



Degree Project in Technology

First cycle, 30 credits

Joint Trajectory Designing and Association Scheduling for UAV-Enabled Data Collection and Charging System in Space-Air-Ground Integrated Networks

SHEN LIU

Joint Trajectory Designing and Association Scheduling for UAV-Enabled Data Collection and Charging System in Space-Air-Ground Integrated Networks

SHEN LIU

Master's Programme, Communication Systems, 120 credits

Date: December 9, 2025

Supervisor: Irshad Ahmad Meer

Examiner: Cicek Cavdar

School of Electrical Engineering and Computer Science

Swedish title: Gemensam banutformning och associationsschemaläggning för ett UAV-baserat system för datainsamling och laddning i ett rymd-luft-mark-integrerat nätverk

Abstract

In remote areas, ground users (GUs) often lack direct access to terrestrial network infrastructure, which poses significant challenges for data transmission. To address this issue, we propose a solution based on a Space-Air-Ground Integrated Network (SAGIN), where an Unmanned Aerial Vehicle (UAV) is deployed to simultaneously collect data from GUs and provide wireless power transfer (WPT). Furthermore, a Low Earth Orbit (LEO) satellite network is incorporated to enable the UAV to forward the collected data to the satellite, thereby connecting to the core network.

To improve data collection efficiency while maintaining user fairness, we formulate an optimization problem in which the UAV dynamically selects its flight direction and determines its associations with multiple users. The objective is to maximize the total amount of collected data while ensuring that as many users as possible are served. Additionally, the wirelessly transferred energy to each user must exceed their communication energy consumption.

To capture the real-time dynamics of satellite communication, we consider a dynamic LEO satellite constellation and model the instantaneous positions of all satellites, leading to a time-varying UAV-satellite uplink data rate. Due to the high mobility of both the UAV and the satellites, the considered optimization problem is highly dynamic and complex, making it difficult to solve using conventional mathematical methods. Therefore, we employ the Deep Reinforcement Learning (DRL) algorithm Proximal Policy Optimization (PPO) to address this problem.

To evaluate the effectiveness of our proposed algorithm, we design two benchmark methods: Optimized Trajectory and Single Association (OTSA) and Optimized Trajectory and Conditional Association (OTCA). The proposed algorithm outperforms OTCA and OTSA by 21.8% and 224.01%, respectively, in terms of overall performance in data collection and user fairness. Moreover, we observe a trade-off between data collection and user fairness due to the limited number of time steps. Specifically, our algorithm achieves 9.33% lower performance in data collection compared to OTCA, but outperforms OTCA by 64.53% in user fairness. This is because, within the limited time steps, it is not feasible for the UAV to approach every user. Instead, our algorithm intelligently directs the UAV to associate with more distant users when necessary, thereby ensuring that more users receive service. These results further demonstrate the effectiveness of our proposed method.

Keywords

Deep Reinforcement Learning (DRL), Low Earth Orbit (LEO) satellite, Unmanned Aerial Vehicle (UAV), Trajectory optimization, Wireless Power Transfer (WPT)

Sammanfattning

I avlägsna områden saknar markanvändare (GUs) ofta direkt tillgång till markbunden nätverksinfrastruktur, vilket utgör betydande utmaningar för datakommunikation. För att hantera detta problem föreslår vi en lösning baserad på ett rymd-luft-mark-integrerat nätverk (Space-Air-Ground Integrated Network, SAGIN), där en obemannad luftfarkost (Unmanned Aerial Vehicle, UAV) används för att samtidigt samla in data från GUs och tillhandahålla trådlös energitransfer (Wireless Power Transfer, WPT). Dessutom integreras ett nätverk av låg omloppsbanasatelliter (Low Earth Orbit, LEO) för att möjliggöra att den insamlade datan överförs från UAV till satellit, vilket i sin tur kopplar upp systemet mot kärnnätverket.

För att förbättra datainsamlingens effektivitet samtidigt som användarrättvisa bevaras, formulerar vi ett optimeringsproblem där UAV:n dynamiskt väljer sin flygriktning och bestämmer associeringar med flera användare. Målet är att maximera den totala mängden insamlad data samtidigt som så många användare som möjligt betjänas. Dessutom måste den trådlöst överförda energin till varje användare överstiga deras energiförbrukning för kommunikation.

För att fånga den realtidsdynamik som satellitkommunikation innebär, modellerar vi en dynamisk LEO-satellitkonstellation genom att bestämma de aktuella positionerna för samtliga satelliter, vilket resulterar i en tidsvarierande uppgående dataakt mellan UAV och satelliter. På grund av den höga rörligheten hos både UAV och satelliter är det resulterande optimeringsproblemet mycket dynamiskt och komplext, vilket gör det svårt att lösa med traditionella matematiska metoder. Därför tillämpar vi algoritmen Proximal Policy Optimization (PPO), en metod inom djup förstärkningsinlärning (Deep Reinforcement Learning, DRL), för att lösa problemet.

För att utvärdera effektiviteten hos vår föreslagna algoritm designar vi två jämförelsemetoder: Optimized Trajectory and Single Association (OTSA) och Optimized Trajectory and Conditional Association (OTCA). Den föreslagna algoritmen överträffar OTCA och OTSA med 21,8% respektive 224,01% i total prestanda när det gäller datainsamling och användarrättvisa. Vi observerar dessutom en avvägning mellan datainsamling och rättvisa, till följd av det begränsade antalet tidssteg. Specifikt uppnår vår algoritm 9,33% lägre prestanda i datainsamling jämfört med OTCA, men överträffar OTCA med 64,53% i användarrättvisa. Detta beror på att UAV:n inte hinner nå alla användare inom ett begränsat tidsfönster. I stället styr vår algoritm UAV:n att

prioritera även mer avlägsna användare när det behövs, vilket säkerställer att fler användare får service. Dessa resultat bekräftar ytterligare effektiviteten hos vår metod.

Nyckelord

Djup förstärkningsinlärning, Satellit i låg omlopps bana, Obemannat luftfartyg, Optimering av flygbana, Trådlös energioverföring

Acknowledgments

First, I would like to express my sincere gratitude to my examiner, professor Cicek Cavdar, and my supervisors, Dr. Irshad Ahmad Meer and Dr. Shuai Zhang. I have joined the research group in 2023 with no prior experience in academic research. You are always so patient with my questions which are probably basic and stupid. Over the past two years under your guidance, I have made a great progress in research skills and academic writing. I deeply appreciate your always help and support.

Then I would like to thank my parents and my grandparents. Thank you for your love and unwavering support. Whenever I feel anxious, just thinking of you reminds me that I am not alone. There are no words can express how much I appreciate it.

I also want to thank my friends. Thank you for always being there. I cannot imagine the days without you mates. It's truly amazing that we can gather together and become such close friends.

Lastly, I want to thank my dogs and cat. You may not understand everything and probably cannot read this, but I still want to write here how much you mean to me. Hope you can always be healthy and happy.

Stockholm, December 2025

Shen Liu

Contents

1	Introduction	1
1.1	Background	1
1.2	Problem	2
1.2.1	Original problem and definition	2
1.2.2	Scientific and engineering	3
1.3	Purpose	3
1.4	Goals	3
1.5	Research Methodology	4
1.6	Delimitations	5
1.7	Structure of the thesis	5
2	Background	7
2.1	SAGIN System	7
2.1.1	Space Segment	8
2.1.2	Air segment	8
2.1.3	Ground Segment	9
2.2	Related works	10
2.3	Summary	13
3	System Design of SAGIN with dynamic LEO satellite constellation	15
3.1	System Architecture	15
3.1.1	Ground User System	16
3.1.2	UAV System	16
3.1.3	Dynamic Walker-Delta Constellation	18
3.2	System Model	20
3.2.1	Communication Model	20
3.2.2	Energy Consumption and Harvest Model	22
3.2.3	User fairness	23
3.3	Problem Formulation	24

4 PPO-based UAV trajectory designing and user association scheduling algorithm	25
4.1 Proximal Policy Optimization Algorithm	25
4.2 Algorithm Implementation in SAGIN System	29
4.2.1 Summary	33
5 Results and Analysis	35
5.1 Dynamic Multi-orbit LEO Satellite Constellation	35
5.2 Benchmark Algorithms	38
5.3 Sensitivity Analysis for Weight Factors	41
5.4 Performance Analysis	44
5.5 Summary	49
6 Conclusions and Future work	51
6.1 Conclusions	51
6.2 Limitations	52
6.3 Future work	52
References	55

List of Figures

3.1	SAGIN system architecture	16
3.2	UAV flying model	17
3.3	Walker-Delta constellation pattern for LEO satellites	18
3.4	The projection of a satellite onto coordinate planes	19
4.1	The framework of general RL algorithms	26
4.2	Illustration of the Clipped Surrogate Objective in PPO	28
4.3	The framework of the proposed algorithm	31
5.1	Multi-orbit Walker-Delta LEO satellite constellation	36
5.2	Predicted real-time uplink data rate between UAV and satellite	37
5.3	The average returns of the proposed algorithm and benchmarks	39
5.4	The average returns for data collection	40
5.5	The average returns for user fairness	41
5.6	Effect of fairness weight on overall rewards	42
5.7	Effect of fairness weight on data collection rewards	43
5.8	Effect of fairness weight on user fairness rewards	44
5.9	The average overall returns during training period	45
5.10	UAV trajectory with unevenly distributed users and uniform data collection requirements	46
5.11	UAV trajectory with unevenly distributed users and varying data collection requirements	46
5.12	UAV trajectory with evenly distributed users and uniform data collection requirements	47
5.13	UAV trajectory with evenly distributed users and varying data collection requirements	48
5.14	Number of undercharged user	49

List of Tables

2.1	Comparison of three types of satellites	8
5.1	Parameters for Satellite Constellation Construction	36
5.2	Environment parameters	37
5.3	Simulation parameters	38

Chapter 1

Introduction

With the improvement of wireless communication technologies, 6G aims to establish a global ubiquitous mobile broadband communication system [1]. However, traditional Terrestrial Networks (TN) face significant challenges in deploying in remote areas due to inherent limitations of ground infrastructure, where numerous Internet of Remote Things (IoRT) devices could be deployed for certain applications. To address this limitation, Non-terrestrial Networks (NTN), which includes Unmanned Aerial Vehicle (UAV), High Altitude Platform (HAP) and satellite network, have emerged as a promising solution [2, 3].

This thesis focuses on using NTN including UAV and LEO satellite to address the challenge of energy supply and data uploading for IoRT devices. Under the considered setup, we aim to jointly optimize the UAV flying trajectory and user association scheduling to maximize overall data collection and enhance user fairness, while simultaneously satisfying the charging requirements of the IoRT devices. Finally, we employ reinforcement learning to find a near-optimal solution for this optimization problem.

1.1 Background

In recent years, a growing number of IoRT devices have been deployed in remote areas executing tasks such as industrial automation and real-time environment monitoring [4, 5]. However, these devices face the significant challenge of having no access to the conventional terrestrial infrastructure [6], preventing them from transmitting computational tasks to centralized servers for processing. To tackle this problem, UAVs are dispatched as flying base stations that can collect the tasks generated by IoRT devices [7].

Additionally, UAVs can provide charging services for these devices to sustain their operations.

The integration of LEO satellites introduces additional complexities due to their high-speed movement. The fast motion of satellites leads to frequent handovers between ground users and satellites, along with fluctuations in communication data rates [8]. To accurately simulate these dynamic changes in UE-satellite links, it is necessary to model the real-time coordinates of LEO satellites with the constellation and incorporate it into the system model [9].

Furthermore, the deployment of UAVs adds another layer of complexity. The high mobility of UAVs causes their trajectories and user association decisions to significantly impact the real-time throughput of the link between the users and UAVs, which directly affects the efficiency of data collection and device charging [10]. Therefore, the effective design and scheduling of UAV trajectories and user association mechanisms are essential to optimizing overall system performance.

1.2 Problem

This thesis addresses an optimization problem for non-terrestrial network in remote area, aiming to maximize the data collection and user coverage.

1.2.1 Original problem and definition

In remote and unreachable areas, IoRT devices face two major challenges: lack of access to conventional terrestrial network and the difficulty of obtaining a reliable energy supply. To address these issues, we deploy an UAV to hover over a designated region to simultaneously collect data from the IoRT devices and provide energy transfer. To maximize the overall data collection and improve user fairness, we focus on two key optimization variables: UAV flying trajectory designing and user association scheduling.

In addition, the data upload rate from users also depends on the UAV-satellite communication link, which is determined by the minimum data rate between the GU-UAV and UAV-satellite links. Thus, selecting the appropriate satellite for handover is crucial [11, 12]. However, for the UAV to make handover decisions, it must obtain the real-time link states of multiple candidate satellites. This process can introduce considerable communication overhead and may be vulnerable to link quality fluctuations, potentially delaying or degrading decision accuracy. Therefore, modeling the satellite

constellation and representing their real-time positions is essential for enabling timely and reliable handover decisions.

1.2.2 Scientific and engineering

The scenario we consider in this thesis is highly complex and dynamic due to the high velocity of LEO satellites and the mobility of the UAV. Given these challenges, it is crucial for the UAV to make real-time decisions based on the current state of the environment. Consequently, this optimization problem is formulated as a non-convex mix integer nonlinear problem (MINLP) which is extremely difficult to solve due to its inherent computational complexity.

To address this, we employ deep reinforcement learning to solve this problem. However, our objective is to jointly optimize the UAV trajectory and device association, which increases the dimensions of the action space, leading to convergence challenges in Deep Reinforcement Learning (DRL) algorithms. Due to this problem, most existing studies opt to simplify the problem by optimizing only one factor at a time, thereby sacrificing overall system performance.

1.3 Purpose

The purpose of this thesis is to develop a UAV and LEO satellite-based simultaneous data collection and charging system to address the communication and energy supply challenges faced by users or IoRT devices in remote areas.

This thesis introduces an innovative approach by integrating wireless charging and communication into the UAV-user link. Additionally, it accounts for the dynamic motion of LEO satellites, which affect the time-varying data rate in satellite communication, a factor often overlooked in existing papers. Furthermore, DRL is employed to optimize multiple variables simultaneously, enhancing overall system efficiency. By addressing these critical aspects, this thesis fills an important research gap in this field.

1.4 Goals

The main objective of this thesis is to develop a reinforcement learning-based intelligent UAV trajectory designing and user association scheduling

algorithm in complex environments, which enables the UAV to make real-time decision based on the current state of environment in order to maximize data collection and user coverage simultaneously. The tasks involved in the thesis include three main parts:

1. Modeling the dynamic environment
This involves capturing the motion dynamics of LEO satellites within Walker-Delta constellation and incorporating them into the system model.
2. Designing a DRL-based algorithm
A DRL-based framework will be developed to address the proposed scenario, enabling real-time decision-making for UAV-user association and UAV trajectory optimization.
3. Performance evaluation
The proposed algorithm will be evaluated against at least two benchmark methods to validate its effectiveness.

After achieving these goals, the main deliverables of this thesis include:

1. The proposed system enables simultaneous wireless energy transfer via the UAV-user downlink and data collection via the uplink, effectively addressing both communication and energy supply challenges for users in remote areas.
2. Jointly optimizing UAV trajectory and user associations enhances overall system effectiveness compared to approaches that optimizes only a single factor or enables only single user association.
3. Quantifying the UAV's coverage of users and incorporating it into the objective function helps balance the service time allocated to each user, thereby enhancing the overall quality of service (QoS) for all users.

1.5 Research Methodology

Given the orbit inclination, velocity and altitude of LEO satellites in a Walker-Delta constellation are fixed, we model the satellite constellation using geometric mathematics to express the real-time coordinates of each satellite during its periodic motion. These dynamic coordinates of are then used to

calculate the distances between the satellites and the UAV, which determine the UAV's handover decisions and the real-time data rates for communication.

After establishing a comprehensive system model which includes the modeling of satellite movement, UAV mobility and user distribution, we formulate an optimization problem, incorporating UAV trajectory design and UAV-user association scheduling as decision variables. Due to the mix-integer variables and the highly non-linear, time-varying environment, this problem has become a Mix-Integer Non-Linear Programming (MINLP) problem. Solving such a problem using conventional optimization methods is extremely challenging. However, reinforcement learning provides a promising approach for efficiently searching for near-optimal solutions in complex decision spaces. Therefore, we employ a DRL approach to solve this optimization problem.

1.6 Delimitations

A key limitation not considered in this thesis is the scalability of the proposed algorithm. In our approach, the UAV determines the association relationship with all users at each time slot. As the number of users increases, the dimensions of the action space grows accordingly, leading to a dramatically increase in the computational complexity of the DRL-based algorithm. The DRL algorithm we choose to use is Proximal Policy Optimization (PPO), which does not effectively address the scalability issue. Future research could explore more scalable DRL algorithms, such as Independent PPO (IPPO) or hierarchical reinforcement learning, to better handle large-scale scenario.

1.7 Structure of the thesis

The remainder of this thesis structures as follows: Chapter 2 introduces relevant background information about potential optimization problem in Space-air-ground Integrating (SAGIN) architecture and solutions. Chapter 3 presents the methodology used to establish system model. Chapter 4 presents the principle of RL algorithms and the detailed implementation of our proposed algorithm. Chapter 5 shows the simulation results and discussion. Finally, the thesis concludes in chapter 6.

Chapter 2

Background

This chapter introduces the fundamental background of the three-layer architecture of the Space–Air–Ground Integrated Network (SAGIN) and the roles of each segment. It also reviews related work on optimization problems in SAGIN, along with current solution approaches.

2.1 SAGIN System

With the growing number of IoRT devices deployed in remote areas for tasks such as industrial automation and real-time environmental monitoring, the challenges of communication and task computing has become significantly urgent to solve. These challenges mainly arise due to the lack of access to base station and the limited computational resources of the devices. To address these issues, SAGIN has emerged as a promising solution.

SAGIN [13] is an architecture framework that integrates satellite constellation, aerial networks, and terrestrial base stations into a unified heterogeneous system. This integration enables SAGIN to provide seamless global coverage and cross-domain interconnection for users in diverse environments [14], making it well-suited for a wide range of practical applications, including earth observation, geographic mapping and disaster rescue.

In this section, we will introduce the architecture of this multidimensional heterogeneous network and explore the composition and functions of each segment.

2.1.1 Space Segment

Space segment includes three categories of satellites: LEO satellites, Medium Earth Orbit (MEO) satellites and Geostationary (GEO) satellites. Their different orbit heights and stationary nature result in various characteristics of communication applications.

Specifically, GEO satellites have the highest orbits (approximately 36786 km) and remain stationary relative to the earth. Therefore, although communication with GEO satellites face significant latency, they can maintain stable connections with ground users. Additionally, since GEO satellites have fixed coverage area, there is no need to consider handover issues. All these characteristics of GEO satellites make them particularly suitable for broadcasting and uninterrupted coverage communication [15].

Satellites with lower orbit, MEO and LEO satellites, are classified as non-geostationary (NGSO) satellites [2]. MEO satellites have orbit height ranging from 7,000 km to 25,000 km, providing lower latency but smaller coverage areas. They are often used for navigation systems such as GPS. LEO satellites, with orbits between 300 km and 1,500 km, offer the highest throughput and lowest latency compared to other types of satellites due to lower pathloss and shorter propagation distances. Additionally, advancements in relevant technologies such as phased array multi-beam antenna, onboard real-time processing and inter-satellite links (ISL) have made LEO satellite communications more cost-effective and mature. Thanks to these advancements, LEO satellites are expected to play a particularly important role in NTN for 6G technology [16]. However, due to the limited coverage area of individual LEO satellites, the constellations need to be extremely dense, posing significant challenges for proper orchestration and management [17].

Table 2.1: Comparison of three types of satellites

Type	Orbit Height (km)	Applications
GEO	36786	Broadcasting
MEO	7000 - 25000	Navigation systems
LEO	300 - 1500	NTNs

2.1.2 Air segment

The air segment mainly consists of High Altitude Platforms (HAPs), UAVs and other types of aviation equipment. This segment offers several advantages.

Compared to satellites, these equipment are easier to deploy and have shorter transmission delays. In contrast to ground base stations, they are more flexible and provide larger coverage areas. Among these aerial equipments, UAV is the most widely used type in the air segment [18].

UAVs can play two key roles in SAGIN. Firstly, they can act as relays for space-ground links, enabling Internet access and communication service in both urban and remote areas [19]. Secondly, UAVs equipped with sufficient energy storage and computational resources can serve as edge computing nodes, providing network service directly to ground users [20].

Notably, UAVs can either follow predetermined trajectories [21] or dynamically adjust their flight directions and velocities to optimize performance [22]. For instance, a UAV hovering closer to users can achieve higher data rate with reduced transmit power. Another critical factor to affect the performance is the UAV-user association. In existing studies, the association between UAVs and single or multiple users is either determined by predefined rules, such as connecting UAVs only to users within a specific distance [23], or dynamically adjusted based on the current state of environment [24]. However, there remains a research gap, as few studies have addressed the joint optimization of UAV trajectories and multi-user associations, primarily due to the complexity arising from the large number of optimization variables. Nevertheless, such joint optimizations could potentially lead to significant performance improvement in SAGIN.

Moreover, UAVs has also been employed to provide wireless power transfer (WPT) for ground users. Xu *et al* proposed a UAV-enabled WPT system in which the UAV trajectory is optimized to maximize the minimum received energy among all users [25]. In addition, UAVs are capable of simultaneously delivering energy to ground devices in the downlink and collecting data in the uplink [26]. UAV-enabled WPT system has shown great potential for integration into SAGIN architecture, offering significant benefit to devices in remote areas that lack a stable energy supply.

2.1.3 Ground Segment

Ground segment serve as the fundamental layer in the SAGIN architecture, supporting and coordinating with other segments. This segment includes ground base stations, user terminals, data center and other infrastructure that enable ground services [14]. The edge server configured at ground base station, with sufficient energy supply and computational resource, can provide high-bandwidth and low-latency service to nearby users [27].

In addition, data centers offer data storage, management and distribution service for applications that involve large volumes of data. By managing data traffic efficiently and ensuring seamless connectivity, these components enable the ground segment to not only serve terrestrial users effectively but also to collaborate smoothly with the space and air segments of SAGIN. This integration ensures end-to-end communication, optimized resource utilization, and reliable service delivery across all segments.

2.2 Related works

Based on the three-layer heterogeneous structure of the SAGIN system, the collaboration between different segments enables SAGIN to provide seamless coverage across the globe [28]. As a result, SAGIN is widely adopted in various scenarios of mobile edge computing, disaster rescue, task offloading, etc. Cheng *et al* use SAGIN architecture for IoT devices to offload tasks to address their resource constraint. In the system, UAVs are served as edge servers while satellites are relays [13]. Tang *et al* proposed a reinforcement learning-based traffic offloading system that takes into account the high mobility of nodes in SAGIN and their frequent changes in throughput and link state [29]. The study in [30] focuses on remote areas without the access to terrestrial network, and addresses the task scheduling problem for IoRT devices, aiming to reduce energy consumption of UAVs while meeting delay constraints. Similarly, [31] considers a marine scenario and reduces UAV energy consumption by jointly optimizing communication and computation resource allocation. This study also models three connection states between UAVs and LEO satellites based on their real-time relative positions.

Recent initiatives such as the 6G for Connected Sky (6G-SKY) project [32] further highlight the importance of SAGIN in the 6G era. This project aims to merge terrestrial and non-terrestrial networks to ensure continuous connectivity for both aerial and ground users, and identifies Advanced Air Mobility (AAM) as a representative use case requiring ultra-reliable low-latency communication (URLLC) for the safe operation of aerial vehicles (AVs) [33–35]. To meet these demands, optimization approaches for multi-connectivity path selection in combined airspace and non-terrestrial networks (ASN) have been proposed, while cellular-connected AVs remain an active research focus in areas such as handover and trajectory management [36–44]. Satellite communication technologies have also been explored, with studies proposing handover optimization methods for AVs [45]. Beyond conventional direct air-to-ground (DA2G) communication [46], air-to-air (A2A) links are

being considered to extend coverage in challenging environments such as over oceans [47, 48].

Although current existing studies have considered various complex scenarios, many of them overlook the highly dynamic nature and dense deployment of LEO satellites in a constellation. For example, [13], [29], [30] do not model the behavior of LEO satellites and treat them as black box instead, while [31] models only a single satellite in one orbit, which is not realistic. Nguyen *et al* formulated a computation offloading problem to minimize the energy consumption of both ground users and UAVs, where the coordinates of each LEO satellite in the constellation are incorporated into the model [22]. However, their model assumes static satellite coordinates. Given that LEO satellites can travel at velocities up to 7.9 km/s, real-time modeling of their movement is still necessary for accurate system analysis and optimization.

There are already many studies focusing on UAV in SAGIN system. Among them, two key factors, UAV trajectory and user association, have attracted the most attention. For instance, [49] proposed a UAV-enabled data collection system that employs DRL algorithm to optimize UAV trajectory with the objective of maximizing the collected data. While this work considers only a single UAV, recent studies have extended trajectory design problems to multiple-UAV scenarios, which require more coordination and collision avoidance mechanism, for example, Wang *et al* developed a collision-free trajectory designing algorithm for multiple UAVs to reduce task completion time [50]. [51] proposed a 3-D multi-UAV trajectory optimization algorithm for MEC system.

In addition, UAV communication for Beyond Visual Line of Sight (BVLOS) operations has been actively studied. Ozger et al. [52] introduced a framework combining mobile edge computing and augmented reality, highlighting the critical role of URLLC in enabling safe BVLOS operations. Wang et al. [53] examined the use of millimeter-wave (mmWave) cellular networks for UAV piloting, pointing out the potential of mmWave to provide high throughput but also its sensitivity to blockage. These studies collectively show that reliable, low-latency communication is indispensable for UAV operations in SAGIN. However, they also demonstrate that single-connectivity solutions are insufficient to meet the strict URLLC requirements. To address this, recent works have proposed multi-connectivity architectures that integrate DA2G, A2A, high-altitude platforms (HAPs), and LEO satellites [54, 55]. Evaluations under the finite block length regime confirm that such architectures significantly improve end-to-end reliability and latency. In addition, Wang et al. [56] proposed a clustering-based soft handover algorithm

to reduce handover frequency. While effective in lowering signaling overhead, it still cannot guarantee seamless transitions for multi-connectivity users, highlighting ongoing challenges in handover management for highly dynamic aerial networks.

However, few studies have investigated the joint optimization of trajectory designing and association scheduling. For example, [57] considers only one associated node at one time when planning UAV flying trajectory. [58] presents a joint optimization algorithm for both user association and UAVs trajectory design. However, this study does not take into account fairness among users. There remains a notable research gap for this joint optimization problem.

UAVs can also serve as energy suppliers, which is particularly beneficial in remote areas without grid electricity [59,60]. Several studies have examined optimization problem in UAV-enabled WPT systems. Due to the strong distance sensitivity of WPT pathloss, Xu *et al* formulated a UAV trajectory optimization problem to maximize the total received energy of ground users (GU) [25]. In [26], the UAV performs both energy supply and data collection, providing significant benefits for devices in remote areas by reducing the need for frequent battery replacement and lowering dependence on conventional power infrastructures [61]. However, most existing works focus primarily on UAV-GU architectures. Considering the advantages that SAGIN brings to remote areas, combining UAV-enabled WPT with other systems such as MEC in SAGIN could simultaneously address two major challenges in rural regions: stable energy supply and lack of terrestrial network (TN) support.

Nevertheless, due to the highly dynamic and heterogeneous nature of SAGIN, the associated research problems are often time-varying and highly complex. Solving such problems using conventional mathematical methods presents significant challenges. In [22] and [31], successive convex approximation strategies are applied to convexify and solve the non-convex problems. The work in [62] adopts a Lyapunov-based method to decompose the stochastic optimization into separate subproblems. Similarly, [63] tackles the non-convex and non-concave optimization problem by decomposing the original problem into subproblems and solve them iteratively. Hao *et al* [64] proposed a Lagrange dual decomposition-based algorithm to obtain closed-form optimal solutions.

Reinforcement Learning has shown strong potential in addressing the aforementioned complex non-convex problems, particularly those that are difficult to solve with conventional mathematical approaches [65–70]. Owing to its model-free nature, RL model can make real-time decisions based solely on the current environment state, making it well suited for high dynamic

scenarios. For example, [71] employs a Deep Neural Network (DNN) to map current system state to real-time decisions, and subsequently add the newly attained decisions into replay memory. A random batch is sampled from the memory to train the DNN periodically. Different RL algorithms have been utilized to suit various environments. Ullah *et al* formulate the optimization problem as a Markov Decision Process (MDP) and use a Double Deep Q-Network (DDQN) to make adaptive decisions [72]. Compared to standard DQN, DDQN separates the Q-value estimator into two streams during training, which can greatly improve generalization. In [73], it is assumed that each user has independent environment states which cannot be fully observed by other users. The cooperation of users is formulated as a Partially Observable Markov Decision Process (POMDP) and is solved by a multi-agent deep reinforcement learning (MADRL) algorithm.

Building upon prior research, this thesis jointly optimizes the UAV trajectory and multi-user associations with the goal of maximizing data collection while maintaining user fairness. Unlike most existing studies that focus solely on data collection and assume single-user association, our approach incorporates fairness considerations and supports multiple simultaneous user connections. Furthermore, we integrate a dynamic LEO satellite constellation model into the system, enabling the UAV to make handover decisions based on predicted satellite positions without requiring continuous communicating with satellites, thereby reducing signaling overhead. Additionally, we propose a novel system design in which the UAV simultaneously performs data collection and WPT to the associated users, which is rarely addressed in current studies.

2.3 Summary

This chapter provides an overview of background knowledge of SAGIN system, including its three-layer architecture and the functionalities of each segment. The related work section reviews the existing study on various optimization problems in SAGIN and summarizes the methods employed to address them. Furthermore, it identifies the current research gaps and outlines the contributions of this thesis.

Chapter 3

System Design of SAGIN with dynamic LEO satellite constellation

This chapter presents the detailed system setup for the scenario considered in this thesis. Section 3.1 introduces the overall system architecture, including the specific configurations of each segment in SAGIN, with a particular focus on the real-time position representation of the dynamic Walker-Delta constellation in the space segment. Section 3.2 describes the system model in detail, covering the communication model, energy consumption model, and user fairness metrics. Finally, Section 3.3 provides the formal problem formulation.

3.1 System Architecture

As shown in Fig. 3.1, we consider a joint multi-user association and trajectory design problem of a UAV in SAGIN architecture. In this scenario, the LEO satellites are continuously moving within a constellation, providing seamless service to ground users (GU). However, due to the limited transmit power of GUs, an UAV is deployed as a relay between GUs and satellites. The UAV also provide wireless power transfer for GUs, enabling them to have enough energy to transmit data.

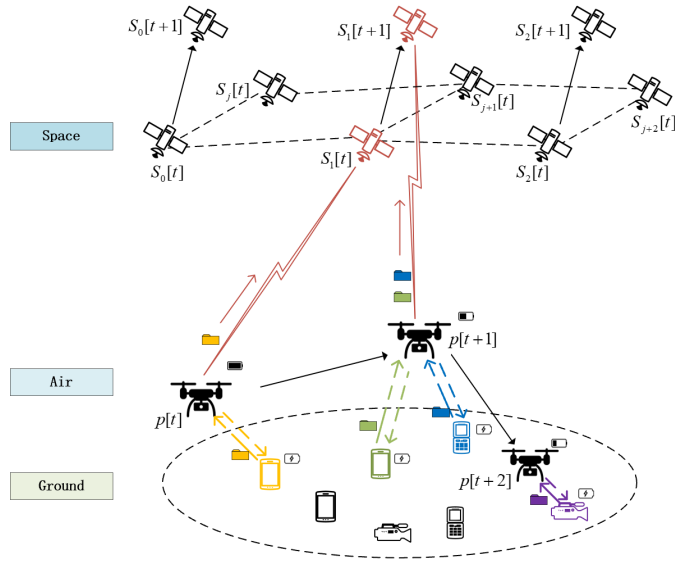


Figure 3.1: SAGIN system architecture

3.1.1 Ground User System

We consider a rectangular service area of fixed size, centered at a geographic location with latitude and longitude coordinates (σ^A, η^A) . A set of heterogeneous users $\mathcal{I} = \{1, 2, \dots, I\}$ are located within this area, and their positions are assumed to remain fixed during the UAV's operation period. To better describe the position of users and the UAV within the considered area, we establish a three-dimensional Cartesian coordinate system. The origin of this coordinate system is set at the lower-left corner of this rectangular area, with the x- and y-axes oriented along the directions of the horizontal and vertical edges adjacent to this corner, respectively. Consequently, the position of any user i within the region can be represented as $(x_i, y_i, 0)$.

At the beginning of the operation period, each user generates a certain amount of data i to be uploaded to the UAV. A user can only transmit data with constant power P^G when it is associated with the UAV.

3.1.2 UAV System

As shown in Fig. 3.2, we consider a single UAV operating over a designated remote area for a fixed duration, during which it simultaneously performs data collection from GUs via the uplink and WPT via the downlink.

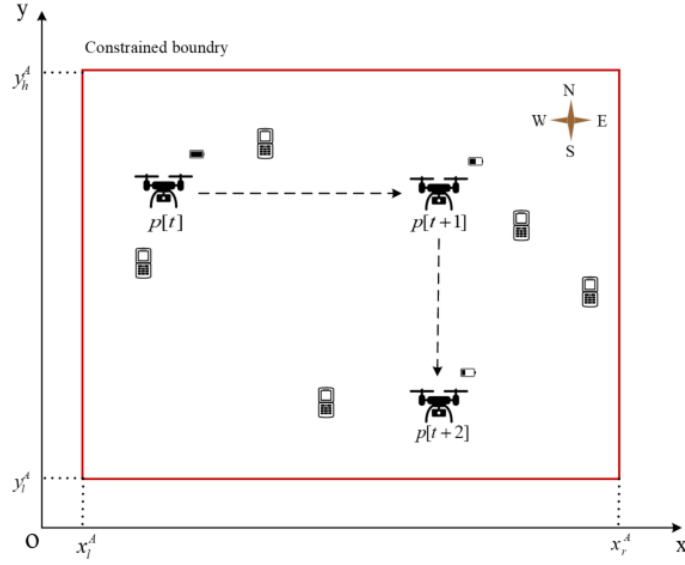


Figure 3.2: UAV flying model

To simplify the research problem, the continuous operation period is discretized into T time slots of equal duration τ . At each time slot, the UAV must decide its user association strategy $\{a_1[t], a_2[t], \dots, a_i[t]\}$ and its flying direction $\varphi^U[t]$. $a_i[t]$ is a binary variable, where $a_i[t] = 1$ indicates the UAV is associated with user i at time slot t , and $a_i[t] = 0$ otherwise. Once associated, the user is allowed to simultaneously transmit data and receive wireless energy from the UAV until the association is terminated. The UAV is capable of associating with multiple users within a single time slot, in which case the communication resources are shared among the associated users.

The UAV has four candidate directions: east, west, north and south, and moves at a constant velocity V^U . The UAV flies at a constant height H^U . Its real-time position $p[t]$ can be represented in the three-dimensional Cartesian coordinate system defined in Section 3.1.1 as $(x^U[t], y^U[t], H^U)$. Moreover, the UAV is constrained to operate within a bounded rectangular area. Specifically, its coordinates must satisfy:

$$\begin{cases} x_l^A \leq x^U[t] \leq x_r^A \\ y_b^A \leq y^U[t] \leq y_t^A \end{cases} \quad (3.1)$$

where x_l^A and x_r^A denote the left and right boundaries of the flight area, and y_b^A and y_t^A represent the bottom and top boundaries, respectively.

As each time slot is assumed to be sufficiently short, we can assume

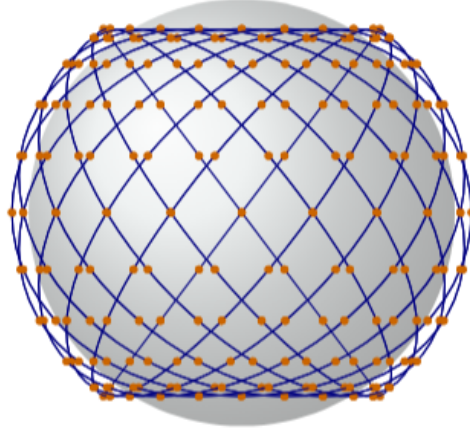


Figure 3.3: Walker-Delta constellation pattern for LEO satellites

UAV's location only updates at the sharp end of each time slot, thus avoiding the frequent changes in user-UAV channel states caused by its continuous movement. Besides, given the dense deployment of LEO satellites in the constellation, it is assumed that the service area is always covered by at least one satellite. The UAV establishes a connection with the nearest satellite based on real-time satellite positions and performs a handover decision every five time slots.

3.1.3 Dynamic Walker-Delta Constellation

We consider LEO satellites $\mathcal{J} = \{1, 2, \dots, J\}$ in a Walker-Delta constellation, where all satellite orbits have the same inclination α . Assume there are N_p orbit planes and M_p satellites in each orbit. Both the orbits and the satellites are considered evenly distributed. The position of each satellite in the constellation can be represented by the latitude σ and longitude η of its sub-satellite point (SSP), i.e., its projection onto the Earth's surface, along with its radial distance from the Earth's center, which can be represented as (σ, η, R) . Owing to the uniform distribution of orbital planes and satellites in the constellation, the longitude of the intersection between the orbital plane and the Earth's equatorial plane, which is Right ascension of Ascending Node (RAAN), denoted by L , of adjacent orbital planes satisfies $\Delta L = 2\pi/N_p$. Similarly, the orbital phase difference between adjacent satellites within the same orbital plane is given by $\Delta u = 2\pi/M_p$. As a result, the positions of all satellites in the constellation can be determined based on the location of any single reference satellite using these geometric relationships.

We then consider the dynamic position representation of an arbitrary satellite j , which can be denoted by $(\sigma_j[t], \eta_j[t], R)$. Similar to the simplification applied to UAV motion, satellite positions only update at the sharp end of each time slot. Assuming that all satellites move with the same angular velocity ω , the orbital phase angle of satellite j can be expressed as:

$$u_j[t] = \text{mod}(\omega t + u_{j0}, 2\pi/M_p), \quad (3.2)$$

where u_{j0} is its initial phase angle and mod is the normalization function.

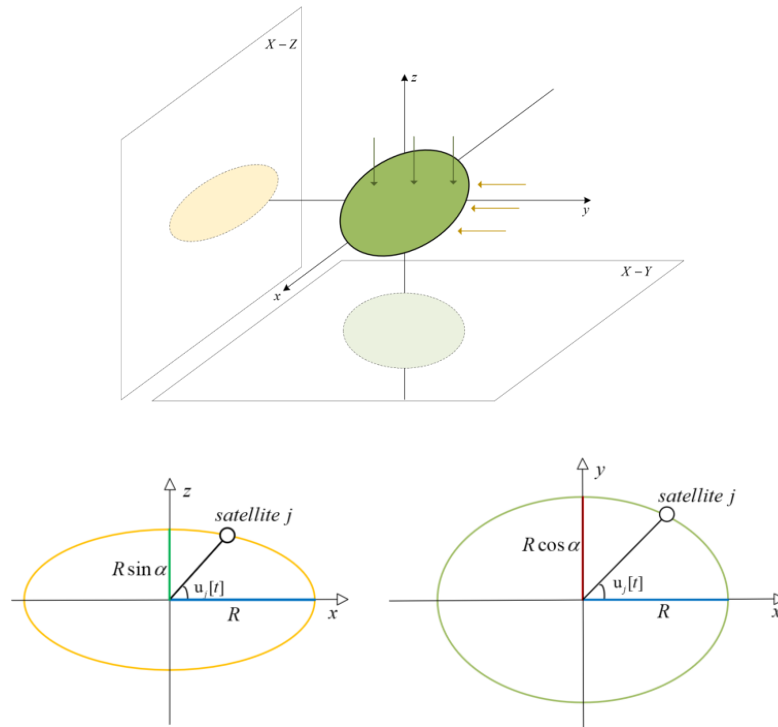


Figure 3.4: The projection of a satellite onto coordinate planes

To express the satellite's position in Cartesian coordinates, we establish a three-dimensional coordinate system centered at the Earth's center. In this system, we define the x -axis as the direction of the intersection between the orbital plane of satellite j and the equatorial plane, the x - y plane as the equatorial plane, and the z -axis as the direction pointing toward the North Pole, perpendicular to the equatorial plane. Under this coordinate system, the

position of satellite j can be expressed as:

$$(R \cos \sigma_j[t] \sin \eta_j[t], R \cos \sigma_j[t] \cos \eta_j[t], R \sin \sigma_j[t]). \quad (3.3)$$

Fig. 3.4 illustrates the projection of an arbitrary orbital plane onto the X-Y and X-Z planes. As shown, the projection onto the X-Y plane forms an ellipse with a semi-minor axis of length $R \cos \sigma$ and a semi-major axis of length R , while the projection onto the X-Z plane results in an ellipse with a semi-minor axis of length $R \sin \sigma$ and a semi-major axis of length R . Therefore, the Cartesian coordinate of satellite j can also be represented as

$$(R \cos(u_j[t]), R \cos \alpha \sin(u_j[t]), R \sin \alpha \sin(u_j[t])). \quad (3.4)$$

By combining the two coordinate representations, we derive equation 3.5, from which the expression for σ and η can be obtained.

$$\begin{cases} x = R \cos \sigma \sin \eta = R \cos(u_j[t]) \\ y = R \cos \sigma \cos \eta = R \sin \alpha \sin(u_j[t]) \\ z = R \sin \sigma = R \cos \alpha \sin(u_j[t]) \end{cases} \quad (3.5)$$

σ_j and η_j can be formulated as:

$$\sigma_j[t] = \arcsin(\sin \alpha \sin(u_j[t])) \quad (3.6)$$

$$\eta_j[t] = \begin{cases} \arctan(\cos \alpha \tan(u_j[t])) + \eta_L, & u_j[t] \in [-\frac{\pi}{2}, \frac{\pi}{2}] \\ \arctan(\cos \alpha \tan(u_j[t])) + \eta_L + \pi, & \text{else} \end{cases} \quad (3.7)$$

where η_L denotes the longitude of the intersection between the orbit of satellite j and the equatorial plane.

3.2 System Model

This section presents the specific system model used in this thesis, including the communication model, the energy consumption model, and the formulation of a variable introduced to quantify user fairness.

3.2.1 Communication Model

Since the considered scenario is located in remote areas such as deserts and oceans, the environment typically contains few or no obstacles. Given the high

mobility and maneuverability of UAVs, the communication between the UAV and ground users is generally assumed to follow a line-of-sight (LoS) model [58]. Furthermore, orthogonal frequency division multiple access (OFDMA) is employed to eliminate inter-channel interference, allowing the uplink and downlink transmissions between the UAV and users to be decoupled and carried out simultaneously. All associated users share the bandwidth of UAV-GU link. Therefore, the channel qualities are basically determined by the distances between the UAV and users and user association. At time slot t , the distance between user i and the UAV can be calculated as:

$$d_i^G[t] = \sqrt{(x^U[t] - x_i)^2 + (y^U[t] - y_i)^2 + (H^U)^2} \quad (3.8)$$

Thus, the uplink channel gain is formulated as:

$$g_i^G[t] = G_T^G + G_R^U - (32.45 + 20 \times \log(d_i^G[t] \times f^G)), \quad (3.9)$$

where G_T^G is the transmitting gain of antennas in GU, and G_R^U denotes the receiving gain of antennas in UAV. f^G is the central carrier frequency. The term $32.45 + 20 \times \log(d_i^G[t] \times f^G)$ represents the free space path loss.

Hence, the transmission data rate can be expressed as:

$$r_i^G[t] = b^G[t] \times \log_2\left(1 + \frac{P^G \times g_i^G[t]}{N_0 \times b^G[t]}\right) \quad (3.10)$$

where $b^G[t]$ denotes the bandwidth allocated to each associated user at time slot t . For associated users, $b^G[t] = \frac{B^G}{\sum_i a_i[t]}$. P^G is the transmitted power for users, and N_0 is the power spectral density of the additive white Gaussian noise (AWGN).

We also assume LoS communication between UAV and LEO satellites. Given the high orbital altitude and large velocity of LEO satellites, the UAV's mobility is negligible. Therefore, when selecting the nearest satellite for connection, the distance is calculated between the satellite and the center of the service area rather than the UAV's real-time position.

The distance between satellite j and the UAV at time slot t can be calculated based on their coordinates in the three-dimensional spherical

coordinate system:

$$d_j^U[t] = \left[(R \cos \eta_j[t] \cos \sigma_j[t] - R_0 \cos \eta^A \cos \sigma^A)^2 + (R \cos \eta_j[t] \sin \sigma_j[t] - R_0 \cos \eta^A \sin \sigma^A)^2 + (R \sin \eta_j[t] - R_0 \sin \eta^A)^2 \right]^{1/2} \quad (3.11)$$

Consequently, the uplink channel gain can be presented as:

$$g_j^U[t] = G_T^U + G_R^S - (32.45 + 20 \times \log(d_j^U[t] \times f^U)) \quad (3.12)$$

where G_T^U is the transmitting antenna gain of UAV, and G_R^S is the receiving gain of satellites. f^U is the central carrier frequency.

Let $g^U[t]$ denote the uplink channel gain from the UAV to the connected satellite. The real-time data rate of the uplink channel can be calculated as:

$$r^U[t] = B^U \times \log_2 \left(1 + \frac{P^U \times g^U[t]}{N_0 \times B^U} \right) \quad (3.13)$$

where B^U is the bandwidth of the uplink channel.

If the UAV's uplink data rate is sufficiently high, it will transmit all the data collected from users in the current time slot to the connected satellite. However, if the uplink data rate is insufficient, the UAV will allocate the available transmission capacity among the users proportionally based on the amount of data collected from each user. Accordingly, the amount of data collected from each user that can be upload to satellite in real time is given by:

$$r_i^U[t] = \frac{r^U[t] \times r_i^G[t]}{\sum_k r_k^G[t]} \quad (3.14)$$

Consequently, the data rate for user i to upload data to LEO satellite j at time slot t is determined by the bottleneck of the two-hop link, and is thus given by the minimum of the data rates over the two links, which can be formulated as:

$$\tilde{r}_i[t] = \min\{r_i^G[t], r_i^U[t]\} \quad (3.15)$$

3.2.2 Energy Consumption and Harvest Model

The energy consumption of the UAV contains three parts: its propulsion, data transmission and for WPT. The propulsion power of a fixed-wing UAV can be

calculated based on a aerodynamics model [74]:

$$P^f = c_1 \|v^U\|^3 + \frac{c^2}{\|v^U\|} \quad (3.16)$$

The first term $c_1 \|v^U\|^3$ denotes the parasitic power, which is the power required to overcome parasitic drag causing by the airflow over the UAV's surface, the shape of the UAV, and etc. The second term $\frac{c^2}{\|v^U\|}$ is the induced power, which is the power needed to overcome induced drag causing by the UAV's wings redirecting wings to generate lift. c_1 and c_2 are the constant parameters related to the UAV's weight, wing area and etc.

The UAV has constant transmit power for communication and wireless power transfer. Consequently, the power consumption of UAV at time slot t can be formulated as follows:

$$p^U[t] = P^f + P^U + P^{tx} \times \sum_i a_i[t] \quad (3.17)$$

where P^U is the transmit power of UAV, and P^{pt} denotes the power for UAV to provide WPT to users.

For GUs, the received power of WPT decreases proportionally to the square of the transmission distance. The power of received energy at an arbitrary user i is given by [25]:

$$p_i^{rx}[t] = \frac{a_i[t] \times \beta_0 P^{tx}}{(d_i^G[t])^2} \quad (3.18)$$

where β_0 is the channel power gain at a reference distance of 1 meter.

3.2.3 User fairness

To ensure fairness among users in terms of data collection service received, we define a metric to quantify this fairness: the number of underserved users. Specifically, a user i is considered underserved if the amount of data collected by the UAV from user i over the total operation period of T time slots is less than a specified proportion of its total data to be transmitted. The indication of underserved user is given by:

$$n_i = \begin{cases} 1, & \text{if } \sum_t (\tilde{r}_i^G[t] \times a_i[t]) \leq \mu \times b_i \\ 0, & \text{otherwise} \end{cases} \quad (3.19)$$

where μ is the required proportion of collected data and b_i denotes the data size to be uploaded for user i .

The user fairness of data collection can thus be presented as $\sum_i n_i$.

3.3 Problem Formulation

In this thesis, we consider an optimization problem to jointly maximize the data collection of UAV and ensure user fairness. This research problem involves two variables: flying direction of the UAV at each time slot and the association between GUs and UAV. Based on the system model introduced before, the research problem can be formulated as:

$$\mathcal{P} : \max_{\mathbf{a}, \boldsymbol{\varphi}} w_r \left(\sum_{t=0}^T \sum_{i=0}^I \tilde{r}_i[t] \right) - w_n \sum_{i=0}^I n_i \quad (3.20)$$

$$s.t. \quad a_i[t] = \{0, 1\}, \quad \forall i \quad (3.20a)$$

$$\varphi[t] = \{0, 1, 2, 3\} \quad (3.20b)$$

$$x_l^A \leq x^U[t] \leq x_r^A \quad (3.20c)$$

$$y_l^A \leq y^U[t] \leq y_r^A \quad (3.20d)$$

$$\sum_{t=0}^T p^U[t] \leq E^U \quad (3.20e)$$

$$\sum_{t=0}^T a_i[t] P^G \leq \sum_{t=0}^T a_i[t] p_i^{rx}[t], \quad \forall i \quad (3.20f)$$

where $\boldsymbol{\varphi} = \{\varphi[t]\}$, $\mathbf{a} = \{a_i[t] | i \in \mathcal{I}\}$. w_r and w_n denote the weight factors. E^U is the maximum energy storage of the UAV. In this optimization problem, (3.20a) implies that multiple GUs can simultaneously associate with the UAV. (3.20b) describes the UAV need to choose from four candidate directions at each time slot. (3.20c) and (3.20d) restrict the flying boundaries of the UAV. (3.20e) imposes limitation on the total energy storage for the UAV. (3.20f) requires the power each user received should larger than the power consumed.

Chapter 4

PPO-based UAV trajectory designing and user association scheduling algorithm

This chapter presents the principles of the reinforcement learning algorithm Proximal Policy Optimization (PPO) and details the implementation of our proposed PPO-based algorithm for optimizing UAV trajectory designing and user association scheduling in the scenario considered in this thesis. Finally, two benchmark algorithms are introduced to validate the performance of the proposed method.

4.1 Proximal Policy Optimization Algorithm

Reinforcement learning (RL) is a type of unsupervised learning that does not require pretrained models or labeled datasets, enabling flexible deployment in various dynamic and complex environments. An RL agent learns to adjust its decisions by interacting with the environment and receiving feedback, aiming to maximize the expected long-term rewards.

In RL algorithm, the interaction between the agent and the environment can be modeled as a Markov Decision Process (MDP), presented by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$. Specifically, \mathcal{S} denotes the state space of the system, which includes all possible states of the environment. \mathcal{A} represents the set of all possible actions. \mathcal{P} is the state transition probability function, and \mathcal{R} denotes the reward function. It's worth noting that the transition probability and achieved reward for moving from the current state of the environment s_t to the next state s_{t+1} depend only on the current state and the selected action

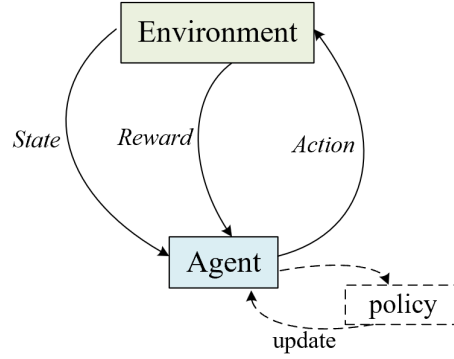


Figure 4.1: The framework of general RL algorithms

a_t rather than history states and actions.

Based on the MDP model, the long-term reward the RL agent can achieved can be defined as the total return accumulated over a specific time horizon:

$$G(\rho) = \sum_{t=0}^{T-1} \gamma^t R_{t+1} = \sum_{t=0}^{T-1} \gamma^t R(s_t, a_t, s_{t+1}) \quad (4.1)$$

where R_{t+1} denotes the reward received at time step $t + 1$, corresponding to the transition of the environment from state s_t to s_{t+1} after the RL agent takes action a_t . The parameter $\gamma \in (0, 1)$ is a discount factor that assigns greater importance to immediate rewards compared to future rewards. $\rho = \{s_0, a_0, r_0, \dots, s_T, a_T, r_T\}$ is the trajectory, which presents the sequence of states, actions and corresponding rewards experienced by the agent while interacting with the environment.

There can be various possible trajectories depending on the policy under which the RL agent operates. The objective of the algorithm is to maximize the expected total reward across all possible trajectories by optimizing the probability distribution over trajectories. Accordingly, the objective function can be formulated as:

$$J(\theta) = \mathbb{E}_{\rho \sim p_{\theta}(\rho)}[G(\rho)] = \int p_{\theta}(\rho) G(\rho) d\rho \quad (4.2)$$

where $p_{\theta}(\rho)$ denotes the probability density of the trajectory ρ , which is determined by the policy selected by the RL agent. It can be expressed as:

$$p_{\theta}(\rho) = p(s_0) \prod_{t=0}^{T-1} \pi_{\theta}(a_t | s_t) P(s_{t+1} | s_t, a_t) \quad (4.3)$$

Here, the policy π_θ defines the action-selection probabilities at each state, and θ represents the set of all parameters that characterize the policy.

In order to find the optimal policy to make appropriate decisions at each state, RL algorithms introduce two key concepts: state value function $V_{\pi_\theta}(s_t)$ and state-action value function $Q_{\pi_\theta}(s_t, a_t)$. The state value function represents the expected return when starting from state s_t at time step t . Similarly, state-action value function represents the expected return when starting from state s_t at time step t and taking action a_t . They can be formulated as follows:

$$V_{\pi_\theta}(s_t) = \mathbb{E}_{\pi_\theta} \left[\sum_{k=0}^T \gamma^k R_{k+t+1} | s = s_t \right] \quad (4.4)$$

$$Q_{\pi_\theta}(s_t, a_t) = \mathbb{E}_{\pi_\theta} \left[\sum_{k=0}^T \gamma^k R_{k+t+1} | s = s_t, a = a_t \right] \quad (4.5)$$

There are two main approaches to enable RL algorithms to maximize the objective function. The first is value-based methods, such as Q-learning algorithm, where actions are selected based on maximizing the value function $V_{\pi_\theta}(s_t)$ of current state. The second is policy-based methods, to which the PPO algorithm belongs. PPO updates the policy parameters θ through gradient ascent, ultimately converging to an optimal policy.

Specifically, the most basic gradient ascend method for policy updates can be described as:

$$\theta_{new} \leftarrow \theta_{old} + \beta \cdot \nabla J(\theta_{old}) \quad (4.6)$$

$$\nabla J(\theta_{old}) = \sum_{t=0}^{T-1} G(\rho) \nabla \ln \pi_{\theta_{old}}(a_t | s_t) \quad (4.7)$$

However, different parameters θ can lead to significant variations in the expected return G obtained by the RL agent, thereby causing large fluctuations in the magnitude of policy updates. To reduce the variance of G , a baseline independent of the action can be introduced. By subtracting this baseline, the policy gradient can be computed with reduced variance. The advantage function is thus defined to represent the immediate reward after subtracting the baseline, which can be formulated as:

$$A_{\pi_\theta}(s_t, a_t) = R_t - (V_{\pi_\theta}(s_t) - V_{\pi_\theta}(s_{t+1})) = Q_{\pi_\theta}(s_t, a_t) - V_{\pi_\theta}(s_t) \quad (4.8)$$

The first term R_t represents the actual immediate reward at time step t , and the

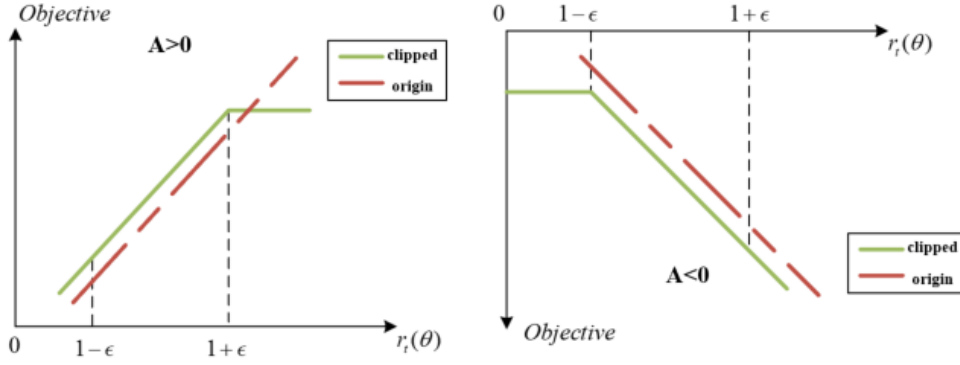


Figure 4.2: Illustration of the Clipped Surrogate Objective in PPO

second term $V_{\pi_{\theta}}(s_t) - V_{\pi_{\theta}}(s_{t+1})$ is the expected reward that can be obtained using policy θ .

Therefore, unlike the most basic RL algorithm equation (4.7), the objective function can be reformulated by using advantage function as:

$$\operatorname{argmax}_{\theta} \mathbb{E}_t \left[\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} A_{\pi_{\theta_{old}}}(s_t, a_t) \right] \quad (4.9)$$

$$\text{s.t.} \quad \mathbb{E}_t [\text{KL}[\pi_{\theta_{old}}(\cdot|s_t), \pi_{\theta}(\cdot|s_t)]] \leq \delta \quad (4.10)$$

where the factor $\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ denotes the importance sampling and KL represents the KL divergence. δ is a fixed threshold parameter. The constraint (4.10) limits the magnitude of policy updates.

Based on the objective function (4.9), PPO introduces a clipped surrogate objective [75], using a penalty rather than a hard constraint. Let $r_t(\theta)$ denotes the probability ratio $\frac{\pi_{\theta}(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$. The objective function is given as:

$$\operatorname{argmax}_{\theta} \mathbb{E}_t [\min(r_t(\theta) A_{\pi_{\theta_{old}}}(s_t, a_t), \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) A_{\pi_{\theta_{old}}}(s_t, a_t))] \quad (4.11)$$

where ϵ is a hyperparameter. The probability ratio $r_t(\theta)$ is restricted between $1 - \epsilon$ and $1 + \epsilon$ by the clipping function to ensure the similarity between the updated and old policy. The final objective is the minimum value of the clipped and unclipped objective as shown in Fig. 4.2.

4.2 Algorithm Implementation in SAGIN System

In our considered scenario in SAGIN system, one single PPO agent is deployed at the UAV, observing the environment states and making appropriate decisions. The optimization problem 3.20 can be converted to a MDP, which is described by tuple $M = (\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$, where \mathcal{S} , \mathcal{A} and \mathcal{R} are the set of environment states, possible actions, and reward function, respectively. \mathcal{P} is the state transition function $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$. $\gamma \in (0, 1)$ denotes the discount factor. The detailed definition of MDP for our research problem are given below.

There exist various environment variables that can influence the agent's decisions at each time step. Let s_t denote the state of environment at time slot t . Specifically, s_t includes the UAV's current location coordinates $(x^U[t], y^U[t], H^U)$, the three-dimensional Euclidean distances between the UAV and each user $d_i^G[t]$, the amount of data each user has remaining to be collected by the UAV $b_i[t]$, the real-time energy level of each user $E_i^G[t]$, and the real-time data rate of the link between the UAV and the connected LEO satellite $r^U[t]$. Therefore, s_t can be expressed as:

$$s_t = \{x^U[t], y^U[t], d_1^G[t], \dots, d_I^G[t], b_0[t], \dots, b_I[t], E_1^G[t], \dots, E_I^G[t], r^U[t]\} \quad (4.12)$$

Therefore, the state space can be represented as $\mathcal{S} = \{s_t | t = 1, 2, \dots, T\}$.

Based on the state of environment, the PPO agent deployed at UAV needs to determine several decisions including the UAV's real-time flying direction $\varphi^U[t]$ and the association relationship with each user $a_i[t]$. The action a_t of the agent at time slot t can thus be defined as:

$$a_t = \{\varphi^U[t], a_1[t], \dots, a_I[t]\} \quad (4.13)$$

with binary variable $a_i[t]$ and $\varphi^U[t] \in \{0, 1, 2, 3\}$. Accordingly, the action space $\mathcal{A} = \{a_t | t = 1, 2, \dots, T\}$.

After executing the actions, current state s_t will transfer to the next state s_{t+1} with probability $P(s_{t+1} | s_t, a_t)$.

The objective of our research problem is to jointly maximize the amount of data collected by the UAV and ensure the user fairness, while satisfying all system constraints. Consequently, we define the reward function at time slot t as the sum of data collection quantity and a user fairness metric. To account for potential constraint violatio, a penalty term is also incorporated into the

formulation. The overall reward function is expressed as:

$$r_t = w_1 \sum_i \tilde{r}_i[t] + w_2 \sum_i \left(\frac{\sum_{k=0}^t (\tilde{r}_i[k]) - \mu \times b_i}{T - t} \right) - w_3 \chi_1[t] - w_4 \chi_2[t] - \sum_i w_5 \chi_{3,i}[t] \quad (4.14)$$

where the first term is the amount of data collected by the UAV and successfully uploaded to the satellite. The second term is the reward associated with user fairness. If the amount of data collected from a user has not yet reached a specified proportion of the total data, this term takes a negative value and acts as a time-increasing penalty until the target proportion is achieved. Conversely, if the collected data from a user reaches the required proportion, this term becomes positive and serves as a time-increasing reward. The last three terms represent penalty functions corresponding to UAV boundary violation (3.20c) & (3.20d), UAV energy constraint violation (3.20e), and insufficient energy at GUs (3.20f), respectively. The weights w_i , $i \in \{1, 2, 3, 4, 5\}$ are positive constants to balance the importance of each term in the reward function. The variables $\chi_1[t]$, $\chi_2[t]$ and $\chi_{3,i}[t]$ are binary indicators, where a value of 1 denotes a constraint violation at time slot t and 0 otherwise. Consequently, the reward function $\mathcal{R} = \{r_t | t = 1, 2, \dots, T\}$.

Due to the high dimensions of the state space \mathcal{S} , we employ a recurrent neural network (RNN) to replace traditional Q-value tables. Given the input environmental state, the RNN is capable of outputting the corresponding value function. The integration of RNN into the PPO framework allows the agent to handle larger state spaces that are difficult to represent in tabular form. RNN further enable the agent to capture temporal dependencies and hidden patterns in the environment by maintaining an internal memory of past observations, which can enhance the performance of the PPO agent in our considered complex and highly dynamic environment. Given the input environmental state, the RNN is capable of outputting the corresponding value function.

We employ an actor-critic method to update the PPO agent. Actor-critic method is a combination of both policy-based and value-based methods. The actor network directly learns a parameterized policy that takes current states as inputs and outputs action, while the critic estimates the value function to provide more informative gradients. This structure enables more stable and sample-efficient learning compared to pure policy gradient methods.

In RNN-PPO, both the policy network (actor) and the value network (critic) are implemented using recurrent architectures. As shown in Fig. 4.3, at

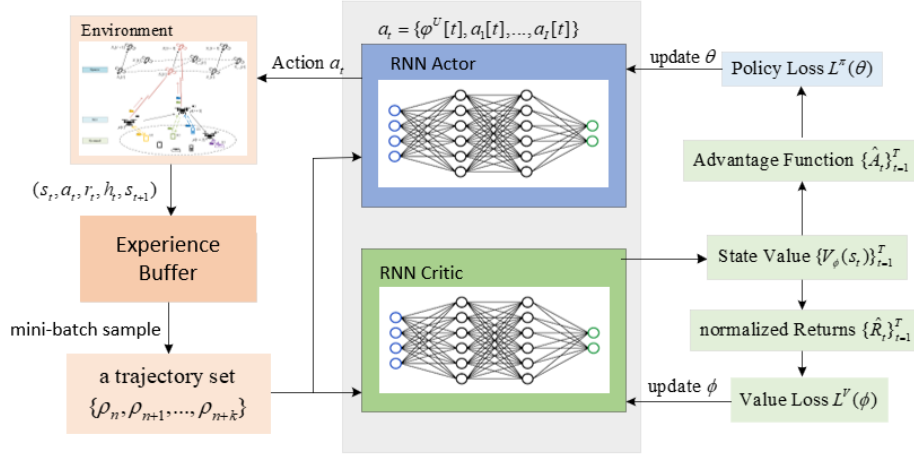


Figure 4.3: The framework of the proposed algorithm

each time step t , the RNN receives the current environment state s_t along with the previous hidden state h_{t-1} , and outputs both an action distribution (policy) $\pi_\theta(a_t | s_t, h_{t-1})$ and the updated hidden state h_t . The action a_t is then sampled by the policy. The critic network estimates the state value $V_{\pi_\theta}(s_t, h_{t-1})$ using a similar recurrent mechanism.

During each episode, the agent interacts with the environment and stores the transitions $(s_t, a_t, r_t, h_t, s_{t+1})$ into the experience buffer. After a specific number of episodes, a mini-batch of trajectories is sampled from the buffer to form a trajectory set $\{\rho_n, \rho_{n+1}, \dots, \rho_{n+k}\}$. This trajectory set is used to compute advantage estimates \hat{A}_t , normalized return \hat{R}_t , and RNN-updated hidden states, which are then used to update the actor and critic networks.

Specifically, as introduced in section 4.1 the actor is updated to maximize a clipped surrogate objective and the critic is updated to minimize the mean-squared error between predicted and empirical returns. The policy parameters θ are updated using the policy loss, which is formulated as equation (4.11). Meanwhile, the value function parameters ϕ are updated by minimizing:

$$L^V(\phi) = \mathbb{E}[(V_\phi(s_t, h_{t-1}) - \hat{R}_t)^2] \quad (4.15)$$

where $V_\phi(s_t, h_{t-1})$ is the estimated value function calculated by critic network. And \hat{R}_t denotes the estimated total return starting from time step t , which is calculated as $\hat{R}_t = \hat{A}_t + V_\phi(s_t, h_{t-1})$. And the advantage function \hat{A}_t is estimated using Generalized Advantage Estimation (GAE) method. Unlike traditional Temporal Difference (TD) learning used in algorithms such as DQN

and A2C, GAE can balance the variance and bias by taking a weighted sum of multi-step TD-errors, which can be formulated as:

$$\hat{A}_t = \sum_l (\gamma\lambda)^l \delta_{t+l} \quad (4.16)$$

where $\delta_t = r_t + \gamma V_\phi(s_{t+1}, h_t) - V_\phi(s_t, h_{t-1})$ is the TD-error, and $\lambda \in [0, 1]$ is a hyperparameter. When $\lambda = 1$, GAE method becomes equivalent to the Monte Carlo method, which has high variance but low bias; when $\lambda = 0$, it reduces to the standard one-step TD method, which has lower variance but higher bias. Thus, selecting an appropriate value of λ is essential to balance the variance and bias in advantage estimation.

The detailed implementation of the RNN-PPO algorithm in the scenario considered in this thesis is presented as follows:

Algorithm 1 RNN-PPO Algorithm for joint UAV trajectory design and user association optimization in SAGIN system

```

1 Initialize policy parameter  $\theta$  and global value function parameter  $\phi$  Using
  orthogonal initialization
2 while  $episode \leq max\_episode$  do
3   Initialize actor RNN states
4   Initialize critic RNN states
5   Each GU generates a task with specific data size
6   for step  $t = 0$  to  $max\_step$  do
7     Update the position of all LEO satellites in the constellation
8     Retrieve current environment state  $s_t$  and RNN hidden state  $h_t$ 
9     Agent selects action  $a_t$  according to policy  $\pi_\theta$ 
10    UAV moves and collects data from associated GUs
11    Observe immediate reward  $r_t$  and next state  $s_{t+1}$ 
12    Store transition  $(s_t, a_t, r_t, h_t, s_{t+1})$  into replay buffer
13    if  $t \% update\_interval == 0$ 
14      Sample a mini-batch of trajectories from replay buffer
15      for mini-batch  $k = 1, \dots, K$  do
16        Compute advantage estimate  $\hat{A}_t$  via GAE method
17        Compute normalized returns  $\hat{R}_t = \hat{A}_t + V_\phi(s_t, h_{t-1})$ 
18        Compute policy Loss  $L^\pi(\theta)$  and value loss  $L^V(\phi)$ 
19        Update policy parameters  $\theta$  in actor network
20        Update value function parameters  $\phi$  in critic network
21      end for
22    end
23  end for
24  if  $episode \% eval\_interval == 0$ 
25    Initialize cumulative return
26    Run current policy  $\pi_\theta$  in each environment without exploration
27    Compute average returns
28  end
29 end while

```

4.2.1 Summary

In this chapter, we first introduced the fundamental principles of reinforcement learning, and then extended them to the PPO algorithm with a clipped objective. Subsequently, we describe the integration of RNNs into PPO and the use of an actor-critic architectures for policy and value updates. Finally, we present the detailed implementation of the RNN-PPO algorithm in the scenario

considered in this thesis, including the configuration of the agent's position, the definitions of the state and action spaces, and the design of the reward function.

Chapter 5

Results and Analysis

In this chapter, we first present the detailed environment settings and parameter configurations used in the simulation. The results are organized into two parts. The first part focuses on the construction of a multi-orbit dynamic LEO satellite constellation, where we demonstrate the constellation generated using the aforementioned method to validate its feasibility and effectiveness. The second part evaluates the performance of the proposed PPO-based algorithm in the designed scenarios. We compare our method with two benchmark algorithms, and observe its performance under various environments to verify its superiority and adaptability.

5.1 Dynamic Multi-orbit LEO Satellite Constellation

Based on the real-time coordinate modeling of LEO satellites described in Section 3.1.3, a dynamic multi-orbit Walker-Delta constellation is constructed. To better approximate practical satellite deployments, all simulation parameters in this part are configured with reference to the Starlink constellation. The table below shows the parameter used to construct the satellite constellation.

Table 5.1: Parameters for Satellite Constellation Construction

Parameter	Value
Orbit number: N_p	20
Satellite number per orbit: M_p	63
Orbit angle: α	1.152 rad
Orbit height: H^S	500 km
Angular velocity of LEO satellite: ω	0.063 deg/s

The satellite constellation is simulated using Matlab. After computing the coordinates of all satellites, their real-time positions are visualized, as illustrated in Fig. 5.1.

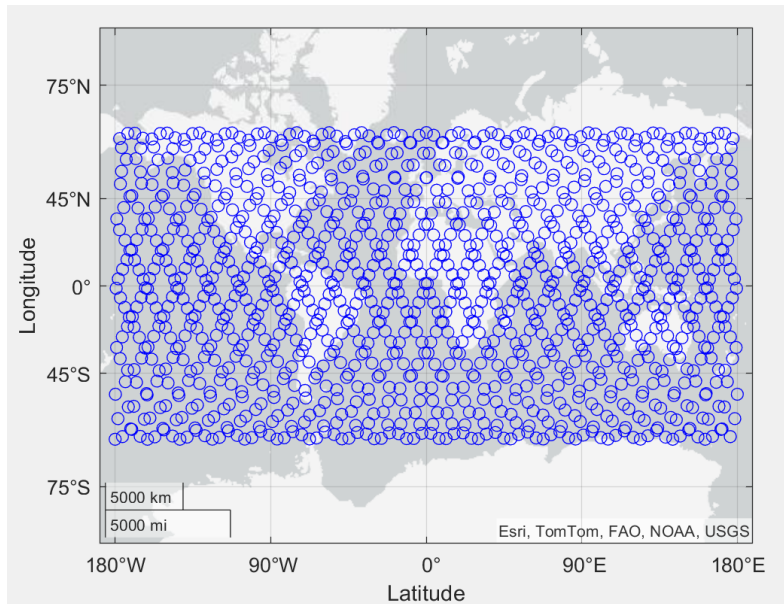


Figure 5.1: Multi-orbit Walker-Delta LEO satellite constellation

Based on the previously computed real-time positions of satellites in the constellation, the uplink data rate between the UAV and the connected satellite can be calculated. The considered remote ground area is located in Lapland, northern Sweden, with a fixed central coordinate. The communication link between the UAV and LEO satellites is assumed to operate in the Ka frequency band. The UAV is connected to the nearest satellite and performs a handover every three time slots. The relevant simulation parameters are summarized in the table below.

Table 5.2: Environment parameters

Parameter	Value
Time slot length τ	1 s
Total number of time slots T	500
UAV-satellite Uplink Carrier Frequency f^U	30 GHz
UAV flying height: H^U	10 m
UAV transmit power: P^U	8 W
GU transmit power: P^G	1 mW
Noise power: N_0	-154 dBm/Hz
UAV transmitting antenna gain: G_T^U	20
UAV receiving antenna gain: G_R^U	8
Satellite Receiving antenna gain: G_R^S	35
GU transmitting antenna gain: G_T^G	5
Total bandwidth between GUs and UAV: B^G	10 MHz
Total bandwidth between UAV and satellite: B^U	30 MHz
Ground area center coordinate: (σ^A, η^A)	(67.75°N, 18°E)

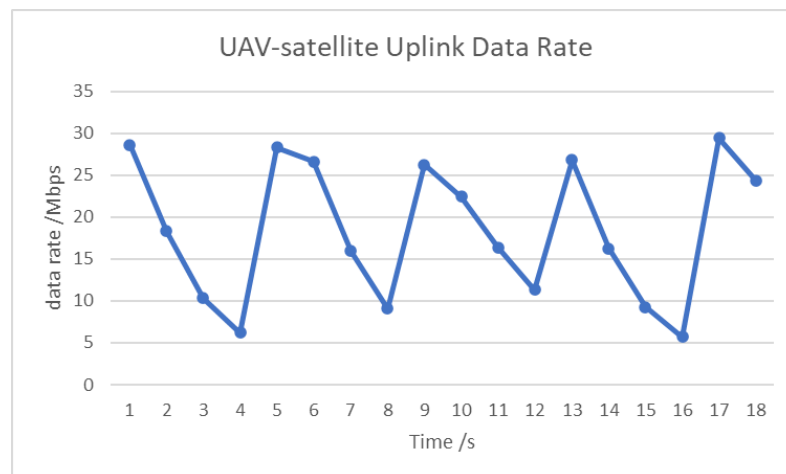


Figure 5.2: Predicted real-time uplink data rate between UAV and satellite

The uplink data rate is illustrated in Fig 5.2. Since the UAV always connects to the nearest satellite at each handover, it obtains the maximum available data rate at the beginning of each handover interval. As time progresses, the currently connected satellite gradually moves away, leading to a continuous decline in the data rate. At the next handover interval, the UAV switches to a closer satellite, resulting in a renewed increase in the uplink data rate.

5.2 Benchmark Algorithms

One of the major contributions of our proposed algorithm is the joint optimization of UAV trajectory and user association using the PPO algorithm, while allowing the UAV to simultaneously associate with multiple users. To validate the effectiveness and necessity of this joint variable design, we implement two representative benchmark algorithms:

1) **Optimized Trajectory and Single Association (OTSA)**: A PPO-based algorithm that jointly optimizes the UAV trajectory and a single-user association strategy.

2) **Optimized Trajectory and Conditional Association (OTCA)**: A PPO-based algorithm that optimizes only the UAV trajectory, with user association determined by predefined rules.

Both benchmark algorithms are designed with the same optimization objective and reward function as our proposed algorithm and are trained under identical environment conditions. The table below summarizes the parameters related to the training environment and algorithm settings.

Table 5.3: Simulation parameters

Parameter	Value
UAV energy storage: E^U	1 kWh
UAV velocity: v^U	5 m/s
channel power gain per meter: β_0	-20 dB
WPT power of UAV: P^{tx}	500 W
Learning rate:	8×10^{-4}
weight factor: w_1	10^{-9}
weight factor: w_2	10^{-3}
weight factor: w_3, w_4, w_5	0.01
Clip parameter:	0.2
Discount factor:	0.99
Update interval:	50
Episode length:	500
Dimensions of hidden layers for actor/critic networks:	256
Number of layers for actor/critic networks:	3
Number of parallel training environments:	8
Number of parallel evaluating environments:	4

The total rewards and the reward responding to data collection and user fairness retrieved during the training process by the proposed algorithm and

the two benchmark algorithms are present in Fig. 5.3, Fig. 5.4 and Fig. 5.5.

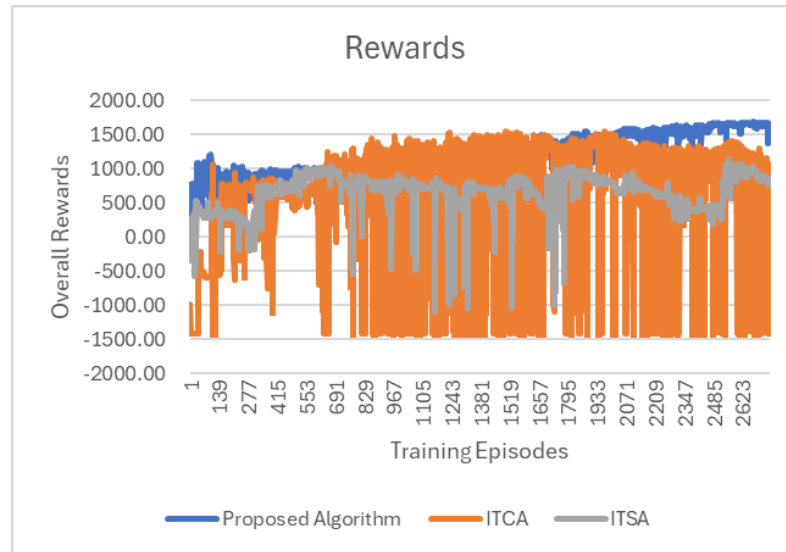


Figure 5.3: The average returns of the proposed algorithm and benchmarks

As shown in Fig. 5.3, the proposed algorithm achieves the highest total reward, followed by the OTCA algorithm, while the OTSA algorithm obtains the lowest. This indicates that the UAV-satellite uplink data rate is not always the bottleneck, and allowing the UAV to associate with multiple users simultaneously leads to more efficient data collection. The results in Fig. 5.4 further support this conclusion. Considering both data collection and user fairness, our proposed algorithm outperforms the two benchmark algorithms by 21.8% and 224.01%, respectively.

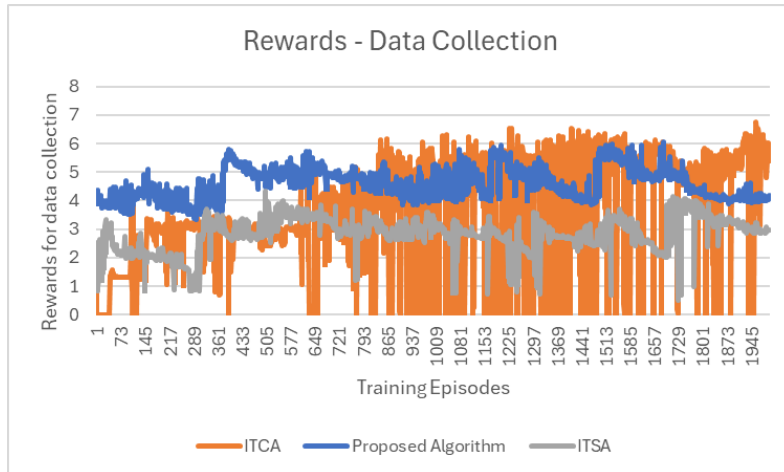


Figure 5.4: The average returns for data collection

As shown in Fig. 5.4, the OTCA algorithm achieves the best performance in data collection, demonstrating the effectiveness of the strategy in which the UAV prioritizes associating with nearby users. Our proposed algorithm collects 9.33% less data than OTCA, while OTSA performs 52.03% worse than our method. This is mainly due to the limited number of time steps, which prevents the UAV from reaching all users. As a result, our algorithm sometimes chooses to associate with more distant users, despite the reduced data collection efficiency. This observation is further supported by the results in Fig. 5.5. The OTSA algorithm is even more constrained by the time limit; since higher UAV-satellite data rates enable more efficient data collection through multi-user association, limiting the UAV to a single user leads OTSA to achieve the lowest overall reward.

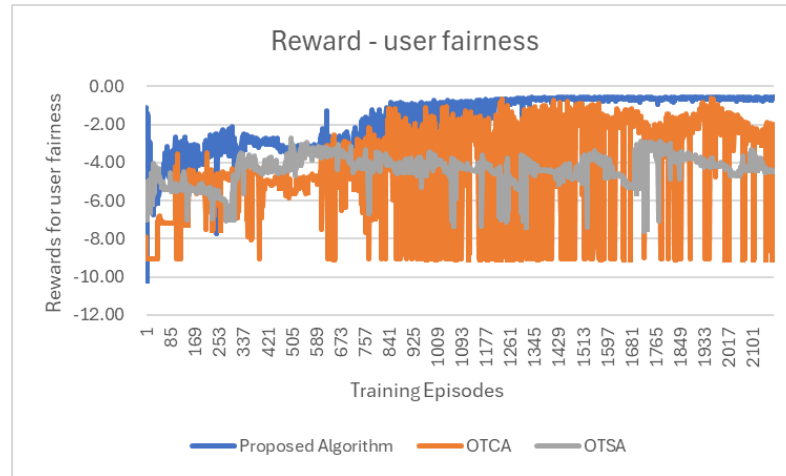


Figure 5.5: The average returns for user fairness

Our proposed algorithm achieves the best performance in terms of user fairness, indicating that the optimized association decisions are more intelligent than the fixed-range user connections adopted by the OTCA algorithm. Although our method sacrifices some data collection performance to ensure user fairness, it ultimately achieves a higher overall reward. This further demonstrates the effectiveness and rationality of the proposed algorithm.

5.3 Sensitivity Analysis for Weight Factors

In this section, to evaluate the sensitivity of the two weight factors in the reward function 4.14, the weight factor for data collection is kept constant while the factor for user fairness is varied. The effects of different weight values on the overall reward, the reward for data collection, and the reward for fairness are then analyzed.

The reward weight factor for data collection is fixed at 10^{-9} while the weight for user fairness is varied among $0, 0.25 \times 10^{-3}, 0.5 \times 10^{-3}$ and 0.75×10^{-3} . Fig.5.6, 5.7 and 5.8 illustrate the variation of rewards with respect to the fairness weight.

As shown in Fig.5.6, the average return decreases for all three algorithms as the fairness weight increases, indicating that a higher emphasis on user fairness leads to a reduction in overall reward. The proposed algorithm consistently outperforms both ITSA and ITCA across all fairness weight settings, achieving the highest average return and the smallest performance

degradation. In contrast ITSA shows the steepest decline, suggesting poor adaptability under higher fairness requirements due to the low efficiency of single user association. ITCA demonstrates intermediate performance, with a decline rate lower than that of ITSA but still notably greater than that of the proposed algorithm. These results indicate that the proposed algorithm maintains better robustness when balancing fairness and overall reward.

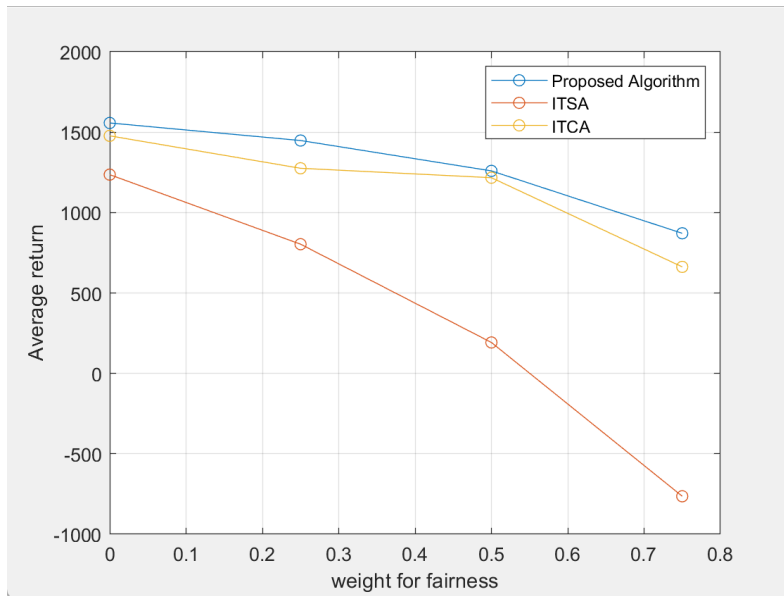


Figure 5.6: Effect of fairness weight on overall rewards

As illustrated in Fig.5.7, since data collection and user fairness are not in an absolute trade-off relationships, the data collection reward of ITSA remains almost unchanged when the fairness weight is small, i.e., from 0 to 0.25×10^{-3} , whereas both ITCA and the proposed algorithm exhibit a noticeable increase. However, when the fairness weight becomes larger, it begins to influence the RL agent's policy decisions, placing greater emphasis on user fairness, potentially compromising data collection performance. Consequently, as the fairness weight increases from 0.5×10^{-3} to 0.75×10^{-3} , the data collection reward of all three algorithms shows a certain degree of decline. Overall, ITSA demonstrates greater sensitivity to the fairness weight in terms of data collection performance, whereas ITCA and the proposed algorithm maintain comparatively higher robustness.

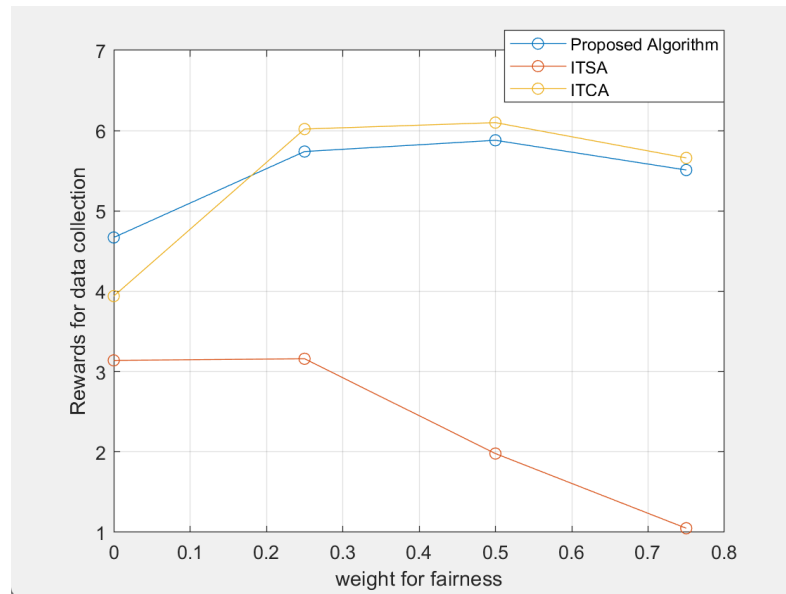


Figure 5.7: Effect of fairness weight on data collection rewards

As in Fig.5.8, the rewards for user fairness decrease across all three algorithms as the fairness weight increases, indicating a negative scaling effect in current reward formulation. The proposed algorithm achieves the highest fairness rewards for most fairness weight settings, demonstrating greater stability compared to the other methods. ITCA follows a similar trend but with slightly lower rewards, particularly at higher fairness weights. In contrast, ITSA exhibits the steepest decline, with fairness rewards dropping sharply and reaching the lowest value among all algorithms. These results suggest that the proposed algorithm is more robust in maintaining fairness-related performance, whereas ITSA is highly sensitive to increase in the fairness weight.

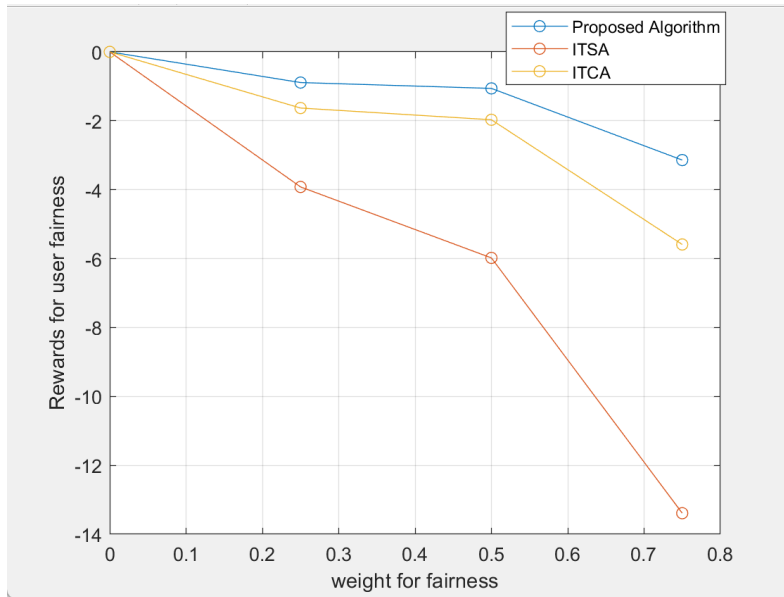


Figure 5.8: Effect of fairness weight on user fairness rewards

All three algorithms exhibit a negative growth trend as the fairness weight factor increases, with the proposed algorithm demonstrating the highest adaptability and ITSA the lowest. The poor adaptability of ITSA is due to its single-user association mechanism, which limits its ability to accommodate multiple users and consequently results in poor fairness. Furthermore, by comparing the curves of ITCA and the proposed algorithm across the three figures, it can be observed that as the fairness weight factor increases, while ITCA's data collection reward surpasses that of the proposed algorithm, its consistently lower fairness performance leads to lower overall rewards. This indicates that our algorithm, by flexibly selecting associated users, is able to better balance the rewards for two terms by sacrificing part of the data collection performance. This observation is consistent with the conclusion in Section 5.2.

5.4 Performance Analysis

In this section, we analyze and explain the final converged policy learned by the proposed algorithm. Furthermore, we examine its performance across different environments to demonstrate the effectiveness and strong adaptability of the algorithm.

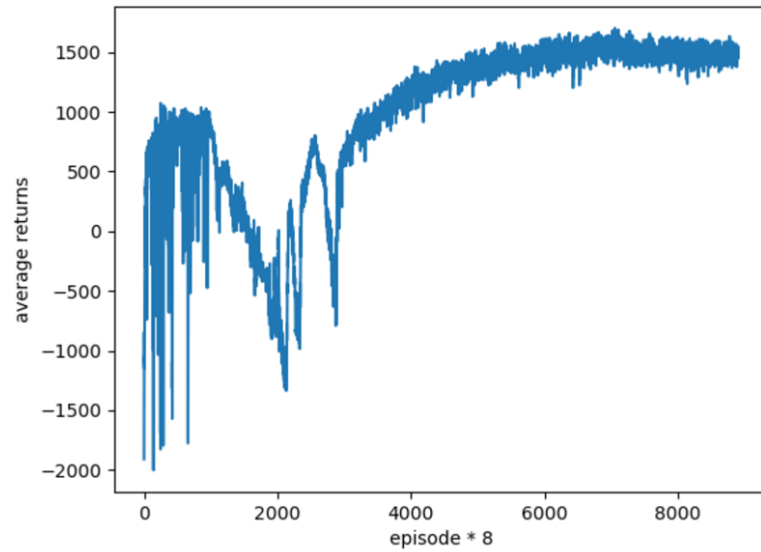


Figure 5.9: The average overall returns during training period

The proposed algorithm is trained over 64000 episodes and evaluated over every set of eight episodes. As shown in Fig. 5.9, the average return exhibits large fluctuations due to extensive exploration and policy updates with high variance in the early stage. A temporary performance drop occurs in the middle stage as the agent explores sub-optimal strategies before stabilizing. As introduced in 4.1, the convergence of the RNN-PPO algorithm is supported by the theoretical foundations of the policy gradient method, where the clipped surrogate objective in PPO constrains policy updates, ensuring monotonic improvement and avoiding destructive parameter shifts. The RNN architectures enhances the policy's ability to capture temporal dependencies, thereby improving state representation over time steps. The use of GAE further reduces the variance of policy gradient estimates. Under these design choices, the average return stabilizes at a high value after 32000 (4000*8) episodes. This indicates that the UAV trajectory designing and association scheduling policy has been converged to a near-optimal solution in the given environment.

Fig. 5.10 illustrates the UAV trajectory learned by our proposed algorithm in the depicted scenario. The red dots represent the locations of the GUs, while the blue line shows the UAV's flight path over the entire time slots. In this scenario, all GUs are assumed to have the same amount of data (0.6 GB) to be collected. It can be observed that the UAV hovers over the areas where users are densely distributed. In contrast, for regions with only a single user, the

UAV simply flies over without stopping, as doing so is sufficient to complete the data collection from that user.

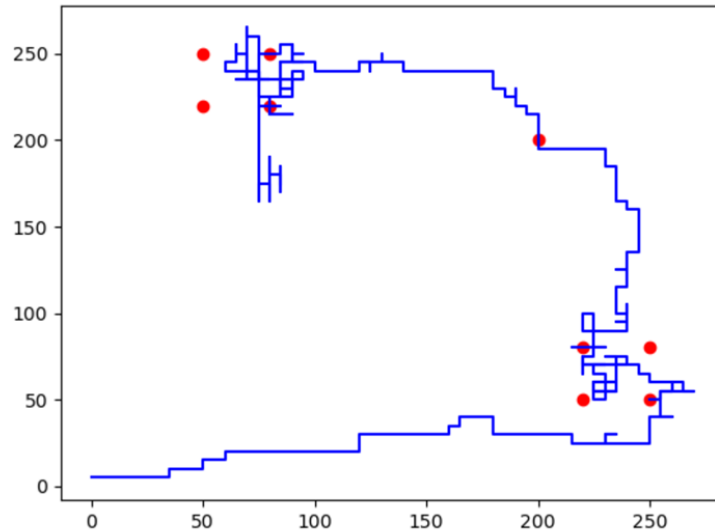


Figure 5.10: UAV trajectory with unevenly distributed users and uniform data collection requirements

Keeping the positions of the GUs unchanged, we increase the amount of data to be collected from the GU in the middle to 1 GB, while reducing that of the other GUs to 0.3 GB.

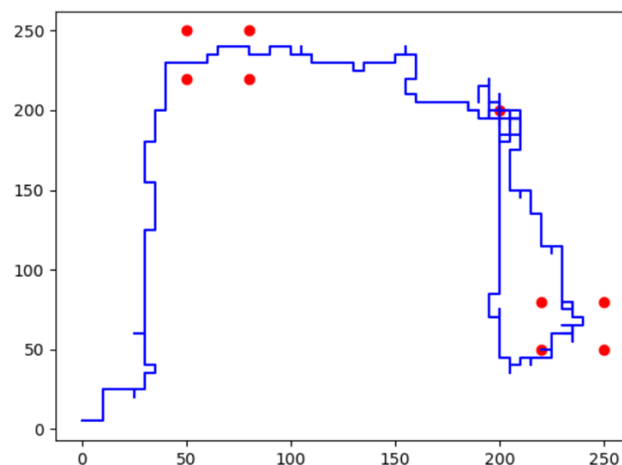


Figure 5.11: UAV trajectory with unevenly distributed users and varying data collection requirements

efficiency.

We increase the amount of data to be collected from the user located in the lower-right corner to 1GB, while reducing the data requirements of the other GUs to 0.3 GB. The UAV trajectory is shown in Fig. 5.13.

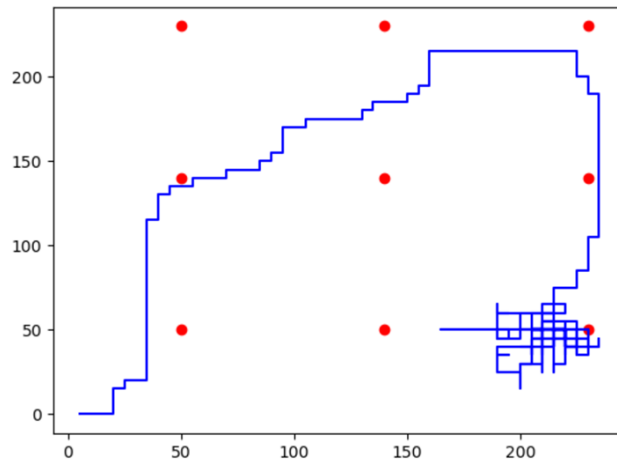


Figure 5.13: UAV trajectory with evenly distributed users and varying data collection requirements

After increasing the data demand of the GU located in the lower-right corner, it is evident that the UAV hovers near this user. However, since the reward function incorporates a user fairness metric, the UAV must also ensure that other users receive data collection services. As a result, the trajectory still shows the UAV flying over multiple users or traversing diagonally through regions containing several users. This again demonstrates that, despite the changes in the spatial distribution of GUs, the UAV can intelligently adjust its flight path and user association based on both user locations and data demands, highlighting the effectiveness and adaptability of our proposed algorithm across varying environments.

Since the UAV is required to provide wireless power transfer to its associated users while collecting data, the total energy supplies must exceed the user's consumption during this period. Fig. 5.14 illustrates the UAV's energy charging performance.

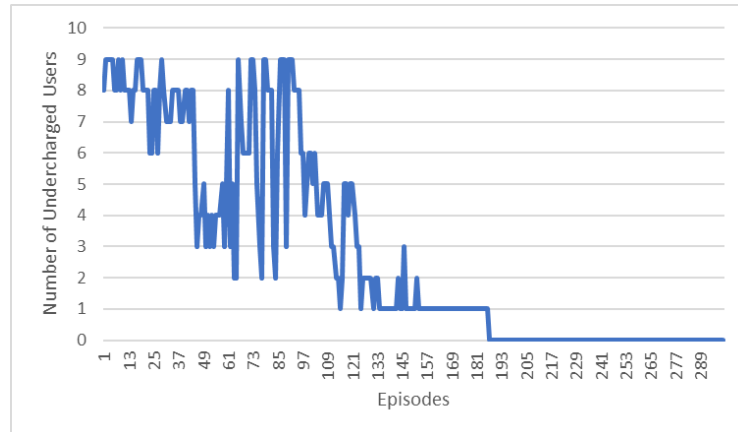


Figure 5.14: Number of undercharged user

It is evident that as the number of training episodes increases, the number of users who do not receive sufficient energy decreases and eventually reaches zero. This indicates that during training, the UAV learns to consider the charging efficiency of each user when making association decisions. Ultimately, the UAV is able to ensure that all users receive sufficient energy.

5.5 Summary

This chapter summarizes the simulation setup and analyzes the results in three parts. First, we simulate a dynamic multi-orbit LEO satellite constellation, showing the varies of UAV-satellite data rate over time due to satellite movement. Second, we compare our proposed algorithm with two benchmarks in terms of convergence performance, reward performance, data collection, and user fairness. Our method, which jointly optimizes UAV trajectory and multi-user association, achieves the highest reward. Finally, we analyze the learned UAV trajectories and charging performance, demonstrating that the algorithm adapts intelligently to user distribution, data demands, and energy needs. Overall, the proposed PPO-based algorithm effectively achieves its objectives and exhibits strong adaptability.

Chapter 6

Conclusions and Future work

This chapter first presents the conclusion based on the simulation results discussed in the previous chapter. It then addressed the limitations and shortcomings of this thesis. Finally, potential directions for future work are outlined.

6.1 Conclusions

In this thesis, we investigate an optimization problem in a SAGIN system, where a UAV is required to design its trajectory and schedule user associations with the objective of maximizing data collection from GUs while ensuring user fairness. A dynamic Walker-Delta LEO satellite constellation is incorporated to model the time-varying uplink data rate from the UAV to LEO satellites. Simulation results show that the uplink data rate fluctuates between 30 Mbps and 8 Mbps, gradually decreasing after each satellite handover due to increasing distance.

To solve this problem, we employ an RNN-based PPO algorithm to obtain a near-optimal solution. Two benchmark algorithms, OTCA and OTSA, are designed to evaluate the performance of the proposed method. Our algorithm achieves the highest overall reward, outperforming OTCA and OTSA by 21.8% and 224.01%, respectively. Due to the limited number of time steps, the UAV sometimes chooses to associate with distant users, which sacrifices some data collection performance in exchange for improved fairness. Specifically, the proposed algorithm performs 9.33% worse in data collection compared to OTCA—the best among the benchmarks—but achieves the highest level of user fairness, outperforming OTCA by 64.53%.

Analysis of the UAV's learned trajectory and association decisions

further demonstrates that the UAV can intelligently adapt its movement and connection strategy based on the users' locations, data transmission demands, and wireless charging efficiency.

6.2 Limitations

In this thesis, all decisions are determined by a single UAV, and the optimization problem is solved using a single-agent PPO algorithm. This leads to a high-dimensional action and state space, which increases further as the number of users grows, resulting in limited scalability. High-dimensional action spaces require more training time and may even hinder convergence, posing a potential risk.

Moreover, due to the complexity of the environment considered, many parameters—such as the size of the service area and the UAV's charging power—may have a significant impact on the UAV's decision-making and overall performance. These factors have not been explored in this work and represent an important limitation of the thesis.

6.3 Future work

To address the limitation that the current algorithm struggles to handle a large number of GUs, we aim to develop a more scalable approach. Specifically, the UAV should not be solely responsible for all decision-making. Therefore, we consider adopting a multi-agent framework in which users make their own decisions about whether to associate with the UAV. Additionally, more scalable learning paradigms such as federated learning can be employed to mitigate the convergence issues commonly encountered in traditional multi-agent reinforcement learning.

In this thesis, the UAV is constrained to fly at a fixed altitude and select its direction from only four predefined options, which limits the flexibility of its trajectory designing. Future work can extend the UAV's movement options to include eight directions, or discretize the flight directions into finer angular intervals. Moreover, enabling the UAV to adjust its flight altitude would allow it to maneuver more dynamically, potentially improving overall performance. The Deep Deterministic Policy Gradient (DDPG) algorithm, which supports continuous action spaces, could also be utilized to allow the UAV to select arbitrary flight directions.

In future work, we can explore how varying environmental parameters

influences the decision-making strategies adopted by the reinforcement learning model. For instance, in the previous chapter, we observed a trade-off between data collection and user fairness due to the limited number of time steps. However, to further validate this trade-off, additional experiments are needed, such as increasing the number of time steps or reducing the size of the service area.

Analyzing how the performance of benchmark algorithms changes with these parameters can also help explain the observed differences in performance and reveal the strategies each algorithm may have adopted. Since reinforcement learning models often operate as black boxes, more empirical results are necessary to interpret and understand their behavior.

References

- [1] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, and M. Peng, “Vision, requirements, and technology trend of 6g: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed,” *IEEE Wireless Communications*, vol. 27, no. 2, pp. 218–228, 2020. [Page 1.]
- [2] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kotheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani *et al.*, “Evolution of non-terrestrial networks from 5g to 6g: A survey,” *IEEE communications surveys & tutorials*, vol. 24, no. 4, pp. 2633–2672, 2022. [Pages 1 and 8.]
- [3] B. Mölleryd, M. Ozger, M. Westring, A. Nordlöw, D. Schupke, U. Engström, C. Cavdar, M. Lindborg, and N. Sciammetta, “Regulatory and spectrum policy challenges for combined airspace and non-terrestrial networks,” *Telecommunications Policy*, vol. 49, no. 1, p. 102875, 2025. [Page 1.]
- [4] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio, and R. Prasad, “Satellite communications supporting internet of remote things,” *IEEE Internet Things J.*, vol. 3, no. 1, pp. 113–123, 2015. [Page 1.]
- [5] A. Baltaci, E. Dinc, M. Ozger, A. Alabbasi, C. Cavdar, and D. Schupke, “A survey of wireless networks for future aerial communications (facom),” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2833–2884, 2021. [Page 1.]
- [6] T. Chen, J. Liu, Q. Ye, W. Zhuang, W. Zhang, T. Huang, and Y. Liu, “Learning-based computation offloading for IoRT through Ka/Q-band satellite–terrestrial integrated networks,” *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12 056–12 070, 2021. [Page 1.]
- [7] S. Zhang, M. Ozger, S. S. Seeram, I. Godor, L. Feltrin, A. Nordlow, J. Pfeifle, L. Toka, G. Biczok, D. A. Schupke *et al.*, “6g for connected

- sky: Holistic adaptive combined airspace and non terrestrial network architecture,” *IEEE Wireless Communications*, 2025. [Page 1.]
- [8] S. S. Sri Ganesh Seeram, L. Feltrin, M. Özger, C. Cavdar *et al.*, “Digital twin-based optimization of service availability in leo mega constellations considering handover delays in open ran,” 2025. [Page 2.]
- [9] S. Liu, S. Zhang, and C. Cavdar, “Task offloading strategy for dynamic leo satellite and cloud networks: A deep reinforcement learning-based approach,” in *ICC 2025-IEEE International Conference on Communications*. IEEE, 2025, pp. 2454–2459. [Page 2.]
- [10] C. Dai, K. Zhu, and E. Hossain, “Multi-agent deep reinforcement learning for joint decoupled user association and trajectory design in full-duplex multi-uav networks,” *IEEE transactions on mobile computing*, vol. 22, no. 10, pp. 6056–6070, 2022. [Page 2.]
- [11] S. S. S. G. Seeram, L. Feltrin, M. Ozger, S. Zhang, and C. Cavdar, “Handover challenges in disaggregated open ran for leo satellites: tradeoff between handover delay and onboard processing,” *Frontiers in Space Technologies*, vol. 6, p. 1580005, 2025. [Page 2.]
- [12] S. S. Sri Ganesh Seeram, L. Feltrin, M. Özger, S. Zhang, C. Cavdar *et al.*, “Handover delay minimization in non-terrestrial networks: Impact of open ran functional splits,” in *12th Advanced Satellite Multimedia Systems Conference and the 18th Signal Processing for Space Communications Workshop, ASMS/SPSC 2025, Sitges, Spain, Feb 26 2025-Feb 28 2025*. Institute of Electrical and Electronics Engineers (IEEE), 2025. [Page 2.]
- [13] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, “Space/aerial-assisted computing offloading for iot applications: A learning-based approach,” *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1117–1129, 2019. [Pages 7, 10, and 11.]
- [14] Y. Qiu, J. Niu, X. Zhu, K. Zhu, Y. Yao, B. Ren, and T. Ren, “Mobile edge computing in space-air-ground integrated networks: Architectures, key technologies and challenges,” *Journal of Sensor and Actuator Networks*, vol. 11, no. 4, p. 57, 2022. [Pages 7 and 9.]
- [15] D.-H. Jung, H. Nam, J. Choi, and D. J. Love, “Modeling and analysis of geo satellite networks,” *IEEE Transactions on Wireless Communications*, 2024. [Page 8.]

- [16] Z. Xiao, J. Yang, T. Mao, C. Xu, R. Zhang, Z. Han, and X.-G. Xia, "Leo satellite access network (leo-san) toward 6g: Challenges and approaches," *IEEE Wireless Communications*, vol. 31, no. 2, pp. 89–96, 2022. [Page 8.]
- [17] S. S. S. G. Seeram, L. Feltrin, M. Ozger, S. Zhang, and C. Cavdar, "Feasibility study of function splits in ran architectures with leo satellites," in *2024 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*. IEEE, 2024, pp. 622–627. [Page 8.]
- [18] F. Zhou, P. Wang, M. Ozger, and C. Cavdar, "Blind detection of drones using ofdm-based zadoff-chu sequences with field tests," in *ICC 2025-IEEE International Conference on Communications*. IEEE, 2025, pp. 1482–1487. [Page 9.]
- [19] M. Vondra, M. Ozger, D. Schupke, and C. Cavdar, "Integration of satellite and aerial communications for heterogeneous flying vehicles," *IEEE network*, vol. 32, no. 5, pp. 62–69, 2018. [Page 9.]
- [20] R. Neetu, O. A. Topal, Ö. T. Demir, E. Björnson, C. Cavdar, G. Ghatak, and V. A. Bohara, "Uav-based cell-free massive mimo: Joint activation and power optimization under fronthaul capacity limitations," *IEEE Wireless Communications Letters*, 2025. [Page 9.]
- [21] C. Zhou, W. Wu, H. He, P. Yang, F. Lyu, N. Cheng, and X. Shen, "Deep reinforcement learning for delay-oriented iot task scheduling in sagin," *IEEE Transactions on Wireless Communications*, vol. 20, no. 2, pp. 911–925, 2020. [Page 9.]
- [22] M. D. Nguyen, L. B. Le, and A. Girard, "Integrated computation offloading, uav trajectory control, edge-cloud and radio resource allocation in sagin," *IEEE Transactions on Cloud Computing*, vol. 12, no. 1, pp. 100–115, 2023. [Pages 9, 11, and 12.]
- [23] A. H. Arani, P. Hu, and Y. Zhu, "Uav-assisted space-air-ground integrated networks: A technical review of recent learning algorithms," *IEEE Open Journal of Vehicular Technology*, 2024. [Page 9.]
- [24] Q. Chen, W. Meng, S. Han, C. Li, and H. H. Chen, "Effect of intelligent multi-association in civil aircraft-augmented sagin," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 1, pp. 223–238, 2022. [Page 9.]

- [25] J. Xu, Y. Zeng, and R. Zhang, "Uav-enabled wireless power transfer: Trajectory design and energy optimization," *IEEE transactions on wireless communications*, vol. 17, no. 8, pp. 5092–5106, 2018. [Pages 9, 12, and 23.]
- [26] S. Zhang, W. Liu, and N. Ansari, "Joint wireless charging and data collection for uav-enabled internet of things network," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 852–23 859, 2022. [Pages 9 and 12.]
- [27] M. Masoudi and C. Cavdar, "Device vs edge computing for mobile services: Delay-aware decision making to minimize power consumption," *IEEE Transactions on Mobile Computing*, vol. 20, no. 12, pp. 3324–3337, 2020. [Page 9.]
- [28] J. Chen, H. Zhang, and Z. Xie, "Space-air-ground integrated network (sagin): A survey," *arXiv preprint arXiv:2307.14697*, 2023. [Page 10.]
- [29] F. Tang, H. Hofner, N. Kato, K. Kaneko, Y. Yamashita, and M. Hangai, "A deep reinforcement learning-based dynamic traffic offloading in space-air-ground integrated networks (sagin)," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 276–289, 2021. [Pages 10 and 11.]
- [30] S. Zhang, A. Liu, C. Han, X. Liang, X. Xu, and G. Wang, "Multiagent reinforcement learning-based orbital edge offloading in sagin supporting internet of remote things," *IEEE Internet of Things Journal*, vol. 10, no. 23, pp. 20 472–20 483, 2023. [Pages 10 and 11.]
- [31] S. Jung, S. Jeong, J. Kang, and J. Kang, "Marine iot systems with space-air-sea integrated networks: Hybrid leo and uav edge computing," *IEEE Internet of Things Journal*, vol. 10, no. 23, pp. 20 498–20 510, 2023. [Pages 10, 11, and 12.]
- [32] M. Ozger, I. Godor, A. Nordlow, T. Heyn, S. Pandi, I. Peterson, A. Viseras, J. Holis, C. Raffelsberger, A. Kercek, B. Mölleryd, L. Toka, G. Biczok, R. de Candido, F. Laimer, U. Tarmann, D. Schupke, and C. Cavdar, "6g for connected sky: A vision for integrating terrestrial and non-terrestrial networks," in *2023 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, 2023, pp. 711–716. [Page 10.]

- [33] I. A. Meer, M. Ozger, M. Lundmark, K. W. Sung, and C. Cavdar, "Ground based sense and avoid system for air traffic management," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2019, pp. 1–6. [Page 10.]
- [34] A. Manzoor, M. Ozger, D. Schupke, and C. Cavdar, "Combined airspace and non-terrestrial 6g networks for advanced air mobility," in *2024 20th International Conference on the Design of Reliable Communication Networks (DRCN)*, 2024, pp. 47–53. [Page 10.]
- [35] I. A. Meer, M. Ozger, and C. Cavdar, "Cellular localizability of unmanned aerial vehicles," *Vehicular Communications*, vol. 44, p. 100677, 2023. [Page 10.]
- [36] Y. Deng, I. A. Meer, S. Zhang, M. Ozger, and C. Cavdar, "D3qn-based trajectory and handover management for uavs co-existing with terrestrial users," in *2023 21st International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2023, pp. 103–110. [Page 10.]
- [37] Y. Deng, S. Zhang, I. A. Meer, M. Ozger, and C. Cavdar, "Joint trajectory and handover management for uavs co-existing with terrestrial users: A multi-agent drl approach," *IEEE Transactions on Cognitive Communications and Networking*, 2025. [Page 10.]
- [38] I. A. Meer, K.-L. Besser, M. Ozger, D. Schupke, H. V. Poor, and C. Cavdar, "Learning based dynamic cluster reconfiguration for uav mobility management with 3d beamforming," in *2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, 2024, pp. 486–491. [Page 10.]
- [39] I. A. Meer, M. Ozger, D. A. Schupke, and C. Cavdar, "Mobility management for cellular-connected uavs: Model-based versus learning-based approaches for service availability," *IEEE Transactions on Network and Service Management*, vol. 21, no. 2, pp. 2125–2139, 2024. [Page 10.]
- [40] I. A. Meer, K.-L. Besser, M. Ozgerl, H. V. Poor, and C. Cavdar, "Reinforcement learning based dynamic power control for uav mobility management," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*, 2023, pp. 724–728. [Page 10.]

- [41] I. A. Meer, B. Hörmann, M. Ozger, F. Geyer, A. Viseras, D. Schupke, and C. Cavdar, “Explainable ai for uav mobility management: A deep q-network approach for handover minimization,” *arXiv preprint arXiv:2504.18371*, 2025. [Page 10.]
- [42] F. Giarre, I. A. Meer, M. Masoudi, M. Ozger, and C. Cavdar, “Hierarchical multi agent drl for soft handovers between edge clouds in open ran,” *arXiv preprint arXiv:2503.08493*, 2025. [Page 10.]
- [43] I. A. Meer, K.-L. Besser, M. Ozger, D. Schupke, H. V. Poor, and C. Cavdar, “Hierarchical multi-agent drl based dynamic cluster reconfiguration for uav mobility management,” *arXiv preprint arXiv:2412.16167*, 2024. [Page 10.]
- [44] I. A. Meer, “Ai assisted mobility management for cellular connected uavs,” Ph.D. dissertation, KTH Royal Institute of Technology, 2025. [Page 10.]
- [45] J. Chen, M. Ozger, and C. Cavdar, “Nash soft actor-critic leo satellite handover management algorithm for flying vehicles,” in *2024 IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, 2024, pp. 380–385. [Page 10.]
- [46] A. E. Garcia, M. Ozger, A. Baltaci, S. Hofmann, D. Gera, M. Nilson, C. Cavdar, and D. Schupke, “Direct air to ground communications for flying vehicles: Measurement and scaling study for 5g,” in *2019 IEEE 2nd 5G World Forum (5GWF)*, 2019, pp. 310–315. [Page 10.]
- [47] S. Hofmann, V. Megas, M. Ozger, D. Schupke, F. H. P. Fitzek, and C. Cavdar, “Combined optimal topology formation and rate allocation for aircraft to aircraft communications,” in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6. [Page 11.]
- [48] V. Megas, S. Hoppe, M. Ozger, D. Schupke, and C. Cavdar, “A combined topology formation and rate allocation algorithm for aeronautical ad hoc networks,” *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 12–28, 2024. [Page 11.]
- [49] S. Li, F. Wu, S. Luo, Z. Fan, J. Chen, and S. Fu, “Dynamic online trajectory planning for a uav-enabled data collection system,” *IEEE Transactions on Vehicular Technology*, vol. 71, no. 12, pp. 13 332–13 343, 2022. [Page 11.]

- [50] X. Wang and M. C. Gursoy, "Learning-based uav trajectory optimization with collision avoidance and connectivity constraints," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4350–4363, 2021. [Page 11.]
- [51] Y. He, Y. Gan, H. Cui, and M. Guizani, "Fairness-based 3-d multi-uav trajectory optimization in multi-uav-assisted mec system," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 383–11 395, 2023. [Page 11.]
- [52] M. Ozger, M. Vondra, and C. Cavdar, "Towards beyond visual line of sight piloting of uavs with ultra reliable low latency communication," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6. [Page 11.]
- [53] P. Wang, M. Ozger, C. Cavdar, and M. Petrova, "Beyond visual line of sight piloting of uavs using millimeter-wave cellular networks," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2019, pp. 1–7. [Page 11.]
- [54] F. Salehi, M. Ozger, N. Neda, and C. Cavdar, "Ultra-reliable low-latency communication for aerial vehicles via multi-connectivity," in *2022 Joint European Conference on Networks and Communications 6G Summit (EuCNC/6G Summit)*, 2022, pp. 166–171. [Page 11.]
- [55] F. Salehi, M. Ozger, and C. Cavdar, "Reliability and delay analysis of 3-dimensional networks with multi-connectivity: Satellite, haps, and cellular communications," *IEEE Transactions on Network and Service Management*, vol. 21, no. 1, pp. 437–450, 2024. [Page 11.]
- [56] X. Wang, L. Wang, C. Cavdar, M. Tornatore, G. B. Figueiredo, H. S. Chung, H. H. Lee, S. Park, and B. Mukherjee, "Handover reduction in virtualized cloud radio access networks using TWDM-PON fronthaul," *Journal of Optical Communications and Networking*, vol. 8, no. 12, pp. B124–B134, 2016, publisher: Optica Publishing Group. [Page 11.]
- [57] A. Ferdowsi, M. A. Abd-Elmagid, W. Saad, and H. S. Dhillon, "Neural combinatorial deep reinforcement learning for age-optimal joint trajectory and scheduling design in uav-assisted networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1250–1265, 2021. [Page 12.]

- [58] C. Dai, K. Zhu, and E. Hossain, “Multi-agent deep reinforcement learning for joint decoupled user association and trajectory design in full-duplex multi-uav networks,” *IEEE transactions on mobile computing*, vol. 22, no. 10, pp. 6056–6070, 2022. [Pages 12 and 21.]
- [59] J. Tepper, A. Baltaci, B. Schleicher, A. Drexler, S. Duhovnikov, M. Ozger, M. Tavana, C. Cavdar, and D. Schupke, “Evaluation of rf wireless power transfer for low-power aircraft sensors,” in *2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC)*. IEEE, 2020, pp. 1–6. [Page 12.]
- [60] M. Tavana, M. Ozger, A. Baltaci, B. Schleicher, D. Schupke, and C. Cavdar, “Wireless power transfer for aircraft iot applications: System design and measurements,” *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 11 834–11 846, 2021. [Page 12.]
- [61] Q. Chen, Z. Guo, W. Meng, S. Han, C. Li, and T. Q. Quek, “A survey on resource management in joint communication and computing-embedded sagin,” *IEEE Communications Surveys & Tutorials*, 2024. [Page 12.]
- [62] Y. Liu, L. Jiang, Q. Qi, K. Xie, and S. Xie, “Online computation offloading for collaborative space/aerial-aided edge computing toward 6g system,” *IEEE Transactions on Vehicular Technology*, vol. 73, no. 2, pp. 2495–2505, 2023. [Page 12.]
- [63] Z. Hu, F. Zeng, Z. Xiao, B. Fu, H. Jiang, H. Xiong, Y. Zhu, and M. Alazab, “Joint resources allocation and 3d trajectory optimization for uav-enabled space-air-ground integrated networks,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 11, pp. 14 214–14 229, 2023. [Page 12.]
- [64] Y. Hao, Z. Song, Z. Zheng, Q. Zhang, and Z. Miao, “Joint communication, computing, and caching resource allocation in leo satellite mec networks,” *IEEE Access*, vol. 11, pp. 6708–6716, 2023. [Page 12.]
- [65] M. Masoudi, M. G. Khafagy, E. Soroush, D. Giacomelli, S. Morosi, and C. Cavdar, “Reinforcement learning for traffic-adaptive sleep mode management in 5g networks,” in *2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications*. IEEE, 2020, pp. 1–6. [Page 12.]

- [66] S. K. G. Peesapati, M. Olsson, M. Masoudi, S. Andersson, and C. Cavdar, "Q-learning based radio resource adaptation for improved energy performance of 5g base stations," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 979–984. [Page 12.]
- [67] M. Masoudi, E. Soroush, J. Zander, and C. Cavdar, "Digital twin assisted risk-aware sleep mode management using deep q-networks," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 1224–1239, 2022. [Page 12.]
- [68] S. Zhang, T. Cai, D. Wu, D. Schupke, N. Ansari, and C. Cavdar, "Iort data collection with leo satellite-assisted and cache-enabled uav: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 4, pp. 5872–5884, 2023. [Page 12.]
- [69] O. A. Topal, Q. He, O. T. Demir, M. Masoudi, and C. Cavdar, "Drl-based joint ap deployment and network-centric cluster formation for maximizing long-term energy efficiency in cell-free massive mimo," in *2023 57th Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2023, pp. 993–999. [Page 12.]
- [70] S. Zhang, T. Cai, Ö. T. Demir, and C. Cavdar, "Multi-agent rl for sleep mode and antenna configuration with user offloading under dynamic traffic in massive mimo networks," *IEEE Transactions on Vehicular Technology*, 2025. [Page 12.]
- [71] D. Zhu, H. Liu, T. Li, J. Sun, J. Liang, H. Zhang, L. Geng, and Y. Liu, "Deep reinforcement learning-based task offloading in satellite-terrestrial edge computing networks," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–7. [Page 13.]
- [72] I. Ullah, H.-K. Lim, Y.-J. Seok, and Y.-H. Han, "Optimizing task offloading and resource allocation in edge-cloud networks: a drl approach," *Journal of Cloud Computing*, vol. 12, no. 1, p. 112, 2023. [Page 13.]
- [73] Z. Qin, H. Yao, T. Mai, D. Wu, N. Zhang, and S. Guo, "Multi-agent reinforcement learning aided computation offloading in aerial computing for the internet-of-things," *IEEE Transactions on Services Computing*, vol. 16, no. 3, pp. 1976–1986, 2022. [Page 13.]

- [74] A. Filippone, *Flight performance of fixed and rotary wing aircraft*. Elsevier, 2006. [Page 23.]
- [75] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv preprint arXiv:1707.06347*, 2017. [Page 28.]

