irrespectively of the number of IoT devices.

Next, we present the achieved sum-rate of the PPO and DDPG algorithm in the hovering UAV scenario comparing with the RSS and without RIS case while the number of IoT devices is fixed at $N = 10$ in Fig. 4.3. The H-PPO again shows the effective results with the different number of RIS elements, K. The sum-rate performance of the H-PPO algorithm improves from 2.0 to 2.8 (bits/s/Hz) following the increase of the RIS elements. The sum-rate performance of the H-DDPG algorithm is slightly higher than the ones using RSS and without RIS schemes. The RIS is a passive reflector; thus, the reflected signal is diverse and not towards the destinations if we cannot control the coefficient of the RIS and select the phase shift randomly. Moreover, the PPO and the DDPG algorithm reach similar results when we only optimise the EH time without the RIS.

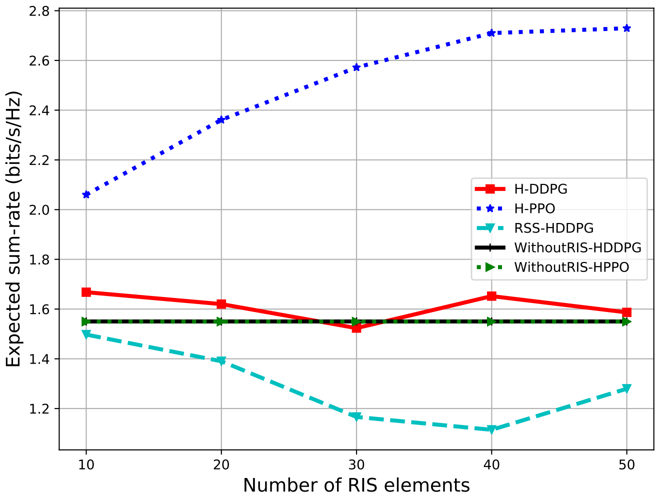


Figure 4.3: The sum-rate performance in the hovering UAV scenario with varying number of RIS elements, K.

Furthermore, in Fig 4.4, we investigate the performance while using the PPO algorithm in the hovering UAV scenario with the different number of RIS elements, K. When we increase the number of RIS elements K, the performance is enhanced. The results suggest that the PPO algorithm can handle the large optimising variables and effective in the hovering UAV scenario.

In Fig. 4.5, we compare the total sum-rate in the mobile UAV with the number of RIS elements $K = 20$ and different numbers of IoT devices, N. In contrast with the hovering scenarios, the method based on the F-DDPG algorithm shows impressive results over other schemes. When using the F-DDPG algorithm, we can achieve a total throughput of around 3.8 bits/Hz. The F-PPO algorithm is not good and trapped in an optimal local value. The reason is that the F-PPO algorithm is an on-policy method and offers less random exploration over the training.

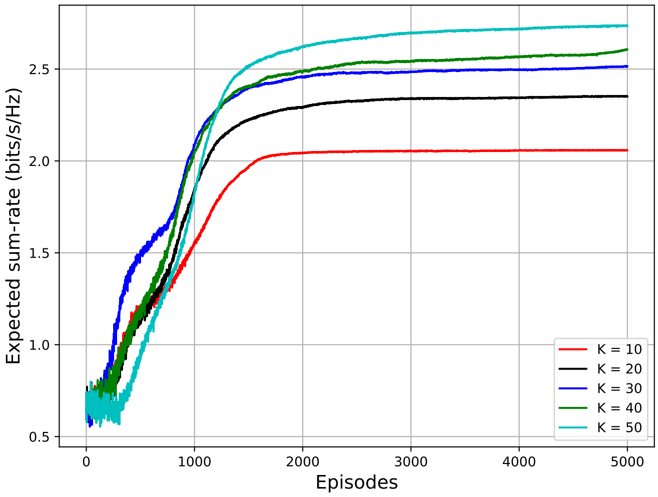


Figure 4.4: The performance while using the PPO algorithm in the hovering UAV scenario with different number of RIS elements, K

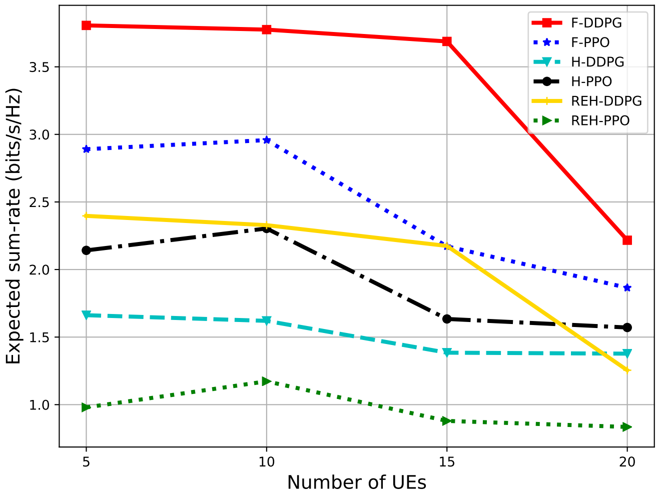


Figure 4.5: The sum-rate performance with different number of IoT devices, N.

We consider the different number of RIS elements K and compare the performance of our proposed algorithms with REH schemes and hovering UAV scenario, as shown in Fig. 4.6. The F-DDPG algorithm-based technique outperforms other

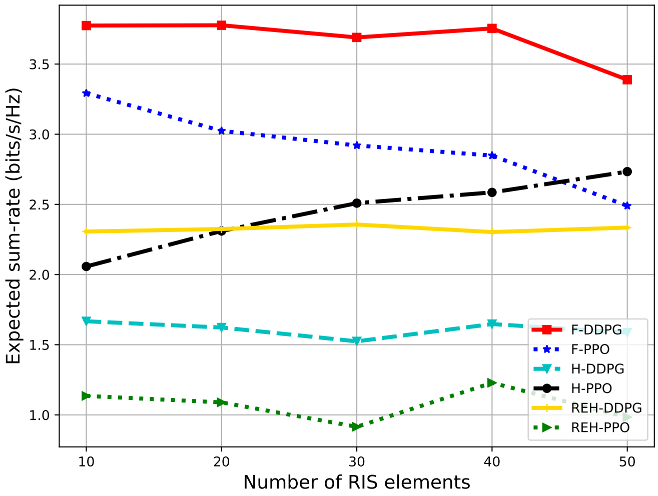


Figure 4.6: The sum-rate performance with different numbers of RIS elements, K.

schemes while it reaches around 3.8 (bits/s/Hz). Following the performance using the F-DDPG algorithm is the ones using the F-PPO algorithm in the mobile UAV. When we jointly optimise the UAV's trajectory, IoT's EH time and RIS's phase shift, the achievable sum-rate is significantly increased in comparison with the case when we optimise only the EH time, RIS phase shift in hovering scenario and when we consider the optimisation of trajectory and EH time in REH-DDPG, REH-PPO algorithm.

We compare the performance of the DDPG algorithm with different values of the discounting factor, ζ in the mobile UAV scenario in Fig. 4.7. When we set the value of ζ too small, the expected sum-rate only achieved a local optimum. The higher values of ζ can help the algorithm converge faster because the next state reward can utilise the previously achieved reward to adjust the network parameters. Based on the results in Fig. 4.7, we set the value $\zeta = 0.9$ for implementing our algorithms.

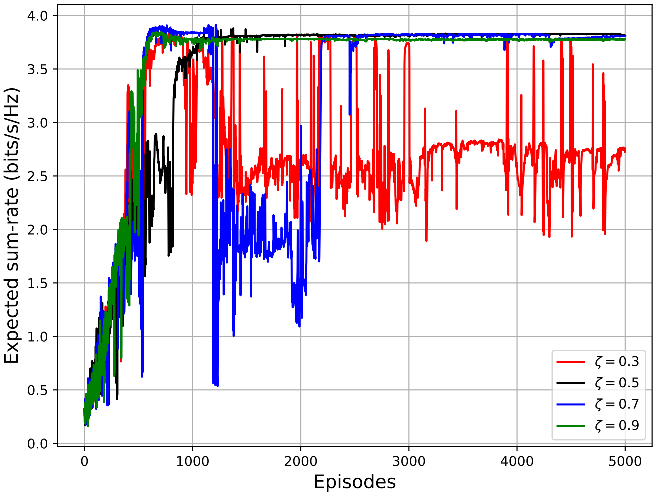


Figure 4.7: The performance with different values of the discounting factor, ζ.

4.7 Conclusion

In this chapter, we have introduced a new system model for RIS-assisted UAV communications with the downlink power transfer and uplink information transmission protocol. By utilising the UAV's mobility, the flexibility of the RIS, and the effectiveness of the protocol, the RIS-assisted UAV network is a promising technique for practical applications. We have proposed two DRL techniques for jointly optimising the UAV's trajectory, IoT's EH time scheduling and the phase shift matrix of the RIS to maximise the network's throughput. The results suggest that the systems learned by the DRL algorithm can deal with dynamic environments and satisfy some power restrictions and processing time in RIS-assisted UAV communications. In the future, we plan to extend our work to include a distributed model and cooperative communications with multiple UAVs.

Chapter 5 is published as

K. K. Nguyen, S. Khosravirad, D. B. da Costa L. D. Nguyen, and T. Q. Duong, "Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning," *IEEE J. Selected Topics in Signal Process.*, 2021 (accepted).

Chapter 5

Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning

5.1 Introduction

The UAVs are also playing a crucial role in bringing beyond fifth generation (5G) network to every corner around the world owing to their low-cost production and flexibility. UAV-assisted wireless networks significantly enhance the network's coverage and improve the information transmit efficiency.

Very recently, RIS has emerged as a cutting-edge technology for beyond 5G and sixth generation (6G) networks. In particular, a massive number of reflective elements are intelligently controlled to reflect the received signal toward the destinations. The controller helps the RIS be dynamically adapted to the propagation environment with the aim to meet different purposes; for example, enhance the arrival signal and mitigate the interference [50, 59, 60, 101, 114–117]. The RIS has been recently deployed efficiently due to its low-cost hardware production, nearly-passive nature, easy deployment, communication without new waves, and energy-saving nature.

Owing to the intrinsic features of RIS and UAVs, the RIS-assisted UAV communications have been recently considered for enhancing network performance. Although the high altitude of the UAV significantly strengthens the channel between the UAV and the users, the connections are sometimes blocked by buildings or other obstacles in specific scenarios. Thus, the RIS attached to the building or on a high place is an option to reflect the channel from the UAV to the users [38,56,57]. Moreover, the data through the RIS will experience fewer intermediate delays and more freshness than when we use a mobile active relay in the middle. On the other hand, the RIS is easily deployed and effective in reducing power consumption.

DRL algorithms have emerged as a powerful method for an embedded optimisation and instant decision-making model in wireless networks. The DRL methods have been used for device-to-device (D2D) communication [27,28], UAV-assisted networks [18], and RIS-assisted wireless networks [131]. Inspiring by the impressive results, in this paper, we use the DRL algorithm to enable the UAVs to choose the proper power level and the RIS to adjust the phase-shift matrix to maximise the reward. The neural networks are trained in the offline phase and then deployed in the terminal devices or controllers. Thus, the proper actions can be chosen in milliseconds or instants in a centralised and decentralised manner to obtain an optimal performance in the multi-UAVs-assisted networks with the support of RIS.

In this chapter, we propose efficient DRL algorithms by jointly optimising the power allocation of the UAV and the RIS's phase-shifts for maximising the EE performance. The DRL approaches bring a flexible and autonomous ability to UAVs and RIS. With trained neural networks, the UAVs and RIS can choose a proper action without delay. Furthermore, continuous learning with up-to-date data by interaction with the environment helps the UAVs and RIS to adapt to the dynamic environment. In this chapter, we exploit the efficiency of DRL techniques in multi-UAV-assisted wireless communications with the support of RISs. Particularly, we conceive a wireless network of multi-UAVs supported by an RIS. Each UAV is deployed for serving a specific cluster of UEs. Due to the severe shadowing effect, the RIS is used to enhance the received signal's quality at the UEs from the associated UAV and to mitigate the interference from others. Next, the EE problem is formulated for the downlink channel with the power restrictions and the RIS's requirement. To optimise the EE network performance, we propose a centralised DRL technique for jointly solving the power allocation at the UAVs and phase-shift matrix of the RIS. Then, a parallel learning is used for training each element in our model to be intelligent and to reduce the delay when transmitting the action between UAV and the RIS. Then, to improve the network performance, we introduce the proximal policy optimisation (PPO) algorithm with a better sampling technique. Finally, through the numerical results, we demonstrate that our proposed methods efficiently solve the joint optimisation problem with the dynamic environmental setting and time-varying CSI and outperform the other benchmarks.

The remainder of this chapter is organised as follows. We present the system model and problem formulation for the energy-efficient multi-UAV-assisted wireless communications with the support of the RIS in Section 5.2. The centralised DDPG approach for joint optimisation of power allocation and phase-shift in multi-UAV-assisted wireless networks is introduced in Section 5.3. We propose parallel learning for our approach to reduce delay in Section 5.4. Moreover, the PPO algorithm is proposed for solving both centralised and decentralised learning in Section 5.5. Numerical results are illustrated in Section 5.6 while the conclusion and future works are presented in Section 5.7.

5.2 System Model and Problem Formulation

We consider a downlink multi-UAV wireless network assisted by one RIS. Each UAV is equipped with a single antenna for serving a specific cluster of users (UEs), in which it is assumed N UAVs corresponding to N clusters of UEs, where each cluster consists of M single-antenna UEs. The UEs are randomly distributed in the coverage C from the centre of each cluster. The channel between the UAV and UEs is blocked by the building, wall and concretes. Thus, we deploy an RIS with K elements for supporting the information transmission from the UAVs to the UEs.

5.2.1 System Model

We assume that the coordinate of the nth UAV and mth UEs in the nth cluster at the time step t are $X_n^t = \left(x_n^t, y_n^t, H_n^t\right)$ and $X_{mn}^t = (x_{mn}^t, y_{mn}^t)$, with $n = 1, \ldots, N$ and $m = 1, \ldots, M$. We consider that the UAVs hover at fixed altitude at the centre of the cluster. The RIS is attached to the building or a high location at (x^t, y^t, z^t), respectively.

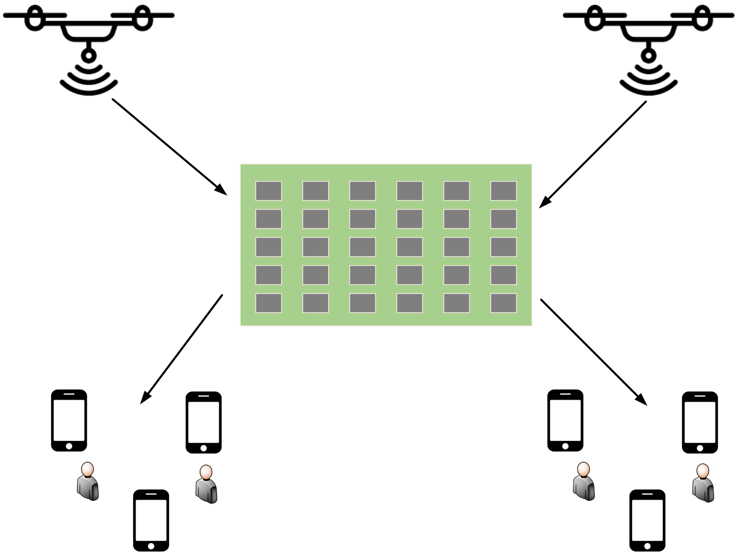


Figure 5.1: System setup.

The distance between the nth UAV and the RIS panel in time step t is denoted by

$$d_n^t = \sqrt{(x_n^t - x^t)^2 + (y_n^t - y^t)^2 + (H_n^t - z^t)^2}. \tag{5.1}$$

Similarly, the distance between the RIS panel and the mth UEs in the nth cluster is written as

$$d^t_{nm} = \sqrt{(x^t - x^t_{nm})^2 + (y^t - y^t_{nm})^2 + (z^t)^2}. \tag{5.2}$$

Due to the high shadowing and severe blocking effect, the direct links between UAVs and UEs do not exist and therefore it is only considered the alternative paths (reflected links) via RIS's reflection. The links between the UAVs and the RIS are modelled as air-to-air (AA) channels whereas the link between the RIS and the UEs is assumed to follow air-to-ground (AG) channel. Following the AA channel model, the channel gain between the nth UAV and the RIS in time step t is formulated as

$$H^t_{n,RIS} = \sqrt{\beta_0 (d^t_n)^{-\kappa_1}} \left[1, e^{-j\frac{2\pi}{\lambda} d \cos(\phi^t_{AoA})}, \ldots, e^{-j\frac{2\pi}{\lambda}(K-1) d \cos(\phi^t_{AoA})}\right]^T, \tag{5.3}$$

where β_0 is the channel gain at the reference distance d_0, κ_1 is the path loss exponent for the UAV-RIS link, d is element spacing, and λ is the carrier wavelength; the right term of (5.3) is the signal from the nth UAV to the RIS, $\cos(\phi^t_{AoA})$ is the cosine of the angle-of-arrival (AoA).

According to the AG channel model, the channel gain between the RIS and the mth UE in the nth cluster can be written as

$$h^t_{RIS,nm} = \sqrt{\beta_0 (d^t_{nm})^{-\kappa_2}} \left(\sqrt{\frac{\beta_1}{1+\beta_1}} h^{LoS}_{RIS,nm} + \sqrt{\frac{1}{\beta_1 + 1}} h^{NLoS}_{RIS,nm} \right), \tag{5.4}$$

where the deterministic LoS component is denoted by $h^{LoS}_{RIS,nm} = \left[1, e^{-j\frac{2\pi}{\lambda} d \cos(\phi^t_{AoD})}, \ldots, e^{-j\frac{2\pi}{\lambda}(K-1) d \cos(\phi^t_{AoD})}\right]$ and the non-light-of-sight

(NLoS) component is modelled as complex Gaussian distribution with a zero-mean and unit-variance $\mathcal{CN}(0,1)$; $\cos(\phi_{AoD})$ is the angle of departure (AoD) from the RIS to the mth UE in the nth cluster; β_1 and κ_2 are the Rician factor and the path loss exponent for the RIS-UEs link, respectively.

The signal from the UAV to UEs is reflected by the RIS. Thus, the received signal at the mth UE in the nth cluster at time step t can be written as

$$y_{nm}^t = H_{n,RIS}^t \Phi^t h_{RIS,nm}^t \sqrt{P_n} x_n + \sum_{l \neq n}^{N} H_{l,RIS}^t \Phi^t h_{RIS,lm}^t \sqrt{P_l^t} x_l + eta, \tag{5.5}$$

where $H_n^t \in \mathbb{C}^{1 \times K}$ is the channel gains array from the nth UAV to the RIS, η is the power noise signal following the complex Gaussian distribution with power α^2; P_n and x_n are the transmit power and the symbol signal sent from the nth UAV, respectively; $\Phi^t = \mathrm{diag}[\phi_1^t, \phi_2^t, \dots, \phi_K^t]$ is the diagonal matrix at the RIS, where $\phi_k^t = e^{j\theta_k^t}, \forall k = 1, 2, \dots, K$ with $\theta_k^t \in [0, 2\pi]$ is the phase-shift of the kth element in the RIS at time step t. In this chapter, we consider a broadcast scenario where the signal from the $n \in N$th UAV, x_n is different from the signal from the $l \in N$th UAV, x_l.

5.2.2 Problem Formulation

In this work, we consider a downlink communications where signal from the UAV is dedicated to a designated UE in the associated cluster. In other words, the mth UE in the nth cluster receives the information from the nth UAV while the signals from other UAVs are considered as interference. Thus, the received signal-to-interference-plus-noise-ratio (SINR) at the mth UE in the cluster n at time step t can be formulated as follows:

$$\gamma_{nm}^{t} = \frac{P_{n}^{t}|H_{n,RIS}^{t}\Phi^{t}h_{RIS,nm}^{t}|^{2}}{\sum_{l\neq n}^{N}P_{l}^{t}|H_{l,RIS}^{t}\Phi^{t}h_{RIS,lm}^{t}|^{2}+\alpha^{2}}. \tag{5.6}$$

The throughput at the mth UEs in the nth cluster at time step t is written as

$$R_{nm}^{t} = W\log_{2}(1+\gamma_{nm}^{t}), \tag{5.7}$$

where W is the bandwidth. The total throughput at time step t is cumulative from the UEs of all clusters and it can be expressed by

$$R_{total}^{t} = \sum_{n=1}^{N}\sum_{m=1}^{M}R_{nm}^{t}, \tag{5.8}$$

and the total power consumption is given by

$$P_{total} = \sum_{n=1}^{N}P_{n}+P_{K}+P_{c}, \tag{5.9}$$

where P_K and P_c are the power consumption at the RIS and the power circuit at the UAV, respectively.

Our objective is to maximise the EE of all UEs by jointly optimising the transmit powers at the UAVs and the phase-shifts at the RIS. In each time step t, each UAV will choose the proper power and each RIS's element will choose the phase-shift value depending on the local information that each component receives from the environment. The optimisation problem of maximising the EE of all UEs subject to the transmit power at UAVs and phase-shifts of RIS can be formulated as

$$\begin{aligned} \max_{P,\Phi} \quad & \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} R_{nm}^{t}}{\sum_{n=1}^{N} P_n + P_K + P_c} \\ s.t. \quad & 0 \leq P_n \leq P_{\max}, \forall n \in N, \\ & \theta_k \in [0, 2\pi], \forall k \in K, \end{aligned} \tag{5.10}$$

where $P = \{P_1, \ldots, P_N\}$ and $P_{\max}$ are the vector of power and the maximum information transmission power at the UAVs, respectively. To solve the EE maximisation problem, we propose two DRL algorithms for centralised approach and then the parallel learning distributed approach is introduced for practical applications.

5.3 Centralised Optimisation for Power Allocation and Phase-shift Matrix

In the centralised approach, we assume that the information is processed at a central point (e.g., cloud server) and the next action for each element in the system will be transferred at the beginning of each time step. Thus, for jointly optimising the power allocation at the UAVs and the phase-shift matrix at the RIS, we consider the central processing point as an agent. The optimisation problem can be formulated by the MDP $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \zeta >$. Particularly, with our centralised optimisation, we formulate the game as follows:

- *State space*: The agent interacts with the environment for maximising the EE performance. Thus, the agent only has knowledge about the local information, e.g., the reflected channel gains. In this chapter, we assume that the UAV and RIS have a perfect channel estimation model. The CSI at the UAV and RIS are the same and defined by the combination of the channel from UAV to RIS, RIS phase shift matrix and RIS to UEs. The state space is defined as follows:

$$\begin{aligned}\mathcal{S} =\{&H_{1,RIS}\ \Phi\ h_{RIS,11},\ H_{1,RIS}\ \Phi\ h_{RIS,12},\\ &\ldots,\ H_{n,RIS}\ \Phi\ h_{RIS,nm},\ \ldots,\ H_{N,RIS}\ \Phi\ h_{RIS,NM}\}.\end{aligned} \tag{5.11}$$

- *Action space*: With the downlink transmission in the RIS-assisted multi-UAV networks, we optimise the power allocation at UAVs and phase-shift matrix at RIS. Thus, the action space is defined as follows:

  $$\mathcal{A} = \{P_1, P_2, \ldots, P_N, \theta_1, \theta_2, \ldots, \theta_K\}. \tag{5.12}$$

  The agent takes the action $a^t = \{P_1^t, P_2^t, \ldots, P_N^t, \theta_1^t, \theta_2^t, \ldots, \theta_K^t\}$ at the state s^t and moves to the next state $s' = s^{t+1}$.

- *Reward function*: Our objective is to maximise the EE performance; thus, we formulate the reward function as

  $$\mathcal{R} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} R_{nm}^t}{\sum_{n=1}^{N} P_n + P_K + P_c}. \tag{5.13}$$

After formulating the EE game, we proposed a DRL algorithm for the agent to interact with the environment to find the optimal policy π^*. Deep deterministic policy gradient (DDPG) is a hybrid model composed of the actor part based on value function and the critic component based on the policy search. In the DDPG algorithm, we use *experience replay buffer* and *target network* techniques to improve the convergence speed and avoid excessive calculation. In the *experience replay buffer*, we use a finite size of a memory size $\mathcal{B}$ to store the executed transition $< s^t, a^t, r^t, s^{t+1} >$. After collecting enough samples, we randomly select a mini-batch D of transitions from buffer $\mathcal{B}$ for training the neural networks. The memory $\mathcal{B}$ is set to a finite size for updating the new sample and discarding the old ones. Otherwise, we use *target networks* for the critic and actor network when calculating the target value.

We denote the critic network as $Q(s, a; \theta_q)$ with the parameter θ_q and the target critic network as $Q'(s, a; \theta_{q'})$ with the parameter $\theta_{q'}$. Similarly, we initialise the actor network $\mu(s; \theta_\mu)$ with the parameter θ_μ and the target actor network $\mu'(s; \theta_{\mu'})$ with the parameter $\theta_{\mu'}$. We train the actor and critic network using the stochastic gradient descent (SGD) over a mini-batch of D samples. The critic network is updated by minimising

$$L = \frac{1}{D} \sum_{i}^{D} \left(y^i - Q(s^i, a^i; \theta_q) \right)^2, \tag{5.14}$$

with the target

$$y^i = r^i(s^i, a^i) + \zeta Q'(s^{i+1}, a^{i+1}; \theta_{q'})|_{a^{i+1} = \mu'(s^{i+1}; \theta_{\mu'})}. \tag{5.15}$$

The actor network parameters are updated by

$$\nabla_{\theta_\mu} J \approx \frac{1}{D} \sum_{i}^{D} \nabla_{a^i} Q(s^i, a^i; \theta_q)|_{a^i = \mu(s^i)} \nabla_{\theta_\mu} \mu(s^i; \theta_\mu). \tag{5.16}$$

The target actor network parameters θ_q and the target critic network parameters $\theta_{\mu'}$ are updated by using soft target updates as follows:

$$\theta_{q'} \leftarrow \varkappa\theta_q + (1-\varkappa)\theta_{q'}, \tag{5.17}$$

$$\theta_{\mu'} \leftarrow \varkappa\theta_\mu + (1-\varkappa)\theta_{\mu'}, \tag{5.18}$$

where $\varkappa$ is a hyperparameter between 0 and 1.

In the DDPG algorithm, the deterministic policy is trained in an off-policy way; thus, for *explorations* and *exploitations* purpose, we add a noise process of $\mathcal{N}(0,1)$ as follows [127]:

$$\mu'(s^t;\theta^t_{\mu'}) = \mu(s^t;\theta^t_\mu) + \psi\mathcal{N}(0,1), \tag{5.19}$$

where ψ is a hyperparameter. The details of our DDPG algorithm-based technique for joint power allocation and phase-shift matrix optimisation in RIS-assisted UAV communications are presented in Algorithm 5, where E and T denote the number of the maximum episode and time step, respectively.

In Algorithm 5, the agent interacts with the environments to maximise the obtained reward (5.13). In time step t, the agent has a local information of channel model $s^t \in \mathcal{S}$. At the state s^t, the agent chooses the action $a^t \in \mathcal{A}$ by the actor networks. By executing the action a^t in the environment, the agent obtains a response following the reward function (5.13). Then, the critic and actor network parameters are updated by training a mini-batch of D transitions by stochastic gradient descent with Adam optimiser [139].

Algorithm 5 Centralised optimisation for joint power allocation and phase-shift matrix in RIS-assisted UAV communications.

```
Initialise the critic network $Q(s, a; \theta_q)$ and the target critic networks $Q'$
Initialise the actor network $\mu(s; \theta_\mu)$ and the target actor network $\mu'$
Initialise replay memory pool $\mathcal{B}$
for episode $= 1, \dots, E$ do
    Initialise an action exploration process $\mathcal{N}$
    Receive initial observation state $s^0$
    for iteration $= 1, \dots, T$ do
        Execute the action $a^t$ obtained at state $s^t$
        Update the reward $r^t$ according to (5.13)
        Observe the new state $s^{t+1}$
        Store transition $(s^t, a^t, r^t, s^{t+1})$ into replay buffer $\mathcal{B}$
        Sample randomly a mini-batch of $D$ transitions $(s^i, a^i, r^i, s^{i+1})$ from $\mathcal{B}$
        Update critic parameter by stochastic gradient descent using loss function in (5.14)
        Update the actor policy parameter in (5.16)
        Update the target networks as in (5.17) and (5.18)
        Update the state $s^t = s^{t+1}$
    end for
end for
```

5.4 Parallel DRL for Joint Power Allocation and Phase-shift Matrix Optimisation

In practical applications, when we process all the data in a centralised manner, the information of the UAV's power and the RIS's phase-shift for the next action need to transfer at the beginning of each time step. The delay will occur and make the system unable to deal efficiently with the dynamic environment. Thus, we propose a parallel DRL (PDRL) technique for joint power allocation and phase-shift matrix optimisation. As the definition of the DRL model, the agents do not know the environmental factor. Thus, in our system, the nth UAV has no knowledge of the power of the mth UAV and the diagonal matrix at the RIS.

Similarly, the RIS controller does not know about the transmit power at the UAV.

To make the UAV and the RIS work cooperatively, we consider a multi-agent learning for our system. In particular, each UAV acts as an agent and the RIS is a separated agent. For all the agents, we define the state space as $\mathcal{S} = \{H_{1,RIS}\ \Phi\ h_{RIS,11},\ H_{1,RIS}\ \Phi\ h_{RIS,12},\ \ldots,$
$H_{n,RIS}\ \Phi\ h_{RIS,nm},\ \ldots,\ H_{N,RIS}\ \Phi\ h_{RIS,NM}\}$ with respect to the channel state information, i.e., the compound of channel gains and phase-shifts of RIS. The UAV and the RIS process independently, thus, the action space for the nth UAV agent is the transmit power $\mathcal{A}_n = \{P_n\}$ and for the RIS agent is the phase-shift matrix $\mathcal{A}_{RIS} = \{\theta_1, \theta_2, \ldots, \theta_K\}$. With the rewards function, we rely on (5.13).

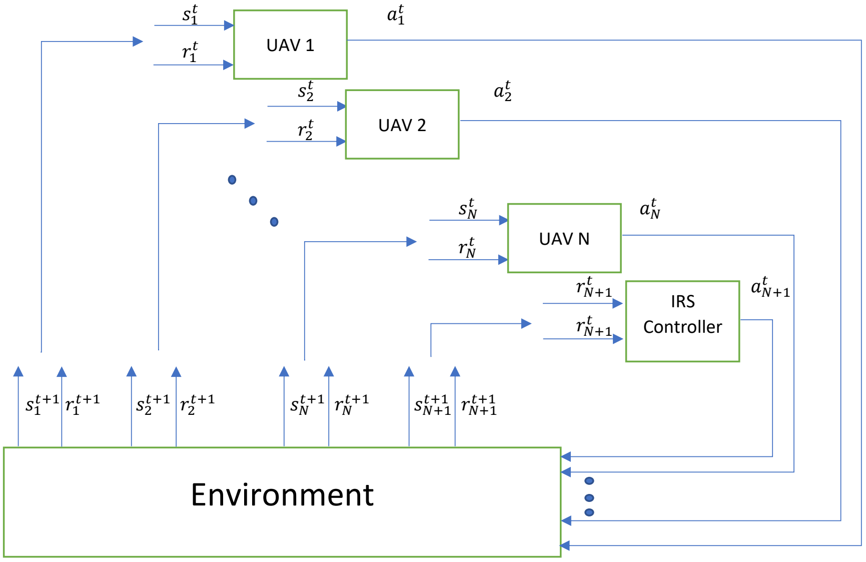


Figure 5.2: A multi-agent learning for the RIS-assisted wireless networks

In Algorithm 6, with the parallel learning, we have $N+1$ policy for the N UAVs and 1 policy for the RIS in the P-PPO algorithm. In time step t, the nth UAV decides the transmit power P_n and the RIS chooses the proper phase-shift matrix Φ^t at the state s^t for maximising the EE performance. In particular, our parallel model is described as in Fig. 5.2. The UAV and the RIS have the local information and interact with the environment to search for an optimal policy π^*. The agents at each time step choose and execute the action toward the environment. Then, the environment will respond by a value of reward toward the agents. Based on the response, the agents adjust the value of parameters in the action-chosen scheme for finding an optimal policy π^*. The details of our proposed techniques for joint optimisation of power allocation at the UAV and phase-shift matrix at the RIS are described in Algorithm 6. The agent $N+1$ represents the RIS controller.

Algorithm 6 Parallel learning for joint power allocation and phase-shift matrix in RIS-assisted UAV communications.

```
for Agent $\varpi = 1, \dots, N, N+1$ do
    Initialise the critic network $Q_\varpi(s, a; \theta_q)$, the target critic networks $Q'_\varpi$ and actor network $\mu_\varpi(s; \theta_\mu)$, target actor network $\mu'_\varpi$ for the agent $\varpi$
    Initialise replay memory pool $\mathcal{B}_\varpi$ for the agent $\varpi$
end for
for episode $= 1, \dots, E$ do
    Initialise an action exploration process $\mathcal{N}$
    Receive initial observation state $s^0$
    for iteration $= 1, \dots, T$ do
        for Agent $\varpi = 1, \dots, N, N+1$ do
            Execute the action $a^t_\varpi$ obtained at state $s^t$
            Update the reward $r^t_\varpi$ according to (5.13)
            Observe the new state $s^{t+1}_\varpi$
            Store transition $(s^t_\varpi, a^t_\varpi, r^t_\varpi, s^{t+1}_\varpi)$ into replay buffer $\mathcal{B}_\varpi$
            Sample randomly a mini-batch of $D$ transitions $(s^i_\varpi, a^i_\varpi, r^i_\varpi, s^{i+1}_\varpi)$ from $\mathcal{B}_\varpi$
            Update critic parameter by SGD using the loss (5.14)
            Update the actor policy parameter (5.16)
            Update the target networks as in (5.17) and (5.18)
            Update the state $s^t_\varpi = s^{t+1}_\varpi$
        end for
    end for
end for
```

5.5 Proximal Policy Optimisation for Centralised and Decentralised Problem.

Instead of using a hybrid model for continuous action space as in the DDPG algorithm, we propose an on-policy algorithm, namely proximal policy optimisation (PPO), with an efficient learning technique to achieve a better performance. In the PPO algorithm, we compare the current policy and obtained policy to find maximisation of the objective function as

$$\begin{aligned}\mathcal{L}(s,a;\theta) &= \mathbb{E}\Bigg[\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}A^{\pi}(s,a)\Bigg]\\ &= \mathbb{E}\Bigg[p^t_{\theta}A^{\pi}(s,a)\Bigg],\end{aligned} \tag{5.20}$$

where $p^t_{\theta} = \frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}$ denote the probability ratio and $A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$ is an estimator of the advantage function defined in [140]. We use SGD for training networks with a mini-batch D to maximise the objective. Thus, the policy is updated by

$$\theta^{t+1} = \text{argmax}\,\mathbb{E}\Big[\mathcal{L}(s,a;\theta^t)\Big]. \tag{5.21}$$

In this work, we use the clipping method function $\text{clip}(p^t_{\theta}, 1-\epsilon, 1+\epsilon)$ for limiting the objective value to avoid the excessive modification as follows [140]:

$$\mathcal{L}^{\mathsf{CLIP}}(s,a;\theta) = \mathbb{E}\Bigg[\min\Big(p^t_{\theta}A^{\pi}(s,a), \text{clip}(p^t_{\theta}, 1-\epsilon, 1+\epsilon)A^{\pi}(s,a)\Big)\Bigg], \tag{5.22}$$

where ϵ is a small constant. We use the upper bound with $1+\epsilon$ when the advantage $A^{\pi}(s,a)$ is positive. In this case, the objective is equal to

$$\mathcal{L}^{\mathsf{CLIP}}(s,a;\theta) = \min\left(\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}, (1+\epsilon)\right)A^{\pi}(s,a). \tag{5.23}$$

While the advantage $A^{\pi}(s,a)$ is positive, the minimum term puts a ceiling on the increased objective. Once $\pi(s,a;\theta) > (1+\epsilon)\pi(s,a;\theta_{old})$, the objective is limited by $(1+\epsilon)A^{\pi}(s,a)$. Similarly, when the advantage is negative, the objective can be written as follows:

$$\mathcal{L}^{\mathsf{CLIP}}(s,a;\theta) = \max\left(\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}, (1-\epsilon)\right)A^{\pi}(s,a). \tag{5.24}$$

When the advantage is negative, if $\pi(s,a;\theta)$ decreases the objective will increase. Thus, the maximum term puts a ceiling and once $\pi(s,a;\theta) < (1-\epsilon)\pi(s,a;\theta_{old})$, the objective is limited by $(1-\epsilon)A^{\pi}(s,a)$. These clipping surrogate methods restrict the new policy not going far from the old policy.

Furthermore, we use an advantage function $A^{\pi}(s,a)$ as follows [142]:

$$A^{\pi}(s,a) = r^{t} + \zeta V^{\pi}(s^{t+1}) - V^{\pi}(s^{t}) \tag{5.25}$$

The PPO algorithm is an on-policy algorithm where the UAVs' power level and the RIS's phase-shift matrix value are chosen by running the current policy $\pi(s,a;\theta)$ to maximise the EE performance in (5.13). In our paper, we use the clipping method in (5.22). For each iteration in the PPO algorithm, a set of trajectory $D = \{\tau_i\}$ are collected by running the current policy $\pi(s,a;\theta)$ in the environment. Then, the policy parameters are updated by running SGD with Adam optimiser.

We deploy the centralised and parallel learning based on the PPO algorithm, namely, centralise PPO (C-PPO) and parallel PPO algorithm (P-PPO). In the C-PPO algorithm, a policy $\pi(s, a; \theta)$ is used and trained to maximise the EE performance while we have $N + 1$ policy for the N UAVs and for the RIS in the P-PPO algorithm.

5.6 Simulation Results

For implementing our algorithms, we use the Tensorflow 1.13.1 [138]. In our paper, we consider a scenario with 3 UAVs, $N = 3$ to serve 3 clusters at the fixed location $(0, 0, 200), (200, 300, 200), (400, 0, 200)$. In each cluster, the number of UEs is set up to 10. Moreover, we assume $d/\lambda = 1/2$ for convenience. The total power consumption at the RIS and non-transmit power of UAV is set to $P_K + P_c = 4$W. For the neural network setting, we run the algorithm with different value of parameters and choose the the best performance with the learning rate $lr1 = 0.001$ and $lr2 = 0.002$ for the actor and critic network in the DDPG algorithm, respectively. In the PPO algorithm, we use learning rate $lr = 0.00001$. Other parameters are provided in Table 5.1. In this section, the four proposed schemes in previous sections are summarised as follows:

- **Our centralised DDPG algorithm (C-DDPG)**: As we explained in Section 5.3, we use the DDPG algorithm for jointly optimising the transmit power of the UAV and the phase-shift matrix of the RIS in a centralised manner.
- **Parallel learning for the DDPG method (P-DDPG)**: We consider

parallel learning to help to reduce the information transmission delay and errors while ensuring the network performance.

- **Our centralised PPO algorithm (C-PPO)**: Instead of using the DDPG algorithm, we use the PPO algorithm for solving the centralised problem.
- **Parallel learning for the PPO algorithm (P-PPO)**: We also deploy the PPO algorithm for parallel learning in our joint power allocation and phase-shift matrix optimisation in multi-UAV and RIS-assisted wireless networks.

In addition, to highlight the advantage of our proposals, we also compare our four proposed methods with the following schemes:

- **Max power transmission (MPT)**: We use the maximal transmit power at the UAV and optimise the phase-shift of the RIS by using the PPO algorithm.
- **Random selection scheme (RSS)**: We select randomly the phase-shift at the RIS and optimise the transmit power at the UAV.

Our proposed approaches achieve a better performance in comparison with MPT and RSS methods. Moreover, by using the neural networks, the processing time is small and the UAVs and RIS can choose the power allocation and phase-shift matrix immediately in milliseconds. Particularly, in Fig. 5.3, we show the EE performance of our proposed method in both centralised and decentralised learning with $M = 10$ and $K = 20$. The methods based on parallel learning reach the best results with the P-PPO algorithm. It is are higher than the ones using the C-DDPG and C-PPO algorithm in the centralised learning. The reason is

Table 5.1: Simulation parameter in Chapter 5

Parameters	Value
Bandwidth (W)	1 MHz
UAV transmission power	5 W
UAV's coverage	500 m
The RIS's position	$(500, 500, 30)$
Path-loss parameter	$\kappa_1 = 2, \kappa_2 = 2.2$
Channel power gain	$\beta_0 = -30$ dB
Rician factor	$\beta_1 = 4$
Noise power	$\alpha^2 = -134$ dBm
Discounting factor	$\zeta = 0.9$
Max number of UEs	30
Initial batch size	$D = 32$

that the P-PPO algorithm uses parallel learning and efficient sampling techniques. In the parallel learning, the optimisations at the UAVs and RIS are small-scale. The UAVs can choose the proper power and the RIS can choose the value of phase shift independently. Moreover, in the DDPG algorithm, we need to use a variable to make randomness to the action for *exploration* and *exploitation*. The performances are also affected by the initial parameter of the networks and the random walk model of users. Thus, the EE of the C-PPO algorithm and P-DDPG algorithm may be stuck in local optimal. The convergence of the P-PPO is fastest and following by the P-DDPG algorithm. As can be observed from this figure our proposed scheme with joint optimisation using the DRL techniques outperform the other approaches using the MPT and RSS methods.

In Fig. 5.4, the EE performance of our methods in comparison with other baseline schemes is presented with the different number of UEs in each cluster, M, for the number of RIS elements $K = 20$. Again, the P-PPO method shows

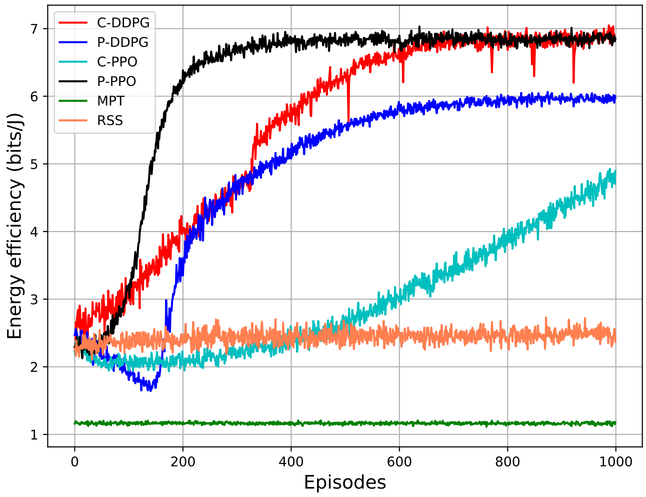


Figure 5.3: The EE with $M = 10$ and $K = 20$.

better EE performance than the centralised C-PPO and the ones using the C-DDPG algorithm. The MPT and RSS methods are less effective for the joint power allocation and phase shift matrix optimisation in the UAV-assisted wireless network with the support of the RIS.

In Fig. 5.5, we plot the EE performance versus the number of the RIS elements (K) when the number of UEs in each cluster equals to ten ($M = 10$). We achieve the best EE performance with the P-PPO algorithm despite the value of K. When the number of RIS elements becomes higher (e.g., $K > 25$), the methods based on the C-PPO algorithm are more effective than the ones using the DDPG algorithm. In contrast, for a smaller value of K, the methods based on the C-DDPG algorithm are better than the centralised learning with the C-PPO algorithm. For all values of K, the best performance can be achieved with P-PPO algorithm, which demonstrates the fact that the P-PPO algorithm is stable and practical for every environmental setting under the joint optimisation of power

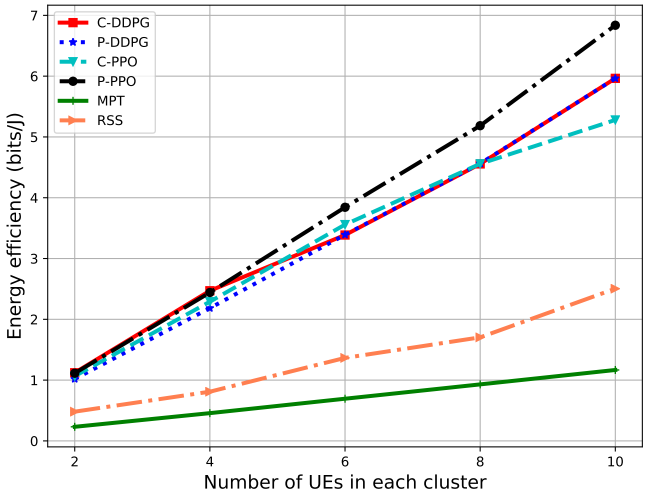


Figure 5.4: The EE versus the number of UEs in each cluster, M.

allocation at UAVs and the phase-shift matrix at RIS.

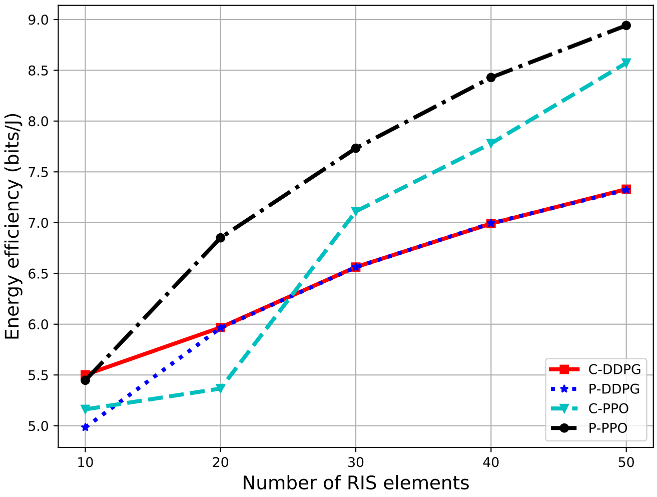


Figure 5.5: The EE versus the number of the RIS elements, K.

The EE performances of the DDPG algorithm versus episodes for different number of RIS elements using the centralised learning and parallel learning are shown in Fig. 5.6 and Fig. 5.7, respectively. With the higher number of RIS elements, the performance increase while the convergence rate is still similar for both centralised and parallel approaches. The result converges after about 600 episodes when the *exploration* is set to 3 and $\psi = 0.99995$. Thus, depending on the specific purpose, we can deploy the configurable RIS with fast learning.

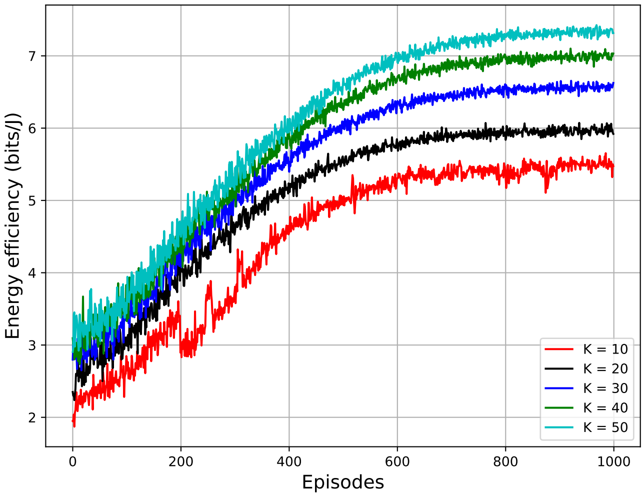


Figure 5.6: The EE of the C-DDPG algorithm with different number of the RIS elements, K.

Similarly, the EE performance of PPO algorithm versus episodes for different number of RIS elements using the centralised learning and parallel earning are plotted in Fig. 5.8 and Fig. 5.9, respectively. While the performance using centralised approach (C-PPO) is unstable and takes around 800 episodes for convergence, the parallel approach (P-PPO algorithm) shows a solid performance even when increasing the number of the RIS elements. The convergence for P-PPO is still stable and even faster with the higher number of RIS elements. We need only about 200 episodes for convergence. Furthermore, we use neural net-

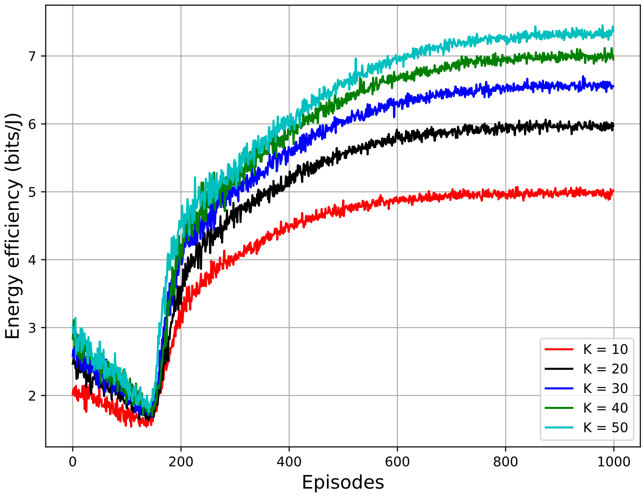


Figure 5.7: The EE of the P-DDPG algorithm with different number of the RIS elements, K.

works for the DDPG and PPO algorithm; thus, the system can be easily deployed after training and the agent can choose the action immediately.

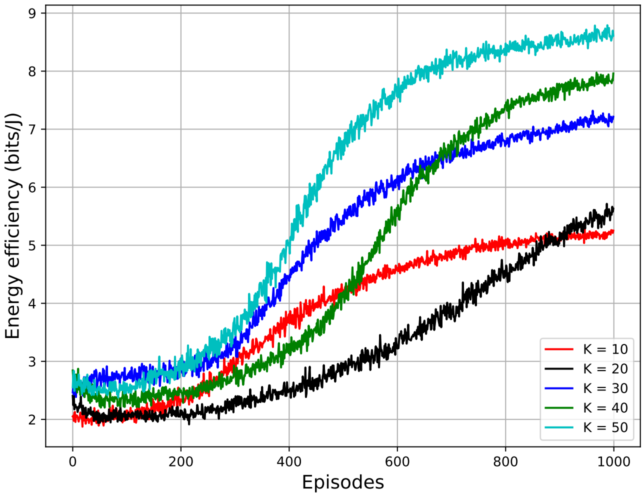


Figure 5.8: The EE of the C-PPO algorithm with different number of the RIS elements, K.

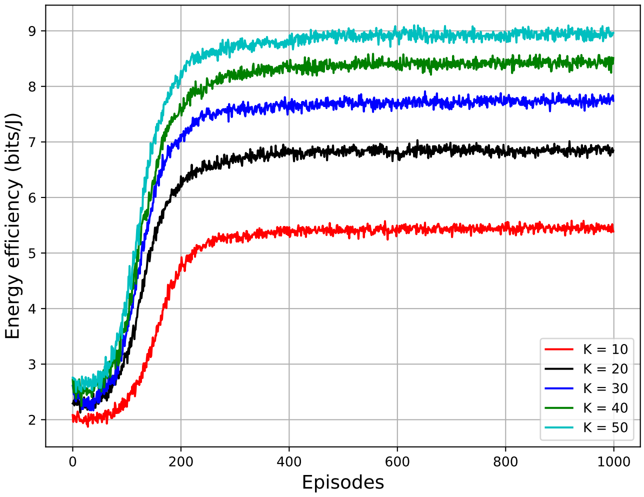


Figure 5.9: The EE of the P-PPO algorithm with different number of the RIS elements, K.

5.7 Conclusions

In this chapter, we have proposed multi-UAV networks supported by a RIS panel to enhance the network performance. To maximise the EE of the considered networks, the transmit power at the UAV and the phase-shift matrix at the RIS were jointly optimised by using the DDPG method and PPO technique in a centralised approach. Moreover, to reduce the network's delay and the power for exchanging the information, we proposed parallel learning for the optimisation problem. The results suggested that we can deploy the DRL algorithms for the real-time optimisation with impressive results compared to other baseline schemes. For the future work, we will improve the model with multiple RIS panels and cooperative communications with a fully autonomous ability in the futures.

Chapter 6

Conclusions and future work

6.1 Summary of the thesis

In this chapter, we highlight key points and major contributions of this thesis. We then present some potential research directions of transfer learning, UAV-assisted communications and RIS-aided networks.

6.1.1 UAVs and RIS-assisted Wireless Communications

In Chapter 1, we have introduced the potential benefits with applications and existing challenges of the UAVs and the RIS in wireless communications. We have also presented the research motivation of using DRL algorithms for solving complex problems in real-time optimisations.

6.1.2 Literature Review

In Chapter 2, we have presented a literature review of resource optimisation in the UAV-assisted wireless communications and the phase shift matrix optimisation in the RIS-aided networks. We have also introduced an overview of applications of the DRL algorithms in wireless communications.

6.1.3 3D UAV Trajectory and Data Collection Optimisation via Deep Reinforcement Learning

In Chapter 3, we have studied the 3D trajectory optimisation in data collection mission of UAV-assisted wireless communications. The DQL and dueling DQL algorithm with a low computational complexity have been presented to maximise the achieved sum-rate while minimising the moving path of the UAV to reach the landing dock. Our simulation results showed the efficiency of our techniques both in simple and complex environmental settings.

6.1.4 RIS-assisted UAV Communications for IoT with Wireless Power Transfer Using Deep Reinforcement Learning

In Chapter 4, we have proposed a novel framework for deploying the UAV in RIS-assisted wireless communications with the downlink power transfer and uplink information transmission protocol. We have proposed two DRL techniques for jointly optimising the UAV's trajectory, IoT's EH time scheduling and the phase shift matrix of the RIS to maximise the network's throughput. The results suggest that the systems learned by the DRL algorithm can deal with dynamic environments and satisfy some power restrictions and processing time in RIS-assisted UAV communications.

6.1.5 Reconfigurable Intelligent Surface-assisted Multi-UAV Networks: Efficient Resource Allocation with Deep Reinforcement Learning

In Chapter 5, we have presented a novel framework for RIS-assisted multi-UAV networks. We have proposed an efficient DRL method for the resource allocation problem to maximise the EE by using the DDPG method and the PPO technique in a centralised approach. Moreover, to reduce the network's delay and the power for the information exchange, we proposed parallel learning for the optimisation problem. The results suggested that we can deploy the DRL algorithms for real-time optimisation with impressive results compared to other baseline schemes.

6.2 Open Problems and Future Works

There are still many open problems that must be investigated in the future.

6.2.1 Multi-UAV Deployment and Trajectory Optimisation

For the multi-UAV networks, there are several key problems. First, there is a need to reduce the delay in the information transmission between UAVs. We can encourage each UAV to cooperate in distributed learning and avoid disruption. Secondly, the resource management and trajectory design in the multi-UAVs-assisted wireless networks is still challenging while we need to minimise the total flying path and energy consumption.

6.2.2 Resource Managements in Multi-RIS-aided Wireless Networks

In terms of open problems in multi-RIS networks, there is a need for new solutions to deploy multiple RISs cooperatively and effectively to achieve optimal performance. In this regard, key problems include: 1) developing robust RIS for massive users node scenarios, 2) designing dynamic and active RISs for improving the signal quality, 3) proposing a new framework for cooperating multi-RISs, 4) designing a framework for utilising the RIS in other emerging research areas, such as UAV-aided communications, mobile edge computing, ultra-reliable and low latency communications

6.2.3 Performance Analysis

There is a need for tractable presentation for the trade-off performance between the energy consumption and sum-rate in the UAV and RIS-assisted network. In addition, more analysis needs to be done to represent the trade-off between the flying energy consumption and EE in the UAV-enable networks. Finally, the performance evaluation of the multi-RISs needs to be analysed. For example, there is a need to study the deployment of the RISs and the number of elements in each RIS impact the performance of the total throughput and latency.

6.2.4 Transfer Learning and Meta Learning in Wireless Communications

There is a huge potential solution for using transfer learning and meta learning in wireless communications for improving the learning process. The essence of transfer learning is using the pre-trained model and knowledge to a new scenario. There are numerous benefits of using transfer learning in comparison to the conventional approaches in wireless communications, such as 1) enhancing quality and quantity of training data, 2) Speeding up the learning process, 3) Reducing computing demands in the training process, 4) mitigating communication overhead, and 5) protecting data privacy [143]. Another potential approach is meta-learning, which refers to learning algorithms that learn from the output of other learning algorithms to rapidly adapt to a new environment with a few training samples. Both transfer learning and meta learning are potentially useful in supporting the UAV-enabled wireless communications when the UAV will instantly have knowledge of optimal solutions of resource management, deployment and trajectory in new scenarios. They are also significantly effective in the RIS-assisted networks while the environment changes with a new user or a new RIS. The pre-trained model or other model output will help the RIS rapidly adjust the phase shift matrix to reach the optimal performance.

References

[1] S. Shakoor *et al.*, “Role of UAVs in public safety communications: Energy efficiency perspective,” *IEEE Access*, vol. 7, pp. 140 665–140 679, Sept. 2019.

[2] K. K. Nguyen, S. Khosravirad, D. B. D. Costa, L. D. Nguyen, and T. Q. Duong, “Reconfigurable intelligent surface-assisted Multi-UAV networks: Efficient resource allocation with deep reinforcement learning,” *IEEE J. Selected Topics in Signal Process.*, pp. 1–1, 2021, early Access.

[3] A. Masaracchia, Y. Li, K. K. Nguyen, C. Yin, S. R. Khosravirad, D. B. D. Costa, and T. Q. Duong, “UAV-enabled ultra-reliable low-latency communications for 6G: A comprehensive survey,” *IEEE Access*, vol. 9, pp. 137 338–137 352, Oct. 2021.

[4] Y. Zeng, J. Xu, and R. Zhang, “Energy minimization for wireless communication with rotary-wing UAV,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2329–2345, April 2019.

[5] 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.

[6] D. Yang, Q. Wu, Y. Zeng, and R. Zhang, "Energy tradeoff in ground-to-UAV communication via trajectory design," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6721–6726, Jul. 2018.

[7] 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.

[8] K. K. Nguyen, A. Masaracchia, T. Do-Duy, H. V. Poor, and T. Q. Duong, "RIS-assisted UAV communications for IoT with wireless power transfer using deep reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/2108.02889

[9] K. K. Nguyen, A. Masaracchia, C. Yin, L. D. Nguyen, O. A. Dobre, and T. Q. Duong, "Deep reinforcement learning for intelligent reflecting surface-assisted D2D communications," 2021. [Online]. Available: https://arxiv.org/abs/2108.02892

[10] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May. 2016.

[11] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys & Tut.*, vol. 21, no. 3, pp. 2334–2360, thirdquater 2019.

[12] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting

surfaces: A contemporary survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2283–2314, Fourthquater 2020.

[13] A. Vacca, H. Onishi, and F. Cuccu, "Drones: military weapons, surveillance or mapping tools for environmental monitoring? the need for legal framework is required," *Transportation research procedia*, vol. 25, pp. 51–62, 2017.

[14] T. Q. Duong, L. D. Nguyen, H. D. Tuan, and L. Hanzo, "Learning-aided realtime performance optimisation of cognitive UAV-assisted disaster communication," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019.

[15] L. D. Nguyen, K. K. Nguyen, A. Kortun, and T. Q. Duong, "Real-time deployment and resource allocation for distributed UAV systems in disaster relief," in *Proc. IEEE 20th International Workshop on Signal Processing Advances in Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.

[16] L. D. Nguyen, A. Kortun, and T. Q. Duong, "An introduction of realtime embedded optimisation programming for UAV systems under disaster communication," *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems*, vol. 5, no. 17, pp. 1–8, Dec. 2018.

[17] 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.

[18] K. K. Nguyen, N. A. Vien, L. D. Nguyen, M.-T. Le, L. Hanzo, and T. Q. Duong, "Real-time energy harvesting aided scheduling in UAV-assisted D2D networks relying on deep reinforcement learning," *IEEE Access*, vol. 9, pp. 3638–3648, Dec. 2021.

[19] "Drone trial to help Isle of Wight receive medical supplies faster during COVID19 pandemic." [Online]. Available: https://www.southampton.ac.uk/news/2020/04/drones-covid-iow.page

[20] "This Chilean community is using drones to deliver medicine to the elderly." [Online]. Available: https://www.weforum.org/agenda/2020/04/drone-chile-covid19/

[21] M. Gao, X. Xu, Y. Klinger, J. van der Woerd, and P. Tapponnier, "High-resolution mapping based on an unmanned aerial vehicle (UAV) to capture paleoseismic offsets along the Altyn-Tagh fault, China," *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, Aug. 2017.

[22] 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.

[23] S. Enayati, H. Saeedi, H. Pishro-Nik, and H. Yanikomeroglu, "Moving aerial base station networks: A stochastic geometry analysis and design perspective," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 2977–2988, June 2019.

[24] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.

[25] X. Jing, J. Sun, and C. Masouros, "Energy aware trajectory optimization for aerial base stations," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3352–3366, May 2021.

[26] Z. Wang, L. Duan, and R. Zhang, "Adaptive deployment for UAV-aided communication networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4531–4543, Sept. 2019.

[27] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and N. M. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100 480–100 490, Jul. 2019.

[28] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and L. D. Nguyen, "Distributed deep deterministic policy gradient for power allocation control in D2D-based V2V communications," *IEEE Access*, vol. 7, pp. 164 533–164 543, Nov. 2019.

[29] J. Gong, T.-H. Chang, C. Shen, and X. Chen, "Flight time minimization of UAV for data collection over wireless sensor networks," *IEEE J. Select. Areas Commun.*, vol. 36, no. 9, pp. 1942–1954, Sept. 2018.

[30] C. Zhong, M. C. Gursoy, and S. Velipasalar, “Deep reinforcement learning-based edge caching in wireless networks,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 48–61, Mar. 2020.

[31] H. Wu, Z. Wei, Y. Hou, N. Zhang, and X. Tao, “Cell-edge user offloading via flying UAV in non-uniform heterogeneous cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2411–2426, Apr. 2020.

[32] H. Huang *et al.*, “Deep reinforcement learning for UAV navigation through massive MIMO technique,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1117–1121, Jan. 2020.

[33] T. Q. Duong, L. D. Nguyen, and L. K. Nguyen, “Practical optimisation of path planning and completion time of data collection for UAV-enabled disaster communications,” in *Proc. 15th Int. Wireless Commun. Mobile Computing Conf. (IWCMC)*, Tangier, Morocco, Jun. 2019, pp. 372–377.

[34] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 3949–3963, Jun. 2016.

[35] B. Wang, Y. Sun, N. Zhao, and G. Gui, “Learn to coloring: Fast response to perturbation in UAV-assisted disaster relief networks,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3505–3509, Mar. 2020.

[36] S. Zhang and J. Liu, “Analysis and optimization of multiple unmanned aerial vehicle-assisted communications in post-disaster areas,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12 049–12 060, Dec. 2018.

[37] E. E. Haber, H. A. Alameddine, C. Assi, and S. Sharafeddine, "UAV-aided ultra-reliable low-latency computation offloading in future IoT networks," *IEEE Trans. Commun.*, vol. 69, no. 10, pp. 6838–6851, Oct. 2021.

[38] A. Ranjha and G. Kaddoum, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, Mar. 2021.

[39] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in UAV-enabled wireless-powered mobile-edge computing systems," *IEEE J. Select. Areas Commun.*, vol. 36, no. 9, pp. 1927–1941, Sept. 2018.

[40] Z. Yang, C. Pan, K. Wang, and M. Shikh-Bahaei, "Energy efficient resource allocation in UAV-enabled mobile edge computing networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4576–4589, Sept. 2019.

[41] Q. Hu, Y. Cai, G. Yu, Z. Qin, M. Zhao, and G. Y. Li, "Joint offloading and trajectory design for UAV-enabled mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1879–1892, Apr. 2019.

[42] M. Li, N. Cheng, J. Gao, Y. Wang, L. Zhao, and X. Shen, "Energy-efficient UAV-assisted mobile edge computing: Resource allocation and trajectory optimization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3424–3438, Mar. 2020.

[43] X. Yuan, T. Yang, Y. Hu, J. Xu, and A. Schmeink, "Trajectory design for UAV-enabled multiuser wireless power transfer with nonlinear energy

harvesting," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1105–1121, Feb. 2021.

[44] Y. Hu, X. Yuan, G. Zhang, and A. Schmeink, "Sustainable wireless sensor networks with UAV-enabled wireless power transfer," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8050–8064, Aug. 2021.

[45] 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.

[46] K. Li, W. Ni, E. Tovar, and A. Jamalipour, "On-board deep Q-network for UAV-assisted online power transfer and data collection," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12 215–12 226, Dec. 2019.

[47] H. Yan, Y. Chen, and S.-H. Yang, "UAV-enabled wireless power transfer with base station charging and UAV power consumption," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12 883–12 896, Nov. 2020.

[48] W. Feng, N. Zhao, S. Ao, J. Tang, X. Zhang, Y. Fu, D. K. C. So, and K.-K. Wong, "Joint 3D trajectory design and time allocation for UAV-enabled wireless power transfer networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9265–9278, Sept. 2020.

[49] 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.

[50] Y. Che, Y. Lai, S. Luo, K. Wu, and L. Duan, “UAV-aided information and energy transmissions for cognitive and sustainable 5G networks,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1668–1683, Mar. 2021.

[51] P. Wu, F. Xiao, C. Sha, H. Huang, and L. Sun, “Trajectory optimization for UAVs’ efficient charging in wireless rechargeable sensor networks,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 4, pp. 4207–4220, Apr. 2020.

[52] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, “UAV-enabled intelligent transportation systems for the smart city: Applications and challenges,” *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 22–28, Mar. 2017.

[53] H. Kim, L. Mokdad, and J. Ben-Othman, “Designing UAV surveillance frameworks for smart city and extensive ocean with differential perspectives,” *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 98–104, Mar. 2018.

[54] Q. Liu, J. Wu, P. Xia, S. Zhao, Y. Yang, W. Chen, and L. Hanzo, “Charging unplugged: Will distributed laser charging for mobile wireless power transfer work?” *IEEE Vehicular Technology Magazine*, vol. 11, no. 4, pp. 36–45, Dec. 2016.

[55] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. D. Renzo, and N. Al-Dhahir, “Reconfigurable intelligent surfaces: Principles and opportunities,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 3, pp. 1546–1577, thirdquater 2021.

[56] L. Ge, P. Dong, H. Zhang, J.-B. Wang, and X. You, “Joint beamforming and trajectory optimization for intelligent reflecting surfaces-assisted UAV communications,” *IEEE Access*, vol. 8, pp. 78 702–78 712, Apr. 2020.

[57] S. Li, B. Duo, X. Yuan, Y.-C. Liang, and M. D. Renzo, “Reconfigurable intelligent surface assisted UAV communication: Joint trajectory design and passive beamforming,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 716–720, May 2020.

[58] X. Hu, C. Masouros, and K.-K. Wong, “Reconfigurable intelligent surface aided mobile edge computing: From optimization-based to location-only learning-based solutions,” *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3709–3725, June 2021.

[59] E. Basar, M. D. Renzo, J. D. Rosny, M. Debbah, M.-S. Alouini, and R. Zhang, “Wireless communications through reconfigurable intelligent surfaces,” *IEEE Access*, vol. 7, pp. 116 753–116 773, Aug. 2019.

[60] S. Atapattu, R. Fan, P. Dharmawansa, G. Wang, J. Evans, and T. A. Tsiftsis, “Reconfigurable intelligent surface assisted two–way communications: Performance analysis and optimization,” *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6552–6567, Oct. 2020.

[61] Y. Li, M. Jiang, Q. Zhang, and J. Qin, “Joint beamforming design in multi-cluster MISO NOMA reconfigurable intelligent surface-aided downlink communication networks,” *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 664–674, Jan. 2021.

[62] Y. Chen, B. Ai, H. Zhang, Y. Niu, L. Song, Z. Han, and H. V. Poor, "Reconfigurable intelligent surface assisted device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2792–2804, May 2021.

[63] S. Jia, X. Yuan, and Y.-C. Liang, "Reconfigurable intelligent surfaces for energy efficiency in D2D communication network," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 683–687, March 2021.

[64] C. Pradhan, A. Li, L. Song, J. Li, B. Vucetic, and Y. Li, "Reconfigurable intelligent surface (RIS)-enhanced two-way OFDM communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 270–16 275, Dec. 2020.

[65] Y. Cao, T. Lv, W. Ni, and Z. Lin, "Sum-rate maximization for multi-reconfigurable intelligent surface-assisted device-to-device communications," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7283–7296, Nov. 2021.

[66] G. Yang, Y. Liao, Y.-C. Liang, O. Tirkkonen, G. Wang, and X. Zhu, "Reconfigurable intelligent surface empowered device-to-device communication underlaying cellular networks," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7790–7805, Nov. 2021.

[67] Z. Zhang, C. Zhang, C. Jiang, F. Jia, J. Ge, and F. Gong, "Improving physical layer security for reconfigurable intelligent surface aided NOMA 6G networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4451–4463, May 2021.

[68] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless

communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[69] S. Li, B. Duo, M. D. Renzo, M. Tao, and X. Yuan, "Robust secure UAV communications with the aid of reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, vol. 20, no. 10, pp. 6402–6417, Oct. 2021.

[70] X. Cao, B. Yang, C. Huang, C. Yuen, M. D. Renzo, D. Niyato, and Z. Han, "Reconfigurable intelligent surface-assisted aerial-terrestrial communications via multi-task learning," *IEEE J. Select. Areas Commun.*, vol. 39, no. 10, pp. 3035–3050, Oct. 2021.

[71] M. Shokry, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[72] Y. U. Ozcan, O. Ozdemir, and G. K. Kurt, "Reconfigurable intelligent surfaces for the connectivity of autonomous vehicles," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2508–2513, Mar. 2021.

[73] Z. Chu, Z. Zhu, F. Zhou, M. Zhang, and N. Al-Dhahir, "Intelligent reflecting surface assisted wireless powered sensor networks for internet of things," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4877–4889, July 2021.

[74] H. Zhang and L. Hanzo, "Federated learning assisted multi-UAV networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14 104–14 109, Nov. 2020.

[75] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, "Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7957–7969, Aug. 2019.

[76] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.

[77] X. Liu, Y. Liu, and Y. Chen, "Reinforcement learning in multiple-UAV networks: Deployment and movement design," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8036–8049, Aug. 2019.

[78] 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.

[79] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE International Conf. Robot. Autom. (ICRA)*, May 2017, pp. 3389–3396.

[80] Q. Cai, A. Filos-Ratsikas, P. Tang, and Y. Zhang, "Reinforcement mechanism design for fraudulent behaviour in e-commerce," in *Proc. Thirty-Second AAAI Conf. Artif. Intell.*, 2018.

[81] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," 2013. [Online]. Available: https://arxiv.org/abs/1312.5602

[82] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Select. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[83] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.

[84] S. Yin, S. Zhao, Y. Zhao, , and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Aug. 2019.

[85] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons, Inc., 1994.

[86] D. P. Bertsekas, *Dynamic Programming and Optimal Control.* Athena Scientific Belmont, MA, 1995, vol. 1, no. 2.

[87] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.

[88] Y. Zeng, X. Xu, S. Jin, and R. Zhang, "Simultaneous navigation and radio mapping for cellular-connected UAV with deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4205–4220, Jul. 2021.

[89] H. Wang, G. Ren, J. Chen, G. Ding, and Y. Yang, "Unmanned aerial vehicle-aided communications: Joint transmit power and trajectory optimization," *IEEE Wireless Commun. Lett.*, vol. 7, no. 4, pp. 522–525, Aug. 2018.

[90] H. Wang, J. Wang, G. Ding, J. Chen, F. Gao, and Z. Han, "Completion time minimization with path planning for fixed-wing UAV communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3485–3499, Jul. 2019.

[91] Z. Wang, R. Liu, Q. Liu, J. S. Thompson, and M. Kadoch, "Energy-efficient data collection and device positioning in UAV-assisted IoT," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1122–1139, Feb. 2020.

[92] J. Li *et al.*, "Joint optimization on trajectory, altitude, velocity, and link scheduling for minimum mission time in UAV-aided data collection," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1464–1475, Feb. 2020.

[93] M. Samir, S. Sharafeddine, C. M. Assi, T. M. Nguyen, and A. Ghrayeb, "UAV trajectory planning for data collection from time-constrained IoT devices," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 34–46, Jan. 2020.

[94] M. Hua, L. Yang, Q. Wu, and A. L. Swindlehurst, "3D UAV trajectory and communication design for simultaneous uplink and downlink transmission," *IEEE Trans. on Commun.*, vol. 68, no. 9, pp. 5908–5923, Sept. 2020.

[95] C. Zhan and Y. Zeng, “Aerial–ground cost tradeoff for multi-UAV-enabled data collection in wireless sensor networks,” *IEEE Trans. on Commun.*, vol. 68, no. 3, pp. 1937–1950, Mar. 2020.

[96] S. Zhang and R. Zhang, “Radio map-based 3D path planning for cellular-connected UAV,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1975–1989, Mar. 2021.

[97] S. H. Chae, C. Jeong, and S. H. Lim, “Simultaneous wireless information and power transfer for internet of things sensor networks,” *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2829–2843, Aug. 2018.

[98] F. Zhu, F. Gao, Y. C. Eldar, and G. Qian, “Robust simultaneous wireless information and power transfer in beamspace massive MIMO,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 9, pp. 4199–4212, Sept. 2019.

[99] K. Li, W. Ni, L. Duan, M. Abolhasan, and J. Niu, “Wireless power transfer and data collection in wireless sensor networks,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2686–2697, Mar. 2018.

[100] J. Chen, L. Zhang, Y.-C. Liang, X. Kang, and R. Zhang, “Resource allocation for wireless-powered IoT networks with short packet communication,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 1447–1461, Feb. 2019.

[101] Y. Zou, S. Gong, J. Xu, W. Cheng, D. T. Hoang, and D. Niyato, “Wireless powered intelligent reflecting surfaces for enhancing wireless communications,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 369–12 373, Oct. 2020.

[102] C. Huang, R. Mo, and C. Yuen, "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1839–1850, Aug. 2020.

[103] C. Pan *et al.*, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[104] H. Yang, X. Yuan, J. Fang, and Y.-C. Liang, "Reconfigurable intelligent surface aided constant-envelope wireless power transfer," *IEEE Trans. Signal Process.*, vol. 69, pp. 1347–1361, Feb. 2021.

[105] S. Lin, B. Zheng, G. C. Alexandropoulos, M. Wen, M. D. Renzo, and F. Chen, "Reconfigurable intelligent surfaces with reflection pattern modulation: Beamforming design and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 741–754, Feb. 2021.

[106] O. Rezaei, M. M. Naghsh, Z. Rezaei, and R. Zhang, "Throughput optimization for wireless powered interference channels," *IEEE Trans. Wireless Commun.*, vol. 18, no. 5, pp. 2464–2476, May 2019.

[107] Z. Chu, F. Zhou, Z. Zhu, R. Q. Hu, and P. Xiao, "Wireless powered sensor networks for internet of things: Maximum throughput and optimal power allocation," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 310–321, Feb. 2018.

[108] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4679–4691, Oct. 2019.

[109] Y. Xu, T. Zhang, Y. Liu, D. Yang, L. Xiao, and M. Tao, “UAV-assisted MEC networks with aerial and ground cooperation,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7712–7727, Dec. 2021.

[110] G. Yang, R. Dai, and Y.-C. Liang, “Energy-efficient UAV backscatter communication with joint trajectory design and resource optimization,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 926–941, Feb. 2021.

[111] A. Alsharoa and M. Yuksel, “Energy efficient D2D communications using multiple UAV relays,” *IEEE Trans. Commun.*, vol. 69, no. 8, pp. 5337–5351, Aug. 2021.

[112] M.-N. Nguyen, L. D. Nguyen, T. Q. Duong, and H. D. Tuan, “Real-time optimal resource allocation for embedded UAV communication systems,” *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 225–228, Feb. 2019.

[113] H. T. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, and W.-J. Hwang, “Joint D2D assignment, bandwidth and power allocation in cognitive UAV-enabled networks,” *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 3, pp. 1084–1095, Sept. 2020.

[114] H. Yu, H. D. Tuan, A. A. Nasir, T. Q. Duong, and H. V. Poor, “Joint design of reconfigurable intelligent surfaces and transmit beamforming under proper and improper Gaussian signaling,” *IEEE J. Select. Areas Commun.*, vol. 38, no. 11, pp. 2589–2603, Nov. 2020.

[115] H. Guo, Y.-C. Liang, J. Chen, and E. G. Larsson, “Weighted sum-rate maximization for reconfigurable intelligent surface aided wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3064–3076, May 2020.

[116] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen, "Reconfigurable intelligent surfaces for energy efficiency in wireless communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 4157–4170, Aug. 2019.

[117] J. Yuan, Y.-C. Liang, J. Joung, G. Feng, and E. G. Larsson, "Intelligent reflecting surface-assisted cognitive radio system," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 675–687, Jan. 2021.

[118] W. Yan, X. Yuan, Z.-Q. He, and X. Kuai, "Passive beamforming and information transfer design for reconfigurable intelligent surfaces aided multiuser MIMO systems," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1793–1808, Aug. 2020.

[119] B. Di, H. Zhang, L. Song, Y. Li, Z. Han, and H. V. Poor, "Hybrid beamforming for reconfigurable intelligent surface based multi-user communications: Achievable rates with limited discrete phase shifts," *IEEE J. Select. Areas Commun.*, vol. 38, no. 8, pp. 1809–1822, Aug. 2020.

[120] P. Luong, F. Gagnon, L.-N. Tran, and F. Labeau, "Deep reinforcement learning-based resource allocation in cooperative UAV-assisted wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7610–7625, Nov. 2021.

[121] Y. Yu, J. Tang, J. Huang, X. Zhang, D. K. C. So, and K.-K. Wong, "Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6361–6374, Sept. 2021.

[122] F. Wu, H. Zhang, J. Wu, Z. Han, H. V. Poor, and L. Song, "UAV-to-device underlay communications: Age of information minimization by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4461–4475, July 2021.

[123] R. Ding, F. Gao, and X. S. Shen, "3D UAV trajectory design and frequency band allocation for energy-efficient and fair communication: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 7796–7809, Dec. 2020.

[124] M. Samir, C. Assi, S. Sharafeddine, D. Ebrahimi, and A. Ghrayeb, "Age of information aware trajectory planning of UAVs in intelligent transportation systems: A deep learning approach," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12 382–12 395, Nov. 2020.

[125] F. Wu, H. Zhang, J. Wu, and L. Song, "Cellular UAV-to-device communications: Trajectory design and mode selection by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4175–4189, July 2020.

[126] H. Jaeger, "The "echo state" approach to analysing and training recurrent neural networks-with an erratum note," *" GMD - German National Research Institute for Computer Science, Tech. Rep.*, vol. 148, no. 34, p. 13, Jan. 2010.

[127] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. 4th International Conf. on Learning Representations (ICLR)*, 2016.

[128] K. K. Nguyen, T. Q. Duong, T. Do-Duy, H. Claussen, and L. Hanzo, "3D UAV trajectory and data collection optimisation via deep reinforcement learning," 2021. [Online]. Available: https://arxiv.org/abs/2106.03129

[129] Y. Chen, B. Ai, H. Zhang, Y. Niu, L. Song, Z. Han, and H. V. Poor, "Reconfigurable intelligent surface assisted device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2792–2804, May 2021.

[130] L. Wang, K. Wang, C. Pan, W. Xu, and N. Aslam, "Joint trajectory and passive beamforming design for intelligent reflecting surface-aided UAV communications: A deep reinforcement learning approach," 2020. [Online]. Available: https://arxiv.org/abs/2007.08380

[131] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems," *IEEE Wireless Commun. Lett.*, vol. 9, no. 5, pp. 745–749, May 2020.

[132] B. Sheen, J. Yang, X. Feng, and M. M. U. Chowdhury, "A deep learning based modeling of reconfigurable intelligent surface assisted wireless communications for phase shift configuration," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 262–272, 2021.

[133] Z. Yang, M. Chen, W. Saad, W. Xu, M. Shikh-Bahaei, H. V. Poor, and S. Cui, "Energy-efficient wireless communications with distributed reconfigurable intelligent surfaces," *IEEE Trans. Wireless Commun.*, 2021, early Access.

[134] X. Liu, Y. Liu, and Y. Chen, “Machine learning empowered trajectory and passive beamforming design in UAV-RIS wireless networks,” *IEEE J. Select. Areas Commun.*, vol. 39, no. 7, pp. 2042–2055, Jul. 2021.

[135] H. Claussen, “Distributed algorithms for robust self-deployment and load balancing in autonomous wireless access networks,” in *Proc. IEEE Int. Conf. on Commun. (ICC)*, vol. 4, Istanbul, Turkey, Jun. 2006, pp. 1927–1932.

[136] X. Li, H. Yao, J. Wang, X. Xu, C. Jiang, and L. Hanzo, “A near-optimal UAV-aided radio coverage strategy for dense urban areas,” *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9098–9109, Sept. 2019.

[137] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, “Dueling network architectures for deep reinforcement learning,” 2015. [Online]. Available: https://arxiv.org/abs/1511.06581

[138] M. Abadi *et al.*, “Tensorflow: A system for large-scale machine learning,” in *Proc. 12th USENIX Sym. Opr. Syst. Design and Imp. (OSDI 16)*, Nov. 2016, pp. 265–283.

[139] D. P. Kingma and J. L. Ba, “Adam: A method for stochastic optimization,” 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[140] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, “High-dimensional continuous control using generalized advantage estimation,” in *Proc. 4th International Conf. Learning Representations (ICLR)*, 2016.

[141] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” 2017.

[142] V. Mnih *et al.*, “Asynchronous methods for deep reinforcement learning,” in *Proc. Int. Conf. Mach. Learn.* PMLR, 2016, pp. 1928–1937.

[143] C. T. Nguyen, N. V. Huynh, N. H. Chu, Y. M. Saputra, D. T. Hoang, D. N. Nguyen, Q. Pham, D. Niyato, E. Dutkiewicz, and W. Hwang, “Transfer learning for future wireless networks: A comprehensive survey,” 2021. [Online]. Available: arXivpreprintarXiv:2102.07572