

Game Theory and Meta Learning for Optimization of Integrated Satellite-Drone-Terrestrial-Communication Systems

Ye Hu

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Electrical Engineering

Walid Saad, Chair
T. Charles Clancy
Jeffrey H. Reed
Kekatos Vassilis
Craig A. Woolsey

Aug 2, 2021
Blacksburg, Virginia

Keywords: Game theory, Machine Learning, Meta Learning, Performance Optimization, Resource allocation, Unmanned Aerial Vehicle Communications, Satellite communication

Copyright 2021, Ye Hu

Game Theory and Meta Learning for Optimization of Integrated Satellite-Drone-Terrestrial-Communication Systems

Ye Hu

ABSTRACT

Emerging integrated satellite-drone-terrestrial communication (ISDTC) technologies are expected to contribute to our life by bringing forth high speed wireless connectivity to every corner of the world. On the one hand, the Internet of Things (IoT) provides connectivity to various physical objects by enabling them to share information and to coordinate decisions. On the other hand, the non-terrestrial components of an ISDTC system, i.e. unmanned aerial vehicles (UAVs), and satellites, can boost the capacity of wireless networks by providing services to hotspots, disaster-affected, and rural areas. Despite the several benefits and practical applications of ISDTC technologies, one must address many technical challenges such as, resource management, trajectory design, device cooperation, data routing, and security. The key goal of this dissertation is to develop analytical foundations for the optimization of ISDTC operations, and the deployment of non-terrestrial networks (NTNs). First, the problem of resource management within ISDTC systems is investigated for service-effective cooperation among the terrestrial networks and NTNs. The performance of a multi-layer ISDTC system is analyzed within a competitive market setting. Using a novel decentralized algorithm, spectrum resources are allocated to each one of the communication links, considering the fairness among devices. The proposed algorithm is proved to reach a Walrasian equilibrium, at which the sum-rate of the network is maximized. The results also show that the proposed algorithm can reach the equilibrium with a practical convergence speed. Then, the effective deployment of NTNs under environmental dynamics is investigated using machine learning solutions with meta training capabilities. First, the use of satellites for on-demand coverage to unforeseeable radio access needs is investigated using game theory. The optimal data routing strategies are learned by the satellite system, using a novel reinforcement learning approach with distribution-robust meta training capability. The results show that, the proposed meta training mechanism significantly reduces the learning cost on the satellites, and is guaranteed to reach the maximal service coverage in the system. Next, the problem of control of UAV-carried radio access points under energy constraints is studied. In particular, novel frameworks are proposed to design trajectories for UAVs that seek to deliver data service to distributed, dynamic, and unforeseeable wireless access requests. The results show that the proposed approaches are guaranteed to converge to an optimal trajectory, and can get a faster convergence speed and lower computation cost using decomposition, cross validation and meta learning. Finally, this dissertation looks at the security of an IoT system. In particular, the impact of human intervention on the system security is analyzed under specific resource constraints. Psychological game theory frameworks are proposed to analyze the human psychology and its impact on the security of the system. The results show that the proposed solution can help the defender optimize its connectivity within the IoT system by estimating the attacker's behavior. In summary, the outcomes of this dissertation provide key guidelines for the effective deployment of ISDTC systems.

Game Theory and Meta Learning for Optimization of Integrated Satellite-Drone-Terrestrial-Communication Systems

Ye Hu

General Audience Abstract

In the past decade, the goal of providing wireless connectivity to all individuals and communities, including the most disadvantaged ones, has become a national priority both in the US and globally. Yet, remarkably, until today, there is still a great portion of the Earth's population who falls out of today's wireless broadband coverage. While people who live in under-developed or rural areas are still in "wireless darkness", communities in megacities often experience below-par wireless service due to their overloaded communication systems. To provide high-speed, reliable wireless connectivity to those on the less-served side of the digital divide, an integrated space-air-ground communication system can be designed. Indeed, airborne and space-based non terrestrial networks (NTNs) can enhance the capacity and coverage of existing wireless cellular networks (e.g., 5G and beyond) by providing supplemental, affordable, flexible, and reliable service to users in rural, disaster affected, and over-crowded areas.

In order to fill the coverage holes and bridge the digital divide, seamless integration among NTNs and terrestrial networks is needed. In particular, when deploying an integrated communication system, one must consider the problems of spectrum management, device cooperation, trajectory design, and data routing within the system. Meanwhile, with the increased exposure to malicious attacks on high altitude platforms and vulnerable IoT devices, the security within the integrated system must be analyzed and optimized for reliable data service.

To overcome all the technological challenges that hinder the realization of global digital inclusion, this dissertation uses techniques from the fields of game theory, meta learning, and optimization theory to deploy, control, coordinate, and manage terrestrial networks and NTNs. The anticipated results show that a properly integrated satellite-drone-terrestrial communication (ISDTC) system can deliver cost-effective, high speed, seamless wireless service to our world.

To my parents, my brother, and my dear husband

Acknowledgments

First and foremost, I want to express my deepest gratitude to my Ph.D. advisor, Dr. Walid Saad, for his continuous support and guidance of my Ph.D study. His immense knowledge, patience, motivation, and experience have been of great help to me in my Ph.D. work, including this dissertation¹. I am grateful to him for always believing in me and giving me the freedom to pursue diverse, yet coherent research directions, which has made my Ph.D. experience stimulating and productive. I would also like to thank you for all the encouragement, trust, and priceless advice, which have helped me to grow as a better research scientist, and enlightened me to find the right path for my future career.

I would like to thank the members of my Ph.D. advisory committee, Dr. T. Charles Clancy, Dr. Jeffrey H. Reed, Dr. Kekatos Vassilis, and Dr. Craig Woolsey for their valuable, insightful comments which have helped me to substantially improve the quality of this dissertation. I greatly appreciate the invaluable contribution of my co-authors, Dr. Nof Abuzainab, Dr. Anibal Sanjab, Dr. Mingzhe Chen, Dr. H. Vincent Poor, and Dr. Shuguang Cui. I am extremely grateful to have the privilege to collaborate with them and improve my research work throughout the Ph.D. study with their insightful advice.

I would like to thank all my friends and colleagues at the NEWS Lab and Wireless@VT for their kind, warm-hearted help. Their support throughout these years has made my time at Virginia Tech much more enjoyable than I ever expected.

Last but not the least, I want to express heartiest gratitude to my family members. Words fail me to express my appreciation to my parents and my brother. Without their unconditional love and persistent support, I will never be able to finish this Ph.D study. I am also greatly thankful to my dear husband, who has always been supporting me, and believing in me through up and downs. He has always been the source of my strength. It is my family's encouragement and support made this dissertation possible.

¹This research has been supported by the U.S. National Science Foundation under Grants CNS-1909372, CNS-1836802, CNS-1617896, and by the Army Research Laboratory under Grant W911NF-17-1-0021.

Contents

1	Motivation, Background, and Contributions	1
1.1	Benefits and Applications of ISDTC system Communications	3
1.1.1	ISDTC for Coverage and Capacity Enhancement	3
1.1.2	ISDTC System for Energy-Efficient Information Dissemination and Data Collection	4
1.1.3	ISDTC System-Assisted Battlefield Scenarios	5
1.2	Challenges and Related Studies in ISDTC systems	6
1.2.1	Cooperation between NTN and Terrestrial Networks	7
1.2.2	Dynamic Resilience Service Routing	8
1.2.3	Trajectory Optimization and Cooperation	9
1.2.4	Security Optimization at ISDTC Aided Military Networks	10
1.3	Limitations of Existing Works	11
1.4	Summary of Contributions	13
1.4.1	Resource Management within Integrated Satellite-Drone Network	13
1.4.2	Dynamic Resilient Service Routing	14
1.4.3	On Demand Trajectory Design and Cooperation	14
1.4.4	Security Optimization for IoBT	15
1.5	List of Publications	15
1.5.1	Journal Publications	15
1.5.2	Conference Publications	16
2	Joint Access and Backhaul Resource Management in Satellite-Drone Networks	17

2.1	Background, Related Works, and Contributions	17
2.2	System Model	18
2.2.1	Channel Model	20
2.2.2	Terrestrial Link Data Rate Analysis	22
2.2.3	Satellite Communication Link Data Rate Analysis	23
2.3	Problem Formulation and Proposed Resource Allocation Framework	24
2.3.1	Problem Formulation	24
2.3.2	Proposed Resource Allocation Algorithm	27
2.3.3	Dual Decomposition	29
2.3.4	Heavy Ball Method	32
2.3.5	Complexity and Convergence	34
2.4	Simulation Results and Analysis	35
2.5	Conclusion	39
3	Competitive Market for Joint Access and Backhaul Resource Allocation in Satellite-Drone Networks	44
3.1	Background, Related Works, and Contributions	44
3.2	System Model and problem formulation	45
3.2.1	Terrestrial Links Analysis	46
3.2.2	Satellite Communication Links Analysis	48
3.2.3	Problem Formulation	48
3.3	Proposed Solution	49
3.3.1	Satellite Sub-channel Allocation and User Association	51
3.3.2	Terrestrial Sub-channel Allocation and User Association	51
3.4	Simulation Results and Analysis	53
3.5	Conclusion	54
4	Distribution Robust-Meta Reinforcement Learning for Dynamic Resilient Contact Plan Design in Cube Satellite Networks	55
4.1	Background, Related Works, and Contributions	55

4.1.1	Related Works	55
4.1.2	Contributions	57
4.2	System Model and Problem Formulation	58
4.2.1	Network Topology with Resource Limitation	59
4.2.2	Communication Performance Analysis	60
4.2.3	Utility Function Model	62
4.2.4	Problem Formulation	62
4.3	Proposed Distribution Robust Meta Reinforcement Learning Algorithm	65
4.3.1	Distribution-Robust Meta Reinforcement Learning Components	65
4.3.2	Reinforcement Learning for Strategy Optimization	66
4.3.3	Meta Trained Learning Initialization for Task Robust Performance	68
4.3.4	Pre-training for Distribution Robust Meta Learning Performance	70
4.3.5	Convergence and complexity Analysis	73
4.4	Simulation Results and Analysis	75
4.5	Conclusion	77
5	Meta-Reinforcement Learning for Trajectory Design in Wireless UAV Networks	79
5.1	Background, Related Works, and Contributions	79
5.1.1	Related Works	79
5.1.2	Contributions	81
5.2	System model	82
5.2.1	Communication Performance Analysis	83
5.2.2	Utility Function Model	84
5.2.3	Problem Formulation	85
5.3	Proposed solutions	86
5.3.1	Value Decomposition based Reinforcement Learning Algorithm	87
5.3.2	Meta Training Procedure	92
5.4	Simulation results	95
5.5	Conclusion	102

6	Meta-Reinforcement Learning for Trajectory Design in Wireless UAV Networks	105
6.1	Background, Related Works, and Contributions	105
6.2	System Model and Problem Formulation	106
6.2.1	Communication Performance Analysis	107
6.2.2	Utility Function Model	108
6.2.3	Problem Formulation	109
6.3	Meta Gradient Policy Gradient Meta-Reinforcement Learning	109
6.3.1	MGPG Components	110
6.3.2	Policy training procedure	110
6.4	Simulation Results and Analysis	113
6.5	Conclusion	115
7	Psychological Game Theory for Secured Internet of Battlefield Things Systems	116
7.1	Background, Related Works, and Contributions	116
7.2	System Model and Problem Formulation	118
7.2.1	Soldier's communication delay	118
7.2.2	Strategies of the players	119
7.2.3	Material payoff	120
7.2.4	Game formulation	123
7.3	Dynamic Psychological Game	124
7.3.1	Psychology in the battlefield	125
7.3.2	Soldier and attacker's frustration	126
7.3.3	Dynamic Psychological game	127
7.4	Learning to Be Rational in the Psychological Game	129
7.5	Simulation Results and Analysis	131
7.6	Conclusion	137
8	Dynamic Psychological Game for Adversarial Internet of Battlefield Things Systems	139
8.1	Background, Related Works, and Contributions	139

8.2	System Model and Problem Formulation	141
8.3	Psychological Game Formulation and Solution	143
8.3.1	Psychological Game Formulation	143
8.3.2	Psychological Forward Induction Solution	146
8.4	Simulation Results and Analysis	148
8.5	Conclusion	151
9	Conclusion and Future Works	152
9.1	Summary of Contributions	152
9.2	Meta-Federated Learning for Beam Training within Satellite Drone Networks . . .	153
9.3	Trajectory design with Variable Drone Functionality	154
9.4	Secure Satellite Drone Systems from Data Leakage	154
	Appendices	156
A	Appendices for Chapter 2	157
A.1	Proof of Proposition 3	157
A.2	Proof of Lemma 3	158
A.3	Proof of Lemma 2	158
B	Appendices for Chapter 4	159
B.1	Proof of Proposition 2	159
C	Appendices for Chapter 5	161
C.1	Proof of Proposition 3	161
D	Appendices for Chapter 7	163
D.1	Proof of Theorem 1	163
D.2	Proof of Theorem 2	165
	Bibliography	181

List of Figures

2.1	Illustration of the studied integrated satellite-drone network topology.	19
2.2	Convergence of the heavy ball algorithm.	36
2.3	Convergence of heavy ball algorithm as the number of DBSs varies.	37
2.4	Price as the normalized data rate varies.	38
2.5	Transmitted data at radio access links as the number of DBSs and MBSs varies. . .	39
2.6	Average data rate as the number of DBSs varies.	40
2.7	Average data rate as the number of DBSs varies.	41
2.8	Transmitted data at radio access links as the number of DBSs and MBSs varies. . .	41
2.9	Average data rate as the altitude of DBSs varies.	42
3.1	Integrated satellite-drone network topology	45
3.2	Radio access data rate CDF vs.different algorithms.	53
3.3	Backhaul data rate CDF vs.different algorithms.	54
4.1	Topology of our considered network.	58
4.2	Illustration of the proposed T-RMRL algorithm.	71
4.3	Convergence of all considered RL algorithms.	76
4.4	Convergence of all considered meta training scheme.	77
4.5	Convergence of all considered RL algorithms as the number of unexpected user requests varies.	78
5.1	Snapshot of one user access request. DBSs only connect with active users that have not been served by other DBSs.	85
5.2	Illustration of one iteration of the VD-RL algorithm defined in Algorithm 6.	88

5.3	Illustration of the proposed meta-training method at DBS n with $J = 3$. Here, 3 different user requests are drawn from $p(\mathcal{L})$ at each training episode. $\tilde{V}_{\theta_{c,n,1}^{(1)'}}$ and $\pi_{\theta_{a,n,1}^{(1)'}}$ are the updated policy and value functions in the first sampled environment,	94
5.4	Successful service rate achieved at each DBS over time.	96
5.5	The value decomposition learned by the proposed VD-RL algorithm.	97
5.6	Value decomposition within 5 VD-RL iteration.	98
5.7	Convergence of all considered multi-agent RL algorithms.	99
5.8	Snapshots of trajectories resulting from all meta-trained and pre-trained algorithms in a sampled unseen environment.	100
5.9	Convergence of the meta-trained VD-RL.	101
5.10	Convergence of the meta-trained VD-RL algorithm as the number of DBSs varies.	102
5.11	Successful service rate as the number of DBSs varies.	103
6.1	Network topology.	107
6.2	Snapshot of trajectories resulting from all considered algorithms in a sample unseen environment.	113
6.3	Convergence of all considered algorithms.	114
6.4	CDF in terms of service success rate.	115
7.1	Soldier battlefield security graph.	118
7.2	The soldier and attacker's strategy in the battlefield.	132
7.3	The soldier and the attacker's PE strategies as the frustration weights vary.	133
7.4	Convergence of the BU strategies.	134
7.5	Convergence of the BU expected material payoffs.	134
7.6	The soldier's and the attacker's expected material payoffs as the frustration weights vary (BU results from 10 iterations in Algorithm 1).	135
7.7	The soldier's expected material payoff as the attacker's assigned weight, θ_1 , on maximizing the soldier's communication delay increases.	135
7.8	The soldier's expected material payoff as its tolerable transmission delay varies. . .	136
7.9	Frustration of the soldier and the attacker at the different steps in the battlefield. . .	137
7.10	Frustration of the soldier and the attacker resulting from BU, as time elapses. . . .	138

8.1	Soldier battlefield security graph.	141
8.2	Illustration of a soldier's movement in a 4×4 battlefield.	142
8.3	The soldier's material payoff as the number of vertices in the battlefield varies. ($X = 20, \omega_s = 0.5$).	149
8.4	The soldier's material payoff as the number of IoBT devices varies. ($v = 3, \omega_s =$ 0.5).	149
8.5	The soldier's material payoff as the weight ω of the attacker's frustration in the soldier's utility varies. ($v = 3, X = 20$).	150
8.6	The soldier and attacker's strategy. ($v = 3, X = 10$).	150
D.1	α' versus β'	165

List of Tables

2.1	List of notations	19
2.2	Simulation Parameters [1]	35
3.1	Simulation Parameters [1]	53
4.1	Simulation Parameters [2]	75
5.1	Simulation Parameters [2]	96
6.1	Simulation Parameters [3]	113
7.1	Simulation Parameters [1]	132
8.1	System parameters [1]	148

Chapter 1

Motivation, Background, and Contributions

Next-generation cellular systems, will experience an exponential increase in the number of connected physical objects due to the emergence of the Internet of Things (IoT). In essence, with IoT, people and things are able to connect to anything and anyone, at any time, using any network so as to make use of various services [4–6]. However, satisfying the unpredictable traffic demand of massive, diverse physical objects is challenging, especially with the IoT physical objects being decentralized, dynamic, and vulnerable to both cyber and physical attacks. Recently, non-terrestrial networks (NTNs) including the aerial networks carried out by unmanned aerial vehicles (UAVs), and satellite constellations have been expected to provide a feasible solution for covering such traffic demand.

In particular, within NTNs, the use of UAVs as aerial base stations for telecommunication applications, which is also popularly called DBSs, has received significant attention [7–11]. Indeed, compared to the traditional, terrestrial base stations, mobile UAVs can fly to various areas, so as to provide cost effective on-demand wireless access to the distributed ground users [12–14]. With their intrinsic features such as mobility and adjustable altitudes, DBSs can also effectively establish line-of-sight (LoS) communication links to ground users, and, thus, can admit many potential use cases in wireless networks. For instance, mobile DBSs can be deployed as wireless access points to enhance the wireless capacity and coverage at temporary events or hotspots such as sport stadiums and outdoor events. Moreover, DBSs can also disseminate information to, and collect data from IoT devices, so as to extend the IoT devices' communication range, which is usually limited due to their small transmit power. However, the performance of UAV-aided wireless systems could be restrained by the DBSs' stringent on board energy, constrained flight routes, and so on.

Meanwhile, the communication system carried out by satellites within NTNs, which are capable of provide enduring, world-wide service, has also received significant attention [15–22]. With their world-wide seamless coverage, satellite constellations are still among the only available communication systems in rural, maritime, and disaster-stricken areas. However, traditional geostationary earth orbit (GEO) satellite communication links are not practical in many communication scenarios, as they impose a large power consumption, and high transmission delays. In contrast, low earth

orbit (LEO) satellite constellations, with orbits at altitudes of less than 2000 km, have emerged as more cost-effective global wireless access points with enhanced data rates. Key examples of recent LEO satellite projects includes the recently launched commercial satellites OneWeb and SpaceX. However, works in [17, 23, 24] have also shown that limited spectrum, low efficiency and high latency are main challenges for satellite-enabled communication networks. Here, we note that, even with its enhanced data rates, a LEO satellite communication link still suffer from high transmission delay with its extremely long transmission distance.

Due to their own limitations, neither the UAVs nor the satellites within NTN can be perfectly fit into all the wireless application alone. In order to fill the coverage holes in future wireless communication networks, the cooperation among the UAV systems, satellite constellations and terrestrial communication systems must be investigated. In particular, a framework that jointly use UAVs and satellites as wireless access points must be considered. This framework shall reap the benefits of the UAV and satellite systems, and avoid their weakness. Yet doing so introduces additional, untackled challenges, such as resource management, device cooperation, data routing, and trajectory design among UAVs and LEO satellites.

Recently, the 3rd generation partnership project (3GPP) has been working on enabling the 5G wireless technology for NTNs. In particular, 3GPP has studied the standardization of the deployment scenarios, channel models, network architecture, high-layer protocols, RF requirements, as well as the use case and function identification of 5G supported NTNs [25–32]. Further, this dissertation considers the deployment of NTNs under environmental dynamics. In particular, this dissertation studies the problems of device cooperation, trajectory control, on-demand data routing within NTNs, as well as the integration of NTNs and traditional terrestrial communication systems, for on-demand service delivery to the users with dynamic and unforeseeable needs.

Moreover, the implementation of the IoT itself entails considerable security challenges, as the distributed physical objects are vulnerable to cyber and physical attacks [33, 34]. Moreover, the human operation within the IoT system could cause destructively effect on the system security, as a faulty operation can expose the system to the cyber and physical attacks. This is certainly true in the adversarial military scenarios where defenders' sophisticated psychology sometimes causes irrational behaviors.

In this chapter, we first overview the main benefits and applications of integrated satellite-drone-terrestrial-communication (ISDTC) system in Section 1.1. In Sections 1.2, we present the challenges and related studies on the ISDTC systems. Subsequently, we present the main contribution of this dissertation in Sections 1.3. In the end, the list of publications is provided in Section 1.4.

1.1 Benefits and Applications of ISDTC system Communications

The ISDTC system correspond to internet of flying DBSs, communication satellites, and physical objects connected in a IoT. In this section, we present the main advantages and potential applications of the ISDTC system in wireless networks.

1.1.1 ISDTC for Coverage and Capacity Enhancement

The IoT technology introduces an idea of extensively connecting a variety of wireless devices for diverse applications, and, thus, yields an ever-growing need for a worldwide coverage and a high speed wireless service. Nevertheless, with the geographical constraints (e.g. mountain and forest), economic considerations, and unexpected disasters, there still are many regions in this world that are suffering from poor connectivity, low quality-of-service, and even no wireless service. Example of such regions includes rural areas and countries (e.g. maritime cruises, rural national parks) where a complete terrestrial infrastructure is economically infeasible, and major public events (e.g. concert, football games), in which the surge in wireless access requests requires supplemental and substantial services. Moreover, the existing terrestrial communication infrastructure and cell phone towers can be compromised by unexpected events such as natural disasters. In such critical scenarios, there is a vital need for public safety communications between first responders and victims for search and rescue operations.

In such scenarios, a robust, fast, and capable communication system is needed to enable affordable and effective communications. One promising solution is to deploy UAVs that act as DBSs. These drone networks can be further integrated with communication satellites in order to provide both substantial and seamless global coverage. Using UAVs as radio access points is a quite natural and judicious choice, as a mobile UAV can boost the coverage and capacity of current wireless communication system, by intelligently changing their locations to provide on-demand service for specific areas that are not effectively covered by ground networks [13, 35]. Moreover, mobile DBSs can continuously move to establish public safety communications in a disaster affected areas within a minimum possible time [36, 37]. DBSs can also potentially assist in coverage and rate enhancement by providing LoS connections towards ground users [38, 39]. In particular, due to the aerial nature of UAVs, they can be effectively deployed to reduce the shadowing and blockage effects. In addition, the coverage range of UAVs can be intelligently optimized by adjusting the UAVs' altitude, in which case the UAVs can provide a maximum coverage or capacity for ground users. In summary, the optimal, on-demand deployment of an internet of flying UAVs can effectively provide fast, reliable and cost-effective connectivity to areas which are either congested (e.g. hotspots) or poorly covered by terrestrial networks. Recently, several drone communication prototypes have been presented, such as the Facebook Aquila project, Google Loon project, Nokia F-Cell project and the Eurecom Perfume project [40–43]. However, the flexible, affordable communication services provided by mobile UAVs are not endurable, as an UAV's on board energy can only support

the UAV for up to several hours.

Compared to the UAV aided networks, LEO satellite systems provide a less cost-effective but more durable solution to boost the coverage and capacity of future wireless networks [15–17, 20, 44–46]. In particular, different from the UAV aided networks, LEO satellite networks can provide ubiquitous network access services all over the world, and they are still the unique network systems in rural, sea, and disaster-affected areas where no network infrastructure can be easily deployed [15–17]. Nevertheless, the solar powered LEO satellite have extreme high endurance with their up to 10 year lifespan. Orbiting around the earth in a high altitude, the communication satellites can provide strongly reliable (in regards to coverage), and jam-resistant communication. The use of Ka-band and the the promising vision of further broadband satellite networks allow us to achieve high performance communication service from LEO satellite systems at everywhere on the earth [19, 47, 48]. For instance, Space X has recently constructed a LEO satellite internet constellation that consists of thousands of mass-produced small satellites (i.e. the starlink) to provide broadband internet access [49]. Despite the decreases on launching costs and sizes of satellites are expected to promote the development of satellite communication systems, one must cope with the problems including spectrum management, efficiency enhancement, latency reduction and data routing, before effectively implementing such a system [44].

In summary, the UAVs and satellites within NTN all have their own merits and shortcomings. Thus, an integration of these networks that can reap the benefits and avoid the weakness of UAV and satellite system can be a promising solution for enhancing the coverage and capacity of current network systems. Indeed, the integration of these networks can be both cost effective, in regards to energy consumption, and time delay, and highly endurable. The cooperation among the UAVs and satellite also boosts beyond line of sight (BLoS) operations, which can overcome LoS data links limitations, including the low coverage range, high power attenuation, and local weather influences [50]. The integration of these networks has attracted interest from academics and industry experts, several commercial integrated systems, such as the Iridium system, have emerged in recent decades [18–22].

1.1.2 ISDTC System for Energy-Efficient Information Dissemination and Data Collection

In scenarios where the cellular networks are not fully operational or are overloaded, the device-to-device (D2D) communication technologies can provide a more cost-effective approach for distributing the common files over a geographical area among a variety of terrestrial IoT devices [51–53]. For example, multiple sensors, meters, and street lights can be used to disseminate data to improve infrastructure, services, and public utilities in cities. However, the performance of a stand-alone D2D connected IoT network is limited due to the short communication range of the connected devices. In order to spread information within a large geographical area, aerial platforms embedded with UAVs and satellites can be used for information dissemination [54, 55]. The UAVs

can fly over the area and transmit the required data to a number of ground devices. The LEO satellite communication systems with seamless wide-area coverage can be leveraged to provide high data rate backhaul services to UAVs, and can also disseminate information to the UAVs' coverage holes caused by the UAVs' flight limitations and communication constraints. Therefore, information dissemination can be effectively performed by integrating mobile DBSs, LEO satellites, and ground IoT network. When integrating these networks (i.e. a ISDTC system) for information dissemination, the UAVs can transmit data to ground devices. The satellite system can provide backhaul connectivity to UAVs in rural or overloaded areas, and also provide supplemental service to devices that can not be served by terrestrial communication infrastructure or DBSs. Then, the ground D2D communication system can distribute common files within a specific area.

IoT systems are also popularly used to collect data from the environment and transfer such data to central controllers for further processing. In particular, IoT systems use sensor nodes (SNs) to sense information from the surrounding environment, and make optimal decisions without or with only minimal human intervention with their constrained resources and the dynamic of the environment for the requested IoT services [56]. However, when designing a IoT based data collection system, one must consider the energy management within data collection tasks, since battery replacement or recharging for distributed SNs is typically inconvenient, costly, and sometimes impossible. In particular, the energy-limited IoT devices may not be able to successfully report their collected data to distant terrestrial base stations. In such a case, the UAVs and satellites within NTN can be used as gathering entities to collect information from various, energy and communication capability limited IoT devices. NTN can also provide a computing hub at the edge to run several data analytic algorithms on the collected data, therefore achieving the low latency required by several critical IoT applications [20, 57]. Owing to their mobility, UAVs can flexibly move to the locations that enable LoS communications to the ground devices, thus, improving data rates and conserving energy at the less capable IoT devices. The satellite system can also supplement the UAVs to collect data in rural or overloaded areas to provide a global seamless service, at a higher price, i.e. higher energy consumption assumption, and higher transmission delay. In summary, by integrating the IoT, UAV and satellite data collection systems, a ISDTC system can introduce service that is cost-effective, flexible, reliable and seamless, and as such, it is considered among the contending enabling technologies for building networks of the future.

1.1.3 ISDTC System-Assisted Battlefield Scenarios

The emerging ISDTC system technologies also provide new ways for managing and operating a battlefield by interconnecting combat equipment, IoT devices, and other battlefield resources [58–60, 60–62]. In military scenarios where cellular networks are not fully operational, D2D communication technologies can be a great fit for providing situational awareness to the forces such that better-informed decisions can be made and implemented faster than an adversary can react. In modern battlefields, military uniforms, battlefield entities and weapon systems are often equipped with sensors capable of collecting and processing information on the status of inspected objects and environments [63, 64]. By integrating all these objects, the ISDTC system technologies can

provide situation awareness data for mission-critical battlefield operations, and, thus, contribute to cyber defense capabilities. However, the connectivity between these objects is highly vulnerable to cyber and physical attacks, given the the adversarial nature of the battlefield coupled with the limitations of the IoT devices' security mechanisms [65, 66]. The communication range is also limited by the energy capacities of the terrestrial military devices. The enhance the capability of the military network, the UAVs and satellites within NTN can be introduced.

The internet of flying UAVs is an indispensable part of such NTNs [67]. Mobile UAVs not only can provide real time situation awareness in military areas where terrestrial sensing infrastructures have not implement or overloaded, but also assure long distance communications in remote military areas. Moreover, satellite communication systems can provide a more resilient communication coverage to the adversarial military areas. Satellite communication systems are advantageous at connecting scattered military devices over large distances, and the United States Army's PM WIN-T program is currently developing a satellite-based network-centric waveform [68]. In a nutshell, by integrating the military IoT communication systems with NTNs, the ISDTC system provides an effective approach for managing and operating a battlefield.

In summary, we have presented the main advantages and potential applications of ISDTC system technologies in wireless networks. In particular, we have discussed the effective use of ISDTC system technologies, for coverage and capacity enhancement in cellular networks and public safety scenarios. Furthermore, in this section, the promising application of ISDTC system technologies for information dissemination and data collection and battlefield operations were discussed. Despite such benefits and applications of ISDTC system technologies in wireless networks, there are several technical challenges that must be also taken into account. Next, we discuss various ISDTC system-related challenges in detail.

1.2 Challenges and Related Studies in ISDTC systems

Beyond the numerous advantages of the ISDTC system, effectively utilizing ISDTC techniques requires overcoming several key challenges. The first challenge is the implementation of ISDTC system with the cooperation among terrestrial communication system and UAVs and satellites within NTNs. Indeed, multiple factors have to be considered when merging a group of UAVs, a satellite constellation, and terrestrial wireless networks. Such factors includes spectrum resource management, network modeling, cross layer power management, routing, handover management, and security guarantee. In particular, spectrum sharing and resource management among terrestrial and NTN communication links is a major challenge. This stems from the fact that, a variety of physical constraints including the energy limits of UAVs, the mobility of satellite, the interference among the nonlinear coupled terrestrial links and NTN links must be considered while preforming resource management within an ISDTC system.

The effective deployment satellite constellation is also an important challenge. Compared to the terrestrial communication systems, the satellite communication is characterized by limited con-

tact chances, dynamic system topology, high service latency (thus, slow responses), and globally seamless access. Using the slow response satellites, any changes on the service needs, service location or propagation environment can cause ineffective data delivery. Therefore, the data routing among the resource constraint satellites is only effective when it can fully utilize the limited contact chances in the system, and be resilient to environmental dynamics. Indeed, when implemented in real world, the satellites must be prepared not only for the diversity on user needs, but the fact that the user needs can change over time and location in various patterns.

Then the trajectory optimization on the UAVs is also a challenging task, especially when the UAVs are used as DBSs. Naturally, optimizing the flight path of UAVs is challenging as it requires considering many physical constraints and parameters such the energy and flight limits of UAVs. This challenge is exacerbated when UAVs are used as DBSs and must therefore optimize wireless-related performance metrics such as rate and coverage. Furthermore, solving a continuous UAVs' trajectory optimization problem is known to be analytically challenging as it involves finding an infinite number of optimization variables (i.e. UAVs' locations) with the objective of serving access requests. The cooperation among multiple UAVs make the trajectory planning problem even more challenging. This is because that the trajectory optimization problem at multiple UAVs could be extremely complex with the exponentially increased dimension of the optimization variables. Designing trajectories of a group of UAVs will be even more complex when the UAVs operate asynchronously or are dispersed into different areas which makes inter-device communication impossible. In such case, the UAVs can only be imperfectly informed with the states of the environment, which makes the trajectory problem intractable. In practical wireless environments, the problem of trajectory design also has to consider the dynamic, unforeseeable terrestrial access requests. Recently, multiple machine learning (ML) tools have been introduced for the intractable trajectory design problem in UAV aided-communication systems [69–75]. However, the machine learning solutions must perform training procedure on UAVs and, thus, introduce cost on energy, time, and hardware at UAVs. Such overhead is also an important challenge one must

Moreover, while implementing ISDTC system communication systems, a resilient security mechanism must be introduced, especially when the ISDTC system is implemented in an adversarial battlefield. Naturally, devising a security mechanism for military networks will substantially differ from conventional commercial networks due to the inherent dynamic, adversarial features of a battlefield as well as the unavoidable human intervention in the operation. In particular, the psychology of the soldiers and attackers in the adversarial battlefield environments could significantly influence their behavior, and, subsequently influence the security of the battlefield ISDTC system.

Next, we present the existing works that have addressed some of the challenges pertaining to spectrum management, data routing, trajectory design, and security optimization in ISDTC systems.

1.2.1 Cooperation between NTN and Terrestrial Networks

A fundamental analysis of the implementation of ISDTC systems is required when deploying both satellites and DBSs for coverage and capacity enhancement. In particular, the problems

of performance analysis, network modeling, cross layer power management, and spectrum resource management must be solved before the ISDN can be implemented in real life. The works in [19–22, 76–79] have studied a number of problems related to ISDNs. The authors in [19] propose a blind beam tracking approach for a Ka-band UAV-satellite communication system. The work in [21] studies the Doppler effect, the pointing error effect, and the atmospheric turbulence effect on the communication performance of a UAV-to-satellite optical communication system. In [22], the authors propose a statistical dynamic channel model to explore the three-dimensional (3D) channel tracking for a Ka-band UAV-satellite communication system. The work in [76] introduces an innovative architecture in which high altitude platform drones and UAVs are deployed to improve the visibility of satellites. The authors in [77] develop an architecture in which high altitude platform drones are connected to a satellite to enhance telecommunication capabilities. The work in [78] discuss the possible solutions and potential gains of deploying ISDTC networks in maritime environment. The work in [79] introduces a joint altitude and power control solution for the in the ISDTC system. While the authors in [20] provide an energy efficient solution to the problem of trajectory design at the satellite assisted UAVs.

Clearly, in the interest of implementing an ISDN system in real world, one must carry out effective resource allocation schemes for the ISDN assisted communication services, on top of the existing fundamental performance analysis and network modeling of the ISDN system. Indeed, under an ISDTC system, the Earth will be surrounded by massive communication satellites and aerial DBS infrastructures. To coordinate these devices for communication purposes, it is necessary to improve spectrum efficiency since the frequency bands are already congested.

1.2.2 Dynamic Resilience Service Routing

To make the data service provided by satellites accessible for daily usage, dynamic resilience routing solutions must be studied considering the dynamic topology, resource constraints, propagation performance, unforeseeable user demands, and device cooperation within the satellite system. Prior works on satellite communications [80–91] studied a number of problems related to contact plan design to route data service in resource constrained cube satellite networks. The work in [80] studies the problem of contact plan design within LEO satellite networks while jointly considering the satellites' on-board energy capacity and their stochastic solar energy infeed. The authors in [81] design contact plan within a resource constrained LEO satellite network to deliver the satellite's observation to fixed ground base station. In [82], the problem of contact selection within a multi-layer satellite-terrestrial network is investigated with the goal of improving connectivity among the resource constrained satellites. The authors in [83] optimize contact plan design, power allocation, and mission schedule in data relay satellite networks for fair communication performance on satellites. The authors in [84] investigate the suitability of the contact graph routing solution for inter-satellite communication using a Linux testbed. Despite their promising results, these existing works [80–84] do not consider the time-varying, unforeseeable service requests in real world satellite networks. Indeed, the optimization-based solutions in [80–84] are not suitable to design contact plans when the user requests are highly dynamic and unpredictable. To serve

the unforeseeable user requests, the authors in [85–91] developed ML based networking solutions. In particular, the work in [85] employs a machine learning method to predict the unforeseeable service needs within the satellite communication networks. The authors in [86] develop a centralized machine learning algorithm that enables real-time estimation on the satellite communication environment, specifically, the ever-changing rain intensity on broadband satellite communication links. Meanwhile, in [87], a deep reinforcement learning (DRL) based intelligent satellite communication configuring solution is deployed on the national aeronautics and space administration (NASA)’s space communication and navigation testbed. Yet when managing operations on every satellite, a centralized solution can cause significant communication overheads, especially for the high latency satellite systems. Thus, to intelligently classify dynamic Internet traffic flows on satellites with low overhead, the authors in [88] develop a distributed hierarchical classification solution using the deep learning technique. The work in [89] employs a distributed extreme learning machine solution to route data among LEO satellites by relying on the forecast of dynamic traffic density. In [90], the authors develop a multi agent reinforcement learning (MARL) solution that enables multiple satellites cooperatively manage the spectrum resource in the system.

Even though the machine learning based solutions developed in [85–90] are capable of accomplishing certain networking tasks with unknown service needs or unknown communication environment, they are mostly overfitted to the target tasks. In particular, the learning solutions studied in [85–90] are trained to serve specific data needs within specific communication environment. When serving unseen data needs, the learning solution must be retrained within the satellite network, which causes excessive computation costs. As the data needs change constantly in practical applications, the satellite networks must spent a great amount of time and energy on training the machine learning based networking solutions. Thus, training mechanisms that can reduce such cost for dynamic resilient operations must be studied.

1.2.3 Trajectory Optimization and Cooperation

The optimal path planning for UAVs as DBSs is an important challenge in the UAV-based communication systems. The trajectory of an UAV is significantly affected by different factors such as energy constraints, communication performance, dynamic ground users’ demands, and the cooperation among multiple UAVs. The existing literature in [8, 69–74, 92–97] studied a number of problems related to trajectory design for drone-based wireless networks. The work in [92] studies the drone trajectory optimization problem by jointly considering both the drone’s communication throughput and its energy consumption. The authors in [93] design the trajectory of a solar-powered DBS to enhance its wireless communication performance. The authors in [94] propose a dynamic trajectory control algorithm to improve the communication performance of the DBSs. The authors in [95] optimize time allocation, reflection coefficient adjustment, and DBS trajectory in backscatter communication networks. The authors in [8] propose a trajectory design that enhances physical layer security. The work in [96] studies how the 3D antenna radiation pattern and backhaul constraint affect a drone’s 3D trajectory design. Recently, there has also been significant interest in using machine learning for drone trajectory optimization for on-demand coverage to

highly unpredictable and dynamic ground wireless access requests [69, 70]. In [69], the problem of DBS trajectory optimization in an uplink non-orthogonal multiple-access (NOMA) system is studied and solved using reinforcement learning (RL). The work in [70] solves the problem of energy-efficient drone trajectory design using deep RL. The cooperation among DBSs is studied in works [71–74, 97]. In [97], the problem of trajectory design and user association in a multi-drone communication system is solved with a block coordinate descent solution. The authors in [71] study the problem of interference-aware path planning for a group of DBSs using machine learning. The work in [72] uses deep RL to design DBS trajectories so as to minimize the age of information (AoI) for sensing tasks in a multi-drone network. Meanwhile, in [73], the authors develop a MARL framework for dynamic resource allocation in wireless networks with multiple drones. The authors in [74] propose a new approach to predict the users' mobility patterns and, then, they use a multi-agent Q-learning algorithm to design trajectories for a group of DBSs.

However, the existing ML solutions, even though succeeded at solving some intractable problem, are still incapable reduce complexity of the multi-UAV trajectory design problem. Complex as they already are, these ML solutions are also overfitted their training tasks, and must spend extra energy and time on find trajectories in unseen environments. Thus, a less complex solution that fit itself into various unseen tasks must be proposed to solve trajectory design problem in real dynamic world.

1.2.4 Security Optimization at ISDTC Aided Military Networks

The optimal security mechanism design ISDTC system design is another important challenge especially in the adversarial military environments. The existing literature has studied a number of problems related to the security of the internet of battlefield things (IoBT) [58–60, 60–62]. In [58], the communications and information management challenges of the IoBT are investigated. The work in [59] integrates IoT and network centric warfare for the enhancement of the IoBT integrity. The authors in [60] use a feedback Stackelberg solution to dynamically optimize the connectivity of an adversarial IoBT network. The work in [60] develops a mean-field game approach to analyze the spread of misinformation in an adversarial IoBT. The authors in [61] propose a novel enemy localization method based on the situational awareness provided by the IoBT system. In [62], an integrated open-loop framework is proposed to quantify the information dissemination among heterogeneous IoBT devices. Moreover, there has been significant interest in studying human behavior and its cyber-psychical security impact [98–101]. Indeed, the behavioral aspect of human decision making processes, leading agents to deviate from the fully rational objective behavior in an IoBT, has a direct impact on the security of the IoBT networks. The authors in [98] study a common-pool resource game that captures the players' risk preference using tools from prospect theory. The work in [99] uses prospect theory to study the effect of a defender's and attacker's subjective behavior on the security of a drone delivery system. The work in [100] uses prospect theory to analyze the interaction between the defender of a cloud storage system and an attacker targeting the system with advanced persistent threats. In [101], a cognitive hierarchy theory based approach is proposed to capture the bounded rationality of defenders and attackers in cyber-physical systems.

These previous works present interesting and novel results. However, despite being identified as a key factor influencing the security of military networks, the role of humans in an IoBT has not been thoroughly investigated in the literature. In particular, prior to our works in this dissertation, the human behavior in military scenarios was not considered in the security optimization problems.

1.3 Limitations of Existing Works

As discussed in Section 1.1 and 1.2, in order to reap the benefits of the ISDTC system, many technical challenges such as dynamic resilient deployment of NTN, cooperation among NTN and terrestrial infrastructure, and human intervention in IoT systems. While the previous studies presented in Section 1.2 have addressed some of these challenges, they have not proposed practical solutions for the implementation of ISDTC system. In particular, current data routing and trajectory design solutions in NTN could not fit into unforeseeable, dynamic environments, the existing multi-agent solutions only can coordinate NTN devices on high computation cost, the fundamental research for ISDTC did not provide effective resource management solutions, and the impact of human intervention on security of the IoT system have not analyzed. In summary, the main limitations of the previous studies on ISDTC system aided communications are as follows:

- Despite their effort on implementing low latency satellite communication, the prior works [80–91] are still constrained by designing networking solutions for specific service requests. Indeed, the optimization-based solutions proposed in [80–84] can only solve the problem of serving known user needs while considering the constraints on the satellites' contact chances, spectrum resource, energy, and storage capacity. The ML-based solutions in [85–91] can help the satellites serve unknown user needs, but are still limited by being overfitted to their training tasks. This means that the ML solutions must be retrained for unseen service tasks. The excess computation cost introduced by repeated training of ML solution in dynamic environment can cause significant service latency within the satellite system. For such reasons, traditional solutions using techniques from optimization, and machine learning are not suitable for solving networking problem on satellites, which are responsible on serving the diverse, and dynamic data service needs around the globe. Thus, a framework that can train the satellites for all unforeseeable, and unexpected user requests must be investigated.
- The advances in UAVs, and satellites' capabilities naturally expanded their networking application domains, especially when they are integrated to enhance the coverage to the intense terrestrial access requests. For instance, the integration of UAVs and satellites have been considered with much attention. Even though some works such as [18–22, 76–79] have studied a number of problems related to ISDN systems, the spectrum sharing and resource management among terrestrial communication systems, DBS communication systems and satellite backhaul communication systems have not been studied. Indeed, the spectrum resource is still one of the most important factors affecting communication performance. To maximize the overall system performance of ISDN, one must provide a holistic framework for efficient

coordination among UAVs, satellites and the existing terrestrial networks. Yet such cooperation must require these networks to find an effective way to share spectrum from the already congested frequency bands.

- Although several works have studied the problem of trajectory design at DBSs [8, 69–74, 92–97], those works have been focused on provide coverage to a specific set of access requests with UAVs. In particular, the existing works [8, 69–74, 92–97] do not consider practical UAV-assisted wireless networks in which the ground user requests for wireless service follow unpredictable patterns. Indeed, the optimization based solutions in [92–95] are not suitable to design DBS trajectories when the user requests are unknown and unforeseeable. In addition, traditional ML solutions such as those proposed in [69–74] cannot efficiently adapt the trajectories of the DBSs to unseen environments, as they are often overfitted to their training tasks. This is because the hyper-parameters, exploration strategies, and initializations of traditional ML algorithms are manually adjusted for fitting the training tasks. Once the UAV faces an unseen task, manually adjusted ML algorithms may not converge to the optimal solution and, even if they do, the convergence speed will be very slow. Thus, existing solutions for trajectory optimization are not practical for use in wireless systems where the user request varies in dynamic, unknown pattern.
- If future networks are to deploy UAVs as flying base stations, the cooperation among a group of UAVs must be thoroughly investigated, as randomly deploy a group of UAVs could causes waste on energy and hardware as well as poor communication performance. For instance, some works such as [71–74, 97] have design trajectories for multiple UAVs with ML tools, yet these solution can not fit UAVs in real, dynamic, unforeseeable wireless environments. In particular, most of the existing ML solutions such as those in [71], [74], [102], and [103] require UAVs to share their states and actions while searching for the optimal strategies. These traditional ML solutions have high complexity, as they solve multi-agent problems by updating strategies based on the entire set of agents' actions and strategies whose dimension increases exponentially with the number of agents. Meanwhile, the ML solutions in [72] and [73] allow the agents to search strategies independently based on their own actions and states. However, using these ML solutions, the UAVs cannot optimize the overall perform of the UAV group, and thus, cannot maximize the their overall coverage to the ground users, since the UAVs are optimizing their individual utilities. In essence, one must provide a practical, low complexity, and cost effective solution to the trajectory design problem at multiple UAVs.
- The human intervention within the IoBT networks has not been studied in the existing literature. Clearly, in the military scenarios where human operation in needed, the behaviors of human operators could directly impact the functionality and security of the IoBT system. In the intensely adversarial military environments, a person will naturally have some sophisticated feelings, and such psychological behaviors can directly lead him to perform irrationally. Yet any none fully rational behavior in the battlefield could lead to serious consequences. Thus, the existing works in [58–60, 60–62], which focus on enhancing physical

devices' security mechanism, can not effectively prevent the IoBT system from poor functionality. A framework that is capable of analyzing the psychological behaviors of the human operators (i.e. the defender and attacker), must be proposed.

In summary, despite the existence of a notable body of work on ISDTC system-aided communications, this existing literature is largely limited in scope and does not provide the necessary fundamentals needed to understand the full potential of the wireless network integrating the UAV-mounted base stations, the satellite constellations, and the IoT system.

1.4 Summary of Contributions

The main contribution of this dissertation is to develop analytical foundations for deployment, and optimization of wireless ISDTC system networks. In particular, we propose various frameworks for efficient resource management, data routing, deployment, mobility, and security optimization of ISDTC system wireless communication systems. By using the proposed frameworks, the performance of the ISDTC system aided systems can be optimized in terms of coverage, capacity, reliability, and energy efficiency, while taking into consideration the unique features of NTN, such as their mobility, altitude, and coverage range. To enable these contributions, this dissertation will weave together notions from machine learning, game theory, and optimization theory. In particular, we propose novel frameworks to capture practical problems faced by the ISDTC system, and take use of the advanced mathematical tools (i.e. ML, game theory and optimization theory) to fit the ISDTC system in the complex, ever-changing, unforeseeable wireless environments. In summary, our contributions are given as follows

1.4.1 Resource Management within Integrated Satellite-Drone Network

As we discussed in Section 1.1.2, aerial communication platforms carried out by UAVs and satellites can disseminate information to IoT devices for enhanced coverage and capacity. In Chapter 2 and Chapter 3, we investigate the problem of resource management for communications in an ISDN. First, we developed a novel framework within which DBSs provide connectivity, supplementally, to ground users whose demand cannot be satisfied by terrestrial small cell base stations (SBSs). Meanwhile, a satellite system and a set of terrestrial macrocell base stations (MBSs) are used to provide resources for backhaul connectivity for both DBSs and SBSs. Then, we formulate this joint ISDN user access and resource allocation problem as a competitive market to analyze the resource management across radio access links, satellite backhaul links, and terrestrial backhaul links, while maximizing the data rates in ISDNs. Within the formulated competitive market, we consider the data demand at each user, the resource budget of each base station (BS), and the fairness among each users, BSs, and the satellite's benefits. Finally, we use the Walrasian equilibrium to solve the studied problem, and propose a distributed, iterative algorithm to find the Walrasian

equilibrium. Our analytical results show that, with well defined update step sizes, the proposed algorithm is guaranteed to reach a Walrasian equilibrium at which the total payoff gained by all the ISDN devices is maximized. Simulation results validate the theoretical results and show that the proposed algorithm yields a significant gain in terms of convergence speed and the data rates at each communication links in the system.

1.4.2 Dynamic Resilient Service Routing

In order to realize low latency service in satellite networks, we introduce a novel learning mechanism that designs contact plans within satellite networks, for maximal service coverage to users with diverse, unpredictable, and dynamic needs. In particular, we propose a framework that considers the unforeseeable user needs within the practical satellite communication systems in a dynamic team game setting, and solves the game with a dynamic resilient ML solution. Within the proposed framework, the problem of contact plan design can be addressed by reinforcing team beneficial routing designs on each data flows using decomposed team utility. Then, to implement this framework with low communication overhead and computation cost, we propose a novel, distribution-robust meta training mechanism. This mechanism can shorten the leaning procedure with prior estimation on possible data needs, and shortens the meta training procedure by estimating how the data needs change over time and location. Simulation results show that proposed algorithm allows the satellites to serve diverse user needs using only a small number of ML and meta training procedure. In fact, these works present one of the initial studies on the satellite communication system in practical unforeseeable environments with meta learning tools.

1.4.3 On Demand Trajectory Design and Cooperation

In Chapter 5 and 6, we propose novel frameworks for designing trajectories for DBSs in unpredictable, dynamic environments. In particular, we consider a practical drone-aided wireless system in which the DBSs navigate in an area, under strict energy constraints and limited information on surrounding environments, with the goal of providing uplink wireless connectivity to ground users. The DBSs can provide on-demand coverage to the ground users while adapting their trajectories to those users' unpredictable access requests. We formulate the trajectory design problem using optimization frameworks whose objective is to maximize DBSs' coverage to the dynamic ground access requests. First, we investigate the cooperation among a group of DBSs and propose a novel, distributed, VD-RL algorithm which is capable of designing trajectories for a group of DBSs, without requiring the DBSs to share their actions, states, or strategies. In particular, the DBSs can update their strategies with their own experience in the studied environment, and, thus, remarkably reduce the complexity of the high dimensional multi-agent problem. Then, we provide a novel, distributed, light-weight, meta-training mechanism that can quickly adapt the DBSs to unseen environments. The proposed meta-training mechanism prepares DBSs for unseen tasks with a VD-RL equipped with meta-trained initialization. Moreover, we proposed an online hyper pa-

parameter tuning scheme that helps the DBS adjust its trajectory for unseen service tasks, for reduced learning cost in the system. The proposed approach exhibits a smaller communications overhead and a lower computational complexity, compared with conventional ML schemes. Simulation results show that compared with the baseline ML solutions, the proposed trajectory design approach enables UAV operators to provide efficient coverage to various requests, with only a small number of experiences in the dynamic environments. In fact, this is one of initial works that study the on demand service delivery with satellite networks under environmental dynamics.

1.4.4 Security Optimization for IoBT

The use of ISDTC system is a promising solution for military communications. As mentioned in Section 1.2.4, the human intervention is a key factor impacting the security of the IoBT system. In Chapter 8 and Chapter 8, we propose security solutions that consider the psychological behavior of human decision makers in an adversarial IoBT network, in the presence of stringent resource limitations (i.e. time and power limitation) for the players. In particular, we consider a battlefield in which a soldier must accomplish a time-critical mission that requires traversing the battlefield while maintaining connectivity with the IoBT network. Within our proposed framework, the soldier, acting as a defender, will selectively connect to certain IoBT devices at each time instant along its mission path, so as to evade the attack from the attacker in the battlefield. Firstly, we investigated the case where a soldier acquires downlink information from IoBT devices. Then, we analyzed the psychology of the soldier and attacker with mathematical tools incorporated from psychological game theory [104]. In particular, the independent decision making process of the soldier and attacker is captured in an imperfect information dynamic psychological game, the soldier and attacker's intention to frustrate each other is formulated based on their estimation on each other's behavior and belief. The psychological equilibrium is utilized to solve the formulated game, and a Bayesian updating based algorithm is proposed to reach a psychological equilibrium of the game. Our results characterize the human psychology's impact on the security of the studied IoBT system. In addition, our results show that our proposed solutions lead to a significant improvement on the IoBT networks' security performance. It should be noted that, this work is the first to study the human behavior in the design of security mechanism in IoBT systems.

1.5 List of Publications

As a byproduct of the above contributions, thus far, this dissertation has led to the following key publications:

1.5.1 Journal Publications

1. **Y. Hu**, M. Chen, W. Saad, H. V. Poor, and C.S. Hong, “Distribution Robust-Meta Reinforcement Learning for Dynamic Resilient Contact Plan Design in Resource Constrained Cube Satellite Networks”, ready to submit.
2. **Y. Hu**, M. Chen, W. Saad, H. V. Poor, and S. Cui, “Distributed Multi-agent Meta Learning for Trajectory Design in Wireless Drone Networks”, *Journal on Selected Areas in Communications (JSAC) Special Issue on UAV Communications in 5G and Beyond Networks*, to appear, 2021.
3. **Y. Hu**, M. Chen, and W. Saad, “Joint Access and Backhaul Resource Management in Satellite-Drone Networks: A Competitive Market Approach”, *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 3908-3923, Jun. 2020.
4. **Y. Hu**, A. Sanjab, and W. Saad, “Dynamic Psychological Game Theory for Secure Internet of Battlefield Things (IoBT) Systems”, *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3712-3726, Apr. 2019.

1.5.2 Conference Publications

1. **Y. Hu**, M. Chen, W. Saad, H. V. Poor, and S. Cui, “Meta-reinforcement Learning for Trajectory Design in Wireless UAV Networks”, in *Proc. of Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.
2. **Y. Hu**, M. Chen, and W. Saad, “Competitive Market for Joint Access and Backhaul Resource Allocation in Satellite-Drone Networks”, in *Proc. of IFIP International Conference on New Technologies, Mobility and Security (NTMS), Mobility & Wireless Networks Track*, Canary Islands, Spain, June 2019.
3. **Y. Hu**, N. Abuzainab, and W. Saad, “Dynamic Psychological Game for adversarial Internet of Battlefield Things Systems”, in *Proc. of the IEEE International Conference on Communications (ICC), Communication and Information Systems Security Symposium*, Kansas City, MO, USA, May 2018.

Chapter 2

Joint Access and Backhaul Resource Management in Satellite-Drone Networks

As discussed in Chapter 1, the aerial communication platform carried by UAVs are not omnipotent, as the UAVs' reliability is limited by the UAVs' energy capacities. Indeed, as the UAVs can supplement ground infrastructure for cost effective radio accesses, but can never cover all the ground requests with their limited on-board energy, even with the best available trajectories they can get. This is the place where the communication satellites can lend a helping hand. Indeed, the communication satellites are capable of providing seamless coverage to every corner of this world. Yet the communication service provided by the satellites also suffer from long propagation latency, high energy consumption, slow reaction, and so on. Thus, as we discussed in Chapter 1, the integrative use of the UAVs and satellite within NTN, along with communication networks on the ground, could provide a promising approach to enhance the coverage and capacity of wireless communication systems.

2.1 Background, Related Works, and Contributions

The existing literature such as in [19] and [77, 105, 106] have studied a number of problems related to ISDNs. The authors in [107] propose a blind beam tracking approach for a Ka-band UAV-satellite communication system. The work in [105] studies the Doppler effect, the pointing error effect, and the atmospheric turbulence effect on the communication performance of a UAV-to-satellite optical communication system. The work in [106] introduces an innovative architecture in which high altitude platform drones and UAVs are deployed to improve the visibility of satellites. The authors in [77] develop an architecture in which high altitude platform drones are connected to a satellite to enhance telecommunication capabilities. However, the works in [107] and [77, 105, 106] do not consider any spectrum sharing problem in an ISDN system. Indeed, sharing the stringent spectrum resources among the communication links has a direct impact on the achievable

ISDN data rates. Hence, the resource allocation problem over all communication links must be jointly studied within the context of an ISDN.

In this chapter, we develop a novel framework that jointly considers resource allocation over satellite-DBS/SBS backhaul links, terrestrial backhaul links, and DBS/SBS-user radio access links as well as user association with DBSs and SBSs within an ISDN. In this system, the SBSs and DBSs provide downlink data service to terrestrial users, while the MBSs, supplemented with a LEO satellite system, provide backhaul service to SBSs and DBSs. We formulate this joint ISDN user access and resource allocation problem as a competitive market, in which the users request downlink data service with high data rates, while the SBSs and DBSs seek to satisfy these data service requests with minimal power consumption. Meanwhile, the SBSs and DBSs request a backhaul service with high data rates, while the MBSs and the LEO satellite seek to provide them with a backhaul service with minimal power consumption. A Walrasian equilibrium, at which the communication performance in the system are optimized, is applied to solve the formulated market. In this regard, we prove the existence of the Walrasian equilibrium, and propose a distributed, iterative algorithm to optimize the data rates at each user, SBS, DBS, MBS and the satellite, based on dual decomposition. Our analytical results show that, with well defined update step sizes, the proposed algorithm is guaranteed to reach a Walrasian equilibrium at which the total payoff gained by all the ISDN devices is maximized. The simulation results verify the outstanding convergence performance of the proposed distributed algorithm, and also show that the Walrasian equilibrium solution can effectively improve the communication performance of the studied ISDN system.

The rest of this chapter is organized as follows. The system model are described in Section 2.2. In Section 2.3, the problem formulation and proposed framework for joint backhaul and access resource management is developed and discussed. In Section 2.4, numerical simulation results are presented. Finally, conclusions are drawn in Section 2.5.

2.2 System Model

Consider a downlink ISND that consists of a set $\mathcal{S}$ of S SBSs, a set $\mathcal{E}$ of E DBSs supplementing the SBSs, a set $\mathcal{M}$ of M MBSs providing terrestrial backhaul links, a LEO satellite constellation providing a satellite backhaul system, and a set $\mathcal{U}$ of U users as shown in Fig. 6.1. In this system, each user requests data service from either an SBS or a DBS whose hover time is limited by its propulsion energy consumption, while the backhaul connections of the SBSs and DBSs are provided by either the LEO satellite or ground MBSs. Let H_n be the hover time of DBS n , defined as the time duration that a DBS can use to hover over a certain area to serve the ground users [108]. Hereinafter, we refer to the DBSs and SBSs in the network as BSs and we define $\mathcal{N}$ as the set of $N = S + E$ BSs. The 3D coordinates of the users, BSs, and MBSs are defined as $\mathbf{S}_1 = [s_{1,1}, \dots, s_{1,U}]$, $\mathbf{S}_2 = [s_{2,1}, \dots, s_{2,N}]$, and $\mathbf{S}_3 = [s_{3,1}, \dots, s_{3,M}]$, respectively. Each element $s_{i,j}$ captures the 3D coordinates of a specific location, i.e., $s_{i,j} = (x_{ij}, y_{ij}, f_{ij})$. The satellite will start at an initial location with 3D coordinate $s_0 = (x_0, y_0, f_0)$, and move in the direction of the

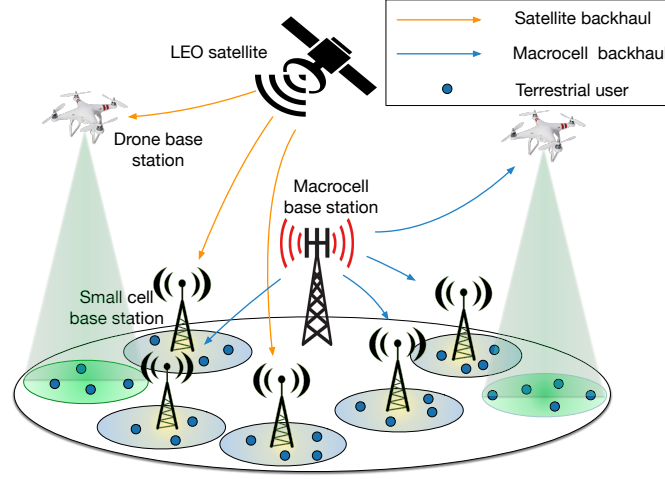


Figure 2.1: Illustration of the studied integrated satellite-drone network topology.

Table 2.1: List of notations

Notation	Description	Notation	Description
U	Number of users	$h(i, j)$	Channel gain between transmitter i and receiver j
S	Number of SBSs	$d_{i,j}$	Distance between transmitter i and receiver j
E	Number of DBSs	L_1	Rician fading channel coefficient
M	Number of MBSs	$L_2(d_{i,j})$	Large-scale channel effects between i and j
N	Number of BSs	$\kappa_{i,j}$	Offset angle of i in direction of j
H_n	Hover time of DBS n	$G_M(\kappa_{i,j})$	Transmit gain at offset angle $\kappa_{i,j}$
Δ	Studied time duration	$G_R(\kappa_{i,j})$	Receive gain at offset angle $\kappa_{i,j}$
C_u	Data service requested by user u	$\Omega_{i,j,t}$	Interference at link between i and j at slot t
T	Number of time slots within Δ	$\gamma_{i,j,t}$	SINR at link between i and j at slot t
T_n	Number of time slots within H_n	$c_{i,j,t}$	Data rate at link between i and j at slot t
τ	Duration of one time slot	Π_i	QoS requirement at device i
$\mathcal{J}^B$	Set of links between BS n and user u	$q_{i,j,t}$	Unit price of time slots at link between i and j
$\mathcal{J}^M$	Set of links between MBS m and BS n	V_i	Payoff of seller i
$\mathcal{J}^S$	Set of links between the satellite and BS n	W_j	Payoff of buyer j
ρ	Vector of allocation scheme at $\mathcal{J}^B$	L_i	Lagrangian function at device i
δ	Vector of allocation scheme at $\mathcal{J}^M$	D_i	Dual function at device i
β	Vector of allocation scheme at $\mathcal{J}^S$	λ	Vector of dual variables at $\mathcal{J}^B$
θ	Vector of request scheme at $\mathcal{J}^B$	ς	Vector of dual variables at $\mathcal{J}^M$
φ	Vector of request scheme at $\mathcal{J}^M$	ξ	Vector of dual variables at $\mathcal{J}^S$
ε	Vector of allocation scheme at $\mathcal{J}^S$	$\pi^{(k)}$	Updating step size at iteration k

x-axis with speed v .

We assume that the amount of data (in bits) that a user u requests is C_u , which has to be provided to the user within a time duration Δ . During Δ , the BSs also request backhaul data service to download the requested data from the core network. Here, we refer the amount of data that each user requests as *data service* and the amount of backhaul data that each BS requests as *backhaul service*. The time required to satisfy each user's data service request and each BS's backhaul service request depends on several factors, including the effective service time and the data rate at the corresponding user and BS. In the considered scenario, all the communication links work over the Ka band (26.5 – 40 GHz), which is a well established millimeter wave (mmWave) range suitable for satellite communication and future 5G links, as discussed in [109]. A time division multiple access (TDMA) scheme is adopted to support directional transmissions at the significantly high path loss mmW band, while maintaining low complex designs for transceivers [110]. Note that, in this model, as done in [111], the downlink interference caused by the satellite on the terrestrial links is considered to be negligible due to the satellite's limited transmit power and long-range transmission. Also, as shown in [112], the reaction time of a satellite generally small, and, thus, we do not account for this factor hereinafter.

Moreover, the considered time duration Δ is further divided, by the BSs, MBSs, and satellite into a set $\mathcal{T}$ of T time slots, each of which has a time duration of τ . Within duration Δ , each user (for radio transmission) or BS (for backhauling) will be served at least once in at least one time slot which is the *effective service time* of this user or BS. Note that, among the T time slots in the studied duration Δ , each DBS $n \in \mathcal{E}$ is only operational for $T_n = \lfloor \frac{H_n}{\tau} \rfloor$ time slots, as limited by its hover time. We also assume that each DBS is operational from the start of the studied duration, and will hover at its location for T_n time slots. We also assume that the SBSs, MBSs, and satellite will be operational during the whole process. In other words, $T_n = T$, $T_m = T$ and $T_S = T$ for all $n \in \mathcal{S}$ and $m \in \mathcal{M}$. The *data rates* at the communication links in the ISDN system are modeled next.

2.2.1 Channel Model

In the studied ISDN, the mmWave links are affected by the surrounding obstacles and, thus, these communication links can be either LoS or NLoS. We assume that the SBSs, DBSs, MBSs, and the satellite are equipped with one antenna. The channel gain between transmitter i , which could be a BS, MBS, or the satellite, and its associated user (or BS for backhaul) j is defined as [110]:

$$h(i, j) = \begin{cases} L_1 10^{\frac{-L_2(d_{i,j})}{10}} G_M(\kappa_{i,j}) G_R(\kappa_{j,i}) & \varpi_{i,j} = 1, \\ 0 & \varpi_{i,j} = 0, \end{cases} \quad (2.1)$$

where L_1 is the Rician fading channel coefficient, and $L_2(d_{i,j}) = \alpha' + \alpha \log_{10} d_{i,j} + \chi$ captures the large-scale channel effects over the mmW link between i and j [113]. Here, α is the slope of the fit and α' , the intercept parameter, is the path loss (dB) for 1 meter of distance. In addition, χ models

the deviation in fitting (dB) which is a Gaussian random variable with zero mean and variance ξ^2 for 1 meter of distance. $d_{i,j}$ is the distance between i and j . For example, if $i \in \mathcal{U}$ and $j \in \mathcal{N}$, $d_{i,j} = \sqrt{(x_{1,i} - x_{2,j})^2 + (y_{1,i} - y_{2,j})^2 + (f_{1,i} - f_{2,j})^2}$. $G_M(\kappa_{i,j})$ is the transmit gain of an antenna at an offset angle $\kappa_{i,j}$, which is defined as:

$$G_M(\kappa_{i,j}) = \begin{cases} Q_{MM}, & \text{if } \kappa_{i,j} \leq \frac{\kappa_T}{2}, \\ Q_{MS}, & \text{otherwise,} \end{cases} \quad (2.2)$$

where Q_{MM} and Q_{MS} are the transmitting antenna gains of the main lobe and the side lobe, respectively. $\frac{\kappa_T}{2}$ is the half power beam width. $\kappa_{i,j}$ is the offset angle (from the boresight direction) of i 's transmitting antenna in the direction of j 's receiver antenna. $G_R(\kappa_{j,i})$ is the receive gain of an antenna at an offset angle $\kappa_{j,i}$, which is defined as:

$$G_R(\kappa_{j,i}) = \begin{cases} Q_{RM}, & \text{if } \kappa_{j,i} \leq \frac{\kappa_R}{2}, \\ Q_{RS}, & \text{otherwise,} \end{cases} \quad (2.3)$$

where Q_{RM} and Q_{RS} are the receiving antenna gains of the main lobe and the side lobe, respectively. $\frac{\kappa_R}{2}$ is the half power beam width. $\kappa_{j,i}$ is the offset angle (from the boresight direction) of the link j 's receiving antenna in the direction of i 's transmitter antenna. $\varpi_{i,j} = 1$ indicates that the link between i and j is LoS, otherwise, it is NLoS. In fact, $\varpi_{i,j}$ is a Bernoulli random variable with probability of success $\iota_{i,j} = e^{\phi_{\text{los}} d_{i,j}}$, where ϕ_{los} is a parameter determined by the density and the average size of the blockage obstacles in the surrounding area. Note that, as done in [114], we assume that all the users have LoS paths to DBSs as the DBSs are assumed to hover at relatively high altitudes and the probability of having scatterers around DBS-user link is very small. That is, the probability of a LoS path $\iota_{i,j} = 1$ when $i \in \mathcal{E}$. Similarly, the probability of LoS path at the satellite backhaul link $\iota_{i,j} = 1$, as $j \in \mathcal{N}$ and i being the satellite.

In our model, the users and BSs will estimate the interference caused by the surrounding communication links in the system to find the the optimal association and allocation scheme that maximizes their data rates. In particular, the path loss between the interfering BSs or MBSs to the victim users or BSs is estimated based on (1), in which the bore sight angles of interfering transmitters shall be estimated. However, estimating interference in an ISDN is not trivial, as it can yield significant overhead and delay for the users and BSs, so that they can gather information on all resources allocated to neighboring links. Thus, instead of gathering such information, the bore sight angles of interfering transmitters are assumed, at each victim BS or user, to be independently and uniformly distributed in $(0, 2\pi)$. In such a case, random transmitting antenna gain $G_M(\kappa_{i,j})$ of interfering transmitters can yield an average value of $\frac{\kappa_T}{2\pi} Q_{MM} + (1 - \frac{\kappa_T}{2\pi}) Q_{MS}$. The random receiving antenna gain $G_R(\kappa_{i,j})$ of interfering transmitters can yield an average value of $\frac{\kappa_R}{2\pi} Q_{RM} + (1 - \frac{\kappa_R}{2\pi}) Q_{RS}$. Note that, the average value of transmitting antenna gain and receiving antenna gain is adopt, respectively, as the value of interfering transmitting antenna gain and victim receiving antenna gain in all the following interference analysis.

2.2.2 Terrestrial Link Data Rate Analysis

Next, we model the transmission links from BSs to users and the backhaul links from MBSs to BSs, which we call *terrestrial links* hereinafter. The link between a BS $n \in \mathcal{N}$ and a user u is denoted by $I_{n,u}^B$. The link between BS n and MBS m is denoted by $I_{n,m}^M$. The set of *transmission links* between BSs and users is defined as $\mathcal{J}^B$, while the set of *backhaul links* between MBSs and BSs is denoted by $\mathcal{J}^M$.

Let $\mathbf{a} = [a_{111}, \dots, a_{NUT}]$ be an allocation vector for the transmission links from BSs to users and let $\mathbf{w} = [w_{111}, \dots, w_{NMT}]$ be an allocation vector for the transmission links from MBSs to BSs. In vectors $\mathbf{a}$ and $\mathbf{w}$, $a_{nut} \in \{0, 1\}$ and $w_{nmt} \in \{0, 1\}$ represent the time slot allocation state at link $I_{n,u}^B$ and $I_{n,m}^M$, with $a_{nut} = 1$ and $w_{nmt} = 1$ indicating that time slot t is assigned to links $I_{n,u}^B$ and $I_{n,m}^M$, respectively. Let $\mathbf{b} = [b_{11}, \dots, b_{NU}]$ and $\mathbf{z} = [z_{11}, \dots, z_{NM}]$ be, respectively, the association vectors for the transmission links in $\mathcal{J}^B$ and backhaul links in $\mathcal{J}^M$. Here, $b_{nu} \in \{0, 1\}$ represents the association state between user u and BS n and $z_{nm} \in \{0, 1\}$ represents the association state between terrestrial MBS m and BS n , with $b_{nu} = 1$ and $z_{nm} = 1$ indicating that user u associates with BS n and BS n associates with MBS m , respectively. We use $\boldsymbol{\rho} = [\rho_{111}, \dots, \rho_{NUT}]$ and $\boldsymbol{\delta} = [\delta_{111}, \dots, \delta_{NMT}]$ to represent, respectively, the allocation schemes adopt by transmission links in $\mathcal{J}^B$ and backhaul links in $\mathcal{J}^M$, with $\rho_{nut} = a_{nut}b_{nu} = 1$ indicating that user u is served by BS n at time slot t , $\delta_{nmt} = w_{nmt}z_{nm} = 1$ indicating that BS n is served by MBS m at time slot t . Recall that each DBS $n \in \mathcal{E}$ is only operational for T_n time slots from the start of the studied period, such that $\rho_{nut} = 0$, $\delta_{nmt} = 0$, for all $u \in \mathcal{U}$, $m \in \mathcal{M}$, $n \in \mathcal{E}$, and $t > T_n$.

In this case, the interference at link $I_{n,u}^B$ that uses time slot t can be expressed as:

$$\begin{aligned} \Omega_{n,u,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) &= \sum_{i \in \mathcal{J}^B, i \neq I_{n,u}^B} \rho_{n_i u_i t} P_{n_i} h(n_i, u) + \sum_{i \in \mathcal{J}^M} \delta_{n_i m_i t} P_{m_i} h(m_i, u), \end{aligned} \quad (2.4)$$

where u_i , n_i , and m_i represent, respectively, the terrestrial user, BS, and MBS in link i . P_{n_i} is the transmit power of the BS at link i and P_{m_i} is the transmit power of the MBS at link i . Thus, the signal-to-interference-plus-noise ratio (SINR) of link $I_{n,u}^B$ is computed as:

$$\gamma_{n,u,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) = \frac{P_n h(n, u)}{\Omega_{n,u,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) + \sigma^2}, \quad (2.5)$$

where σ^2 is the noise power. The data rate at this link $I_{n,u}^B$ will be given by:

$$c_{n,u,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) = B_N \log(1 + \rho_{nut} \gamma_{n,u,t}(\boldsymbol{\rho}, \boldsymbol{\delta})), \quad (2.6)$$

where B_N is the bandwidth allocated to terrestrial link $I_{n,u}^B$, which is assumed to be equal for all terrestrial links.

The interference at backhaul link $I_{n,m}^M$ that uses time slot t is:

$$\begin{aligned} \Omega_{n,m,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) &= \sum_{i \in \mathcal{J}^B} \rho_{n_i u_{it}} P_{n_i} h(n_i, n) + \sum_{i \in \mathcal{J}^M, i \neq I_{n,m}^M} \delta_{n_i m_{it}} P_{m_i} h(m_i, n), \end{aligned} \quad (2.7)$$

Thus, the SINR at this link $I_{n,m}^M$ is given as:

$$\gamma_{n,m,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) = \frac{P_m h(m, n)}{\Omega_{n,m,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) + \sigma^2}. \quad (2.8)$$

The data rate at link $I_{n,m}^M$ is $c_{n,m,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) = B_N \log(1 + \delta_{nmt} \gamma_{n,m,t}(\boldsymbol{\rho}, \boldsymbol{\delta}))$.

2.2.3 Satellite Communication Link Data Rate Analysis

In the Ka band, a LEO satellite provides a supplemental backhaul connection to the BSs. The link between the satellite and BS n is represented by I_n^S . Let $\mathbf{u} = [u_{11}, \dots, u_{NT}]$ be the time allocation vector for the satellite backhaul links, where $u_{nt} \in \{0, 1\}$ represents the time allocation state at link I_n^S at slot t , with $u_{nt} = 1$ indicating that time slot $t \in \mathcal{T}$ is allocated to satellite backhaul link I_n^S , otherwise, we have $u_{nt} = 0$. The satellite backhaul association vector is given by $\mathbf{v} = [v_1, \dots, v_N]$, with $v_n \in \{0, 1\}$ indicating the association state between the satellite and BS n , where $v_n = 1$ indicates that BS n is associated with the satellite. Here, $\boldsymbol{\beta} = [\beta_{11}, \dots, \beta_{NT}]$ is defined as a satellite backhaul allocation vector, with $\beta_{nt} = u_{nt} v_n$ being a single allocation scheme of link I_n^S , and $\beta_{nt} = 1$ indicating that BS n is served by the satellite at time slot t . We do not build backhaul links between the satellite and MBSs, since transferring backhaul data from the satellite to MBSs and then from MBSs to BSs causes higher transmission delays, compared to directly transferring backhaul data from MBSs or the satellite to the BSs.

The interference over the backhaul link between BS n and the satellite at slot $t \in \mathcal{T}$ is:

$$\begin{aligned} \Omega_{n,S,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) &= \sum_{i \in \mathcal{J}^B} \rho_{n_i u_{it}} P_{n_i} h(n_i, n) + \sum_{i \in \mathcal{J}^M} \delta_{n_i m_{it}} P_{m_i} h(m_i, n). \end{aligned} \quad (2.9)$$

The SINR of the backhaul links between BS n and a satellite at time interval $t \in \mathcal{T}$ will be:

$$\gamma_{n,S,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) = \frac{P_S h(s, n)}{\Omega_{n,S,t}(\boldsymbol{\rho}, \boldsymbol{\delta}) + \Omega_c + \sigma^2}, \quad (2.10)$$

where P_S is the transmit power of the LEO satellite, which is assumed to be equal for all satellites in constellation. Thus, the data rate of link I_n^S is:

$$c_{n,S,t}(\boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\beta}) = B_S \log(1 + \beta_{nt} \gamma_{n,S,t}(\boldsymbol{\rho}, \boldsymbol{\delta})), \quad (2.11)$$

where B_S is the bandwidth allocated to link I_n^S , which is assumed to be equal for all satellite backhaul links.

Here, we assume that, in the studied resource stringent ISDN, the BSs and MBSs will keep serving the users or BSs during the whole studied duration Δ . This means that, at link $I_{n,u}^B$, the interference introduced by the backhaul and other transmission links will always exist, regardless of the allocation scheme used by these interfering links. In such a case, the interference in (2.4) terrestrial link $I_{n,u}^B$ can be rewritten as:

$$\Omega_{n,u,t} = \sum_{i \in \mathcal{N}/n} P_i h(i, u) + \sum_{j \in \mathcal{M}} P_j h(j, u), \quad (2.12)$$

Similarly, the interference in (2.7) at backhaul link I_{nu}^M is rewritten as:

$$\Omega_{n,m,t} = \sum_{\substack{i \in \mathcal{N} \\ i \neq n}} P_i h(i, n) + \sum_{j \in \mathcal{M}/m} P_j h(j, n). \quad (2.13)$$

Moreover, the interference in (2.9) at backhaul link I_n^S is rewritten as:

$$\Omega_{n,S,t} = \sum_{\substack{i \in \mathcal{N} \\ i \neq n}} P_i h(i, n) + \sum_{j \in \mathcal{M}} P_j h(j, n). \quad (2.14)$$

Thus, the data rates at a terrestrial link I_{nu}^B and backhaul links I_{nu}^M and I_n^S will only be a function of the allocation scheme ρ_{nut} , δ_{nmt} , and β_{nt} , respectively.

2.3 Problem Formulation and Proposed Resource Allocation Framework

2.3.1 Problem Formulation

Given the system model of Section 6.2, our goal is to find an effective resource allocation and user association scheme that maximizes the system's data rates, while satisfying the user data service request, under a limited available time resource. However, we note that the resource allocation and association scheme at one user or BS can affects the entire network performance. In particular, the optimal allocation and association scheme of any given user or BS could yield significant performance degradation at other communication links by introducing interference to these victim links. This directly leads to a decrease in the amount of data provided by these links which, consequently, cannot meet the data service demand from the users. Moreover, while optimizing its own data rate, a user could continuously occupy a BS without considering its time budget. In

order to solve the joint access and resource allocation problem among the various communication links in the resource-limited ISDN, we formulate a *competitive market* [115] which can take into account the QoS needs of every user, BS, MBS, and, also, the satellite. In such a market, the BSs, MBSs and the satellite submit quotations (or prices) for each slot of service. Meanwhile, the users and BSs find the resource allocation and association scheme that is most beneficial for them, that is, the scheme with highest data rate but lowest charged price. By using a competitive market solution, the intractable many-to-many matching problem, that is, the resource-limited joint access and allocation problem in the ISDN, can be effectively solved. The problem is solved in a distributed way to prevent excessive overhead.

In this model, the primary objective of each user is to find a resource allocation and association scheme that satisfies its data service demand and maximize its data rate, while each BS seeks to find an resource allocation scheme that can provide sufficient data service under the limited time budget. Meanwhile, each BS also seeks to find a resource allocation and association scheme that satisfies its backhaul service demand with maximized data rate. Each MBS, or the satellite, seeks to find a resource allocation scheme that can meet the backhaul service demand within their resource budget. To capture the decision making processes of the users, BSs, MBSs, and the satellite, we introduce a competitive market defined by the tuple $[\mathcal{P}, \rho, \delta, \beta, \theta, \varphi, \varepsilon, \mathcal{W}, \mathcal{V}]$ to capture the dependence between the supply and the demand of data service. Here,

- $\mathcal{P}$ is the set of users, BSs, MBSs and the satellite.
- ρ , δ , and β are the vectors of allocation schemes deployed at BSs, MBSs, and the satellite, respectively.
- $\theta = [\theta_{111}, \dots, \theta_{NUT}]$ is the request schemes deployed at users, whereby $\theta_{nut} = 1$ implies that user u requests data service from BS n at slot t , otherwise, if $\theta_{nut} = 0$, then user u does not request service from BS n at slot t . $\varphi = [\varphi_{111}, \dots, \varphi_{NMT}]$, and $\varepsilon = [\varepsilon_{11}, \dots, \varepsilon_{NT}]$ are request vectors at each BS, such that $\varphi_{nmt} = 1$ and $\varepsilon_{nt} = 1$ represent that, at slot t , BS n requests backhaul service from MBS m , and the satellite, respectively.
- $\mathcal{W}$ is the set of payoffs of the *buyers*, which are the users in the radio access links and BSs in the backhaul links.
- $\mathcal{V}$ is the set of payoffs of the *sellers*, which are the BSs in the radio access links, MBSs and the satellite in the backhaul links.

Note that, in our competitive market, user u will pay $\sum_{t \in \mathcal{T}} q_{u,n,t} \rho_{nut}$ to seller BS n , BS n will pay $\sum_{t \in \mathcal{T}} q_{n,m,t} \delta_{nmt}$ to MBS m and pay $\sum_{t \in \mathcal{T}} q_{n,S,t} \beta_{nt}$ to the satellite, as a return for the provided service. Here, $q_{i,j,t}$ is the cost of the time slot t seller j received from buyer i , noted as unit price. Note that, for each buyer, the unit prices of time slots provided by different sellers will vary, as the communication quality provided by different sellers varies. In particular, the buyer will have to pay a higher price to obtain a higher quality service as captured by the data rates.

Each buyer will use a utility function to measure its profit in the market, which is defined as the normalized difference between its communication quality (i.e. its data rate) and the price it pays to the sellers. Thus, the payoff function of user $u \in \mathcal{U}$ can will be:

$$W_u(\boldsymbol{\theta}) = \frac{1}{T\Pi_u} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} c_{n,u,t}(\boldsymbol{\theta}) - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} q_{u,n,t} \theta_{nut}. \quad (2.15)$$

The payoff function of BS $n \in \mathcal{N}$, as a buyer, is given by:

$$\begin{aligned} W_n(\boldsymbol{\varphi}, \boldsymbol{\varepsilon}) &= \frac{1}{T\Pi_n} \sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} c_{n,m,t}(\boldsymbol{\varphi}) + \frac{1}{T\Pi_n} \sum_{t \in \mathcal{T}} c_{n,S,t}(\boldsymbol{\varepsilon}) \\ &\quad - \frac{1}{T} \sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} q_{n,m,t} \varphi_{nmt} - \frac{1}{T} \sum_{t \in \mathcal{T}} q_{n,S,t} \varepsilon_{nt}. \end{aligned} \quad (2.16)$$

On the other hand, seller j will receive its payment under the cost $P_j A_j \tau$, which is the energy it consumed during the data transmission process. Here, P_j is the transmit power of executor j . A_j is the total number of time slots that seller j has allocated during the whole process. Thus, the payoff of BS $n \in \mathcal{N}$ as an seller is given by:

$$V_n(\boldsymbol{\rho}) = \frac{1}{T} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} q_{u,n,t} \rho_{nut} - \frac{1}{T} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut}. \quad (2.17)$$

Note that, the energy consumption of BS n , $P_n \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut} \tau$, is normalized by $P_n \tau T$, such that the normalized cost of BS n is $\frac{1}{T} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut}$. Similarly, the payoff of seller $m \in \mathcal{M}$ is:

$$V_m(\boldsymbol{\delta}) = \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} q_{n,m,t} \delta_{nmt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \delta_{nmt}. \quad (2.18)$$

The payoff function of the satellite, as a seller, is:

$$V_S(\boldsymbol{\beta}) = \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} q_{n,S,t} \beta_{nt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \beta_{nt}. \quad (2.19)$$

In the formulated market, each one of the buyer or seller, i.e., an user, BS, MBS, or the satellite will seek to maximize its own payoff. To solve this market, we make use of the concept of *Walrasian equilibrium* [115]. At this equilibrium, the supply of data and backhaul service from the sellers exactly matches the demand of data and backhaul service at the buyers, with each participants' payoff being maximized at the same time. The concept of a Walrasian equilibrium is defined as:

Definition 1. Given the price of the goods (i.e. the data service and the backhaul service), a *Walrasian equilibrium* in the considered ISDN competitive market is given by a tuple of vectors $[\boldsymbol{\rho}^*, \boldsymbol{\delta}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \boldsymbol{\varphi}^*, \boldsymbol{\varepsilon}^*]$ that satisfy the following conditions:

1. For radio access links, θ^* is the optimal solution that maximizes the users' payoff in (2.15) and ρ^* is the optimal solution that maximizes the payoff (2.17), of BSs.
2. For backhaul links, φ^*, ϵ^* are the solutions that optimize the BSs' payoff in (2.16), while δ^* and β^* optimize the MBSs' and the satellite's payoffs in (2.18) and (2.19), respectively.
3. The time slots provided by each seller are exactly matched to the number of time slots requested by the buyers. In other words, $\rho^* = \theta^*, \delta^* = \epsilon^*, \beta^* = \epsilon^*$.

At the Walrasian equilibrium, the communication links' optimal data rates are guaranteed and the clearance of the market is ensured by a full utilization of the time resource in the system, with which payoffs (2.15)-(2.19) in the ISDN are optimized [115]. However, improper prices can lead to the mismatch between the time slots provided by each seller and the number of time slots requested by the corresponding buyer. For example, when the prices are low, the users and BSs will be willing to request as much time slots as possible, yet the BSs, MBS, and satellite could rather not to provide any service at all. As such, a set of optimal prices should yield schemes that fully optimize the individual payoffs (2.15)-(2.19) in the system without violating any resource budget and resource demand constraints. In practical scenarios, each user, BS, MBS, or the satellite may not have the ability to analytically characterize such optimal prices by themselves. To this end, a heavy ball based algorithm, which enables a numerical computation of such prices, is proposed next, which as a result, allows computation of Walrasian equilibrium at which both the optimal payoffs at each devices and the optimal total payoff are reached.

2.3.2 Proposed Resource Allocation Algorithm

In this section, the problem of joint access and backhaul resource allocation is solved in the formulated market which considers the payoffs of all the users, BSs, MBSs, and the satellite at the same time. The constraints, including resource budget, service demand, and QoS requirements, is considered in the market. Here, the total payoff gained by all the users, BSs, MBSs and the satellite in the market is given as:

$$J(\rho, \delta, \beta, \theta, \varphi, \epsilon) = \sum_{u \in \mathcal{U}} W_u(\theta) + \sum_{n \in \mathcal{N}} (V_n(\rho) + W_n(\varphi, \epsilon)) + \sum_{m \in \mathcal{M}} V_m(\delta) + V_S(\beta). \quad (2.20)$$

As such, the joint access and backhaul resource allocation problem is formulated as:

$$\max_{\rho, \delta, \beta, \theta, \varphi, \epsilon} J(\rho, \delta, \beta, \theta, \varphi, \epsilon), \quad (2.21)$$

$$\text{s. t.} \quad \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \tilde{c}_{n,u,t}(\theta_{nut}) \tau \geq C_u, u \in \mathcal{U}, \quad (2.21a)$$

$$\sum_{u \in \mathcal{U}} \rho_{nut} + \sum_{m \in \mathcal{M}} \varphi_{nmt} + \varepsilon_{nt} \leq 1, t \in \mathcal{T}, n \in \mathcal{N}, \quad (2.21b)$$

$$\sum_{\eta=1}^t (\hat{c}_{n,t}(\boldsymbol{\rho}, \boldsymbol{\varphi}, \boldsymbol{\varepsilon})) \leq 0, t \in \mathcal{T}, n \in \mathcal{N}, \quad (2.21c)$$

$$0 \leq \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut} \leq T_n, n \in \mathcal{N}, \quad (2.21d)$$

$$0 \leq \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \delta_{nmt} \leq T, m \in \mathcal{M}, \quad (2.21e)$$

$$0 \leq \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \beta_{nt} \leq T, \quad (2.21f)$$

$$0 \leq \sum_{n \in \mathcal{N}} \theta_{nut} \leq 1, u \in \mathcal{U}, t \in \mathcal{T}, \quad (2.21g)$$

$$0 \leq \sum_{m \in \mathcal{M}} \varphi_{nmt} + \varepsilon_{nt} \leq 1, n \in \mathcal{N}, t \in \mathcal{T}, \quad (2.21h)$$

$$\sum_{n \in \mathcal{N}} \delta_{nmt} \leq 1, m \in \mathcal{M}, t \in \mathcal{T}, \quad (2.21i)$$

$$\sum_{n \in \mathcal{N}} \beta_{nt} \leq 1, t \in \mathcal{T}, \quad (2.21j)$$

$$\rho_{nut} = 0, n \in \mathcal{N}, u \in \mathcal{U}, t \in \mathcal{T}, t > T_n \quad (2.21k)$$

$$\varphi_{nmt} = 0, n \in \mathcal{N}, m \in \mathcal{M}, t \in \mathcal{T}, t > T_n \quad (2.21l)$$

$$\varepsilon_{nt} = 0, n \in \mathcal{N}, t \in \mathcal{T}, t > T_n \quad (2.21m)$$

$$\bar{c}_u(\boldsymbol{\theta}) \geq \Pi_u, u \in \mathcal{U} \quad (2.21n)$$

$$\bar{c}_n(\boldsymbol{\varphi}, \boldsymbol{\varepsilon}) \geq \Pi_n, n \in \mathcal{N} \quad (2.21o)$$

$$\rho_{nut}, \theta_{nut} \in \{0, 1\}, n \in \mathcal{N}, u \in \mathcal{U}, t \in \mathcal{T}, \quad (2.21p)$$

$$\delta_{nmt}, \varphi_{nmt} \in \{0, 1\}, m \in \mathcal{M}, n \in \mathcal{N}, t \in \mathcal{T}, \quad (2.21q)$$

$$\beta_{nt}, \varepsilon_{nt} \in \{0, 1\}, n \in \mathcal{N}, t \in \mathcal{T}, \quad (2.21r)$$

$$\boldsymbol{\rho} = \boldsymbol{\theta}, \boldsymbol{\delta} = \boldsymbol{\varphi}, \boldsymbol{\beta} = \boldsymbol{\varepsilon}, \quad (2.21s)$$

where the maximization problem in (4.10) captures the total payoff maximization objective in the market, with (4.10a) indicating that the data service demand of each user, C_u , must be satisfied. Note that, here, $\tilde{c}_{n,u,t} = B_N \log(1 + \theta_{nut} \gamma_{n,u,t})$ is user u 's perceived data rate under request scheme θ_{nut} . Meanwhile, (4.10b) indicates that one time slot cannot be simultaneously used for backhaul links and radio access links at a BS n , so as to avoid interference between the backhaul and radio access links. (4.10c) indicates that the data service each BS can provide must not exceed the capacity of the backhaul of this BS, where $\hat{c}_{n,t}(\boldsymbol{\rho}, \boldsymbol{\varphi}, \boldsymbol{\varepsilon}) = \sum_{u \in \mathcal{U}} c_{n,u,\eta}(\rho_{nut}) - \sum_{m \in \mathcal{M}} \tilde{c}_{n,m,\eta}(\varphi_{nmt}) - \tilde{c}_{n,S,\eta}(\varepsilon_{nt})$. Here, $\tilde{c}_{n,m,t}(\varphi_{nmt}) = B_N \log(1 + \varphi_{nmt} \gamma_{n,m,t})$ and $\tilde{c}_{n,S,t}(\varepsilon_{nt}) = B_N \log(1 + \varepsilon_{nt} \gamma_{n,S,t})$ are BS n 's perceived data rates under request schemes φ_{nmt} , and ε_{nt} , respectively. (4.10d)-(4.10f) indicate that BSs, MBSs, and the satellite must operate under time limitations. (4.10g) and (4.10h) capture the fact that each user or BS can only be serviced at most once in each studied time slot. (4.10i) and (4.10j) captures the fact that, each BS or MBS will serve at most one terrestrial user or BS at each time slot within its effective time. (4.10k)-(4.10m)

capture the fact that each BS can neither receive nor transmit data out of its effective time. (4.10n) and (4.10o) capture the QoS requirements in the ISDN. Here, the QoS requirements at each user in captured in the form of $\bar{c}_u(\boldsymbol{\rho}) = \frac{1}{T} \sum_{t \in \mathcal{T}} \sum_{n \in \mathcal{N}} \tilde{c}_{n,u,t}(\theta_{nut}) \geq \Pi_u$, with $\bar{c}_u(\theta_{nut})$ being user u 's perceived average data rate over the studied duration, Π_u being the minimum rate threshold required by user u . In other words, the average data rate that user u can achieve over the studied duration should never be less than Π_u . Meanwhile, each BS n 's perceived average backhaul data rate should not be less than threshold Π_n , that is, $\bar{c}_n(\boldsymbol{\delta}, \boldsymbol{\beta}) = \frac{1}{T} \sum_{t \in \mathcal{T}} \left(\sum_{m \in \mathcal{M}} \tilde{c}_{n,m,t}(\varphi_{nmt}) + \tilde{c}_{n,S,t}(\varepsilon_{nt}) \right) \geq \Pi_n$. (4.10p)-(4.10r) indicate that (4.10) is a binary problem. (4.10s) captures the market clearance constraints, with which the data service provided by each BS, MBS, and the satellite is consumed by the users and BSs. Notice that, under the market clearance constraint (4.10s), the payments from all buyers are returned to the sellers. Hence, the objective function can be simplified as:

$$\begin{aligned}
& J(\boldsymbol{\rho}, \boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\varepsilon}) \\
&= \frac{1}{T \Pi_u} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \tilde{c}_{n,u,t}(\theta_{nut}) - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut} \\
&+ \sum_{n \in \mathcal{N}} \frac{1}{T \Pi_n} \left[\sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} \tilde{c}_{n,m,t}(\varphi_{nmt}) + \sum_{t \in \mathcal{T}} \tilde{c}_{n,S,t}(\varepsilon_{nt}) \right] \\
&- \frac{1}{T} \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \delta_{nmt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \beta_{nt}. \tag{2.22}
\end{aligned}$$

2.3.3 Dual Decomposition

The integer non-linear programming (INLP) optimization problem in (4.10) can be solved using a dual decomposition method. This stems from the fact that constraints (4.10a)-(4.10r) are local constraints at different users, BSs, MBSs, and the satellite, which means that each of these constraints will only set boundaries to one user, BS, MBS or the satellite. For example, constraints (4.10a), (4.10g), and (4.10n) are local constraints for the terrestrial users, (4.10b)-(4.10d), (4.10h), (4.10k)-(4.10m), and (4.10o) are local constraints at the BSs, (4.10e), (4.10i) are local constraints at MBSs, and (4.10f), (4.10j) are local constraints at the satellite. (4.10p)-(4.10r) are local constraints at different user, BS, MBS, and the satellite, indicating that the allocation schemes are binaries. Note that, hereinafter, constraints (4.10a)-(4.10r) are called local constraints. Note that, solving (4.10) with these local constraints introduces problems including increased amount of information exchanged in the ISDN and also the potential privacy and security crisis. In such a case, the Lagrangian function of the objective function in (4.10) is formulated without considering

(4.10s) as:

$$\begin{aligned}
L = & \frac{1}{T\Pi_u} \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} c_{n,u,t}(\theta_{nut}) - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{nut} \\
& + \sum_{n \in \mathcal{N}} \frac{1}{T\Pi_n} \left[\sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} c_{n,m,t}(\varphi_{nmt}) + \sum_{t \in \mathcal{T}} c_{n,S,t}(\varepsilon_{nt}) \right] \\
& - \frac{1}{T} \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \delta_{nmt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \beta_{nt} \\
& + \sum_{n \in \mathcal{N}} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \lambda_{n,u,t}(\rho_{nut} - \theta_{nut}) \\
& + \sum_{n \in \mathcal{N}} \sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} \varsigma_{n,m,t}(\delta_{nmt} - \varphi_{nmt}) + \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \xi_{n,t}(\beta_{nt} - \varepsilon_{nt}). \tag{2.23}
\end{aligned}$$

Here, $\lambda_{n,u,t}$, $\varsigma_{n,u,t}$, and $\xi_{n,t}$ are the dual variables.

Meanwhile, since the data rate functions $c_{n,u,t}(\theta_{nut})$, $c_{n,m,t}(\varphi_{nmt})$, and $c_{n,S,t}(\varepsilon_{nt})$ are independent, one can obtain the augmented Lagrangian function transformed into a form $L = \sum_{u \in \mathcal{U}} L_u + \sum_{n \in \mathcal{N}} L_n +$

$\sum_{m \in \mathcal{M}} L_m + L_S$. Here, L_u is the Lagrangian function at user u , such that user $u \in \mathcal{U}$ is obligated to solve a maximization problem in the form:

$$\begin{aligned}
D_u(\boldsymbol{\theta}_u, \boldsymbol{\lambda}_u) &= \max_{\boldsymbol{\theta}_u \in \mathcal{X}_u} L_u(\boldsymbol{\theta}_u, \boldsymbol{\lambda}_u) \\
&= \max_{\boldsymbol{\theta}_u} \left\{ \frac{1}{T\Pi_u} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} c_{n,u,t}(\theta_{nut}) - \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \lambda_{n,u,t} \theta_{nut} \right\}, \tag{2.24}
\end{aligned}$$

where $\mathcal{X}_u = \{\boldsymbol{\theta}_u : (4.10a), (4.10g), (4.10n), \theta_{nut} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{T}\}$ includes the local constraints at user u . Here, $\boldsymbol{\theta}_u = [\theta_{1u1}, \dots, \theta_{1uT}, \dots, \theta_{Nu1}, \dots, \theta_{NuT}]$ is the vector of data request schemes of user u , with $\theta_{nut} \in \{0, 1\}$. Note that, $\theta_{nut} = 1$ implies that user u requests data service from BS n at time slot t . $\boldsymbol{\lambda}_u = [\lambda_{1u1}, \dots, \lambda_{1uT}, \dots, \lambda_{Nu1}, \dots, \lambda_{NuT}]$ is the vector of dual variables related to user u . Also, note that $D_u(\boldsymbol{\theta}_u, \boldsymbol{\lambda}_u)$ is the dual function at user u . L_n is the Lagrangian function at BS n , such that the optimization problem at BS n is defined as follows:

$$\begin{aligned}
D_n(\boldsymbol{\varphi}_n, \boldsymbol{\varepsilon}_n, \boldsymbol{\rho}_n, \boldsymbol{\lambda}_n, \boldsymbol{\xi}_n, \varsigma_n) &= \max_{\boldsymbol{\varphi}_n, \boldsymbol{\varepsilon}_n \in \mathcal{X}_n} L_n(\boldsymbol{\varphi}_n, \boldsymbol{\varepsilon}_n, \boldsymbol{\rho}_n, \boldsymbol{\lambda}_n, \boldsymbol{\xi}_n, \varsigma_n) \\
&= \max_{\boldsymbol{\varphi}_n, \boldsymbol{\varepsilon}_n \in \mathcal{X}_n} \frac{1}{T\Pi_n} \sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} c_{n,m,t}(\varphi_{n,m,t}) \\
&+ \frac{1}{T\Pi_n} \sum_{t \in \mathcal{T}} c_{n,S,t}(\varepsilon_{nt}) - \sum_{m \in \mathcal{M}} \sum_{t \in \mathcal{T}} \varsigma_{n,m,t} \varphi_{nmt} \\
&- \sum_{t \in \mathcal{T}} \xi_{n,S,t} \varepsilon_{nt} + \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \lambda_{n,u,t} \rho_{nut} - \frac{1}{T} \sum_{u \in \mathcal{U}} \sum_{t \in \mathcal{T}} \rho_{n,u,t}, \tag{2.25}
\end{aligned}$$

where $D_n(\boldsymbol{\varphi}_n, \boldsymbol{\varepsilon}_n, \boldsymbol{\rho}_n, \boldsymbol{\lambda}_n, \boldsymbol{\xi}_n, \varsigma_n)$ is called the dual function at BS n . Moreover, note that $\mathcal{X}_n =$

$\{\varphi_n, \varepsilon_n, \rho_n : (4.10b) - (4.10d), (4.10h), (4.10k) - (4.10m), (4.10o) - (4.10r), \forall m \in \mathcal{M}, \mu \in \mathcal{U}, t \in \mathcal{T}\}$ includes all the local constraints at BS n . Here, $\varphi_n = [\varphi_{n11}, \dots, \varphi_{n1T}, \dots, \varphi_{nM1}, \dots, \varphi_{nMT}]$, and $\varepsilon_n = [\varepsilon_{n1}, \dots, \varepsilon_{nT}]$ are vectors of the data request schemes at BS n , with $\varphi_{nmt} = 1$ indicating that BS n requests data service from BS m , $\varepsilon_{nt} = 1$ indicating that BS n requests data service from the satellite, at time slot t . $\rho_n = [\rho_{n11}, \dots, \rho_{n1T}, \dots, \rho_{nU1}, \dots, \rho_{nUT}]$ is the allocation scheme vector at BS n in the studied duration. $\lambda_n = [\lambda_{n11}, \dots, \lambda_{n1T}, \dots, \lambda_{nU1}, \dots, \lambda_{nUT}]$, $\varsigma_n = [\varsigma_{n11}, \dots, \varsigma_{n1T}, \dots, \varsigma_{nU1}, \dots, \varsigma_{nUT}]$, and $\xi_n = [\xi_{n1}, \dots, \xi_{nT}]$ are the vectors of dual variables related to BS n . Note that, instead of calculating the data rate it can receive from the satellite at slot t by tracking the location of the satellite, a BS n can estimate the data rate based on Proposition 3.

Proposition 1. Given an allocation vector β , the variation of backhaul data rate $c_{n,S,t}$ caused by the movement of the LEO satellite is:

$$\tilde{c}_{n,S,t} \approx \frac{-\alpha \log_2 10 \beta_{nt} B_S P_S L_1 Q_{MM} Q_{RM} \Gamma (v^2 t + (x_0 - x_{2,n}) v) \tau^2}{10 \Omega_{n,S,t} d_{n,S}^2(t) (1 + \Gamma)}, \quad (2.26)$$

where $\Gamma = 10^{-\frac{0.1\alpha' + 0.1\alpha \log_{10} d_{n,S}(t) + 0.1\chi}{10}}$

Proof. See Appendix A.1. □

As such, using Proposition 1, the BSs can evaluate the satellite backhaul data rates at slot t in the form of $c_{n,S,t-1} + \tilde{c}_{n,S,t}$, without introducing unnecessary overhead while tracking the location of the LEO satellite. Also, the satellite dynamically updates the price of its backhaul data service along its orbit, such that the resource allocation scheme keeps being updated for maximizing communication performance and, hence, reducing instability. The Lagrangian function at MBS m is L_m . The optimization problem at MBS m is:

$$\begin{aligned} D_m(\delta_m, \varsigma_m) &= \max_{\delta_m \in \mathcal{X}_m} L_m(\delta_m, \varsigma_m) \\ &= \max_{\delta_m \in \mathcal{X}_m} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \varsigma_{n,m,t} \delta_{nmt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \delta_{n,m,t}, \end{aligned} \quad (2.27)$$

where $\mathcal{X}_m = \{\delta_m : (4.10e), (4.10i), \delta_{nmt} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{T}\}$ includes all the local constraints at MBS m . $\delta_m = [\delta_{1m1}, \dots, \delta_{1mT}, \dots, \delta_{Nm1}, \dots, \delta_{NmT}]$ is the allocation scheme vector at MBS m . $\varsigma_m = [\varsigma_{1m1}, \dots, \varsigma_{1mT}, \dots, \varsigma_{Nm1}, \dots, \varsigma_{NmT}]$ is the vector of dual variables related to MBS m . Here, $D_m(\delta_m, \varsigma_m)$ is the dual function at MBS m . The Lagrangian function at the satellite is L_S , which is maximized as follows:

$$\begin{aligned} D_S(\beta, \xi) &= \max_{\beta \in \mathcal{X}_S} L_S(\beta, \xi) \\ &= \max_{\beta \in \mathcal{X}_S} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \xi_{n,S,t} \beta_{nt} - \frac{1}{T} \sum_{n \in \mathcal{N}} \sum_{t \in \mathcal{T}} \beta_{nt}, \end{aligned} \quad (2.28)$$

where $\mathcal{X}_S = \{\beta : (4.10f), (4.10j), \beta_{nt} \in \{0, 1\}, \forall n \in \mathcal{N}, t \in \mathcal{T}\}$ includes all the local constraints at the satellite. $\xi = [\xi_{11}, \dots, \xi_{1T}, \dots, \xi_{N1}, \dots, \xi_{NT}]$ is the vector of dual variables related

to the satellite. Here, $D_S(\beta, \xi)$ is the dual function at the satellite. Clearly, these Lagrangian functions are actually the payoff functions at each of the user, BS, MBS, and the satellite, with dual variables $\lambda_{n,u,t}$, $\varsigma_{n,m,t}$, and $\xi_{n,s,t}$ functioning as normalized price $\frac{1}{T}p_{n,u,t}$, $\frac{1}{T}p_{n,m,t}$, and $\frac{1}{T}p_{n,s,t}$. The vectors of dual variables are noted as $\lambda = [\lambda_{1,1,1}, \dots, \lambda_{N,U,T}]$, $\varsigma = [\varsigma_{1,1,1}, \dots, \varsigma_{N,M,T}]$, and $\xi = [\xi_{1,S,1}, \dots, \xi_{N,S,T}]$, respectively.

Lemma 1. There exist at least one Walrasian equilibrium in the ISDN system, if:

1. The relaxed problem of (21) is feasible and admits at least one integer solution.
2. The value of the data rates studied in constraints of (21) are rational numbers.

Proof. See Appendix A.2. □

In summary, given the local constraints, (4.10a)-(4.10r), the feasible region of the INLP problem in (4.10) can be separated into several regions, i.e. $\mathcal{X}_u$, $\mathcal{X}_n$, $\mathcal{X}_m$, and $\mathcal{X}_S$, each of which serves as the domain of the optimization problem at one user, BS, MBS or the satellite. Precisely, the optimization problem at one user, BS, MBS or the satellite is the maximizing the value of its Lagrangian function, or its payoff function. The resulting maximal value of payoff at one user, BS, MBS or the satellite, with given dual variables, is defined as the dual function at this user, BS, MBS or the satellite. The dual of the INLP problem in (4.10), $D = \sum_{u \in \mathcal{U}} D_u + \sum_{n \in \mathcal{N}} D_n + \sum_{m \in \mathcal{M}} D_m + D_S$, is formulated as the sum of maximal payoffs at all the users, BSs, MBSs, and the satellite in the system with given dual variables λ , ς , and ξ . The dual of (4.10) is then minimized at each user, BS, MBS, and the satellite over dual variables λ , ς , and ξ , with the heavy ball method [116]. With the convergence of the heavy ball algorithm, the market clearance constraints is satisfied, which gives rise to a Walrasian equilibrium of the ISDN system, as the maximal values for the payoffs of each user, BS, MBS, and the satellite are reached.

2.3.4 Heavy Ball Method

The heavy ball method [116] is an efficient sub-gradient method that seeks to find the optimal dual variables that minimizes the dual function, D , by iteratively updating the dual variables λ , ς , and ξ in the opposite direction to the gradient $\nabla L(\lambda)$, $\nabla L(\varsigma)$, and $\nabla L(\xi)$, respectively. The dual function D is, firstly, computed with initial dual variables $\lambda^{(0)}$, $\varsigma^{(0)}$ and $\xi^{(0)}$. However, the allocation schemes that maximize payoff functions with such initial dual variables, or normalized prices, may not satisfy the market clearance constraints, that is, the data and backhaul service supply may not match the data and backhaul service demand. The mismatch between supply and demand is referred as to violation of the market clearance constraints, which is:

$$\mathbf{s}_1^{(k)} = \boldsymbol{\theta}^{(k)} - \boldsymbol{\rho}^{(k)}, \quad \mathbf{s}_2^{(k)} = \boldsymbol{\varphi}^{(k)} - \boldsymbol{\delta}^{(k)}, \quad \mathbf{s}_3^{(k)} = \boldsymbol{\varepsilon}^{(k)} - \boldsymbol{\beta}^{(k)}, \quad (2.29)$$

where $\mathbf{s}_1^{(k)}$, $\mathbf{s}_2^{(k)}$, and $\mathbf{s}_3^{(k)}$ are vectors of market clearance violations at radio access links, terrestrial backhaul links, and satellite backhaul links at iteration k of the heavy ball dual variable

updating process, respectively. $\rho^{(k)}$, $\delta^{(k)}$, and $\beta^{(k)}$ are the optimal allocation vectors that maximize Lagrangian functions L_n , L_m , and L_S at each BS, MBS, and the satellite in the ISDN, at iteration k . $\theta^{(k)}$, $\varsigma^{(k)}$, and $\xi^{(k)}$ are the optimal request vectors that maximize Lagrangian functions L_u and L_n , for all the users and BSs in the system, at iteration k . As these market clearance violations are considered to reach 0 at the Walrasian equilibrium of the ISDN, dual variables are updated iteratively at the BSs, MBSs, and the satellite for the decreased violation values. Precisely, in the studied ISDN, each user and BS report their optimal service request scheme $\theta^{(k)}$, $\varsigma^{(k)}$, and $\xi^{(k)}$ to their associated BS, MBS, and satellite, at each iteration k . The BSs will be responsible for updating prices λ based on the market clearance violation $s_1^{(k)}$ derived from (2.29), such that $\lambda^{(k+1)} = \lambda^{(k)} + \pi_1^{(k)} \mu_1^{(k)}$, where $\mu_1^{(k)} = \frac{s_1^{(k)}}{\|s_1^{(k)}\|} + \nu_1^{(k)} \mu_1^{(k-1)}$ with

$$\nu_1^{(k)} = \max \left\{ 0, -1.5 \frac{s_1^{(k)} (\mu_1^{(k-1)})^T}{\|s_1^{(k)}\| \|\mu_1^{(k-1)}\|} \right\}. \text{ MBSs are responsible for updating prices } \varsigma, \text{ based on the}$$

market clearance violation $s_2^{(k)}$ derived from (2.29), with $\varsigma^{(k+1)} = \varsigma^{(k)} + \pi_2^{(k)} \mu_2^{(k)}$, where $\mu_2^{(k)} = \frac{s_2^{(k)}}{\|s_2^{(k)}\|} + \nu_2^{(k)} \mu_2^{(k-1)}$ with $\nu_2^{(k)} = \max \left\{ 0, -1.5 \frac{s_2^{(k)} (\mu_2^{(k-1)})^T}{\|s_2^{(k)}\| \|\mu_2^{(k-1)}\|} \right\}$. The satellite, at the mean time, is

obligated to update prices ξ based on $\xi^{(k+1)} = \xi^{(k)} + \pi_3^{(k)} \mu_3^{(k)}$, where $\mu_3^{(k)} = \frac{s_3^{(k)}}{\|s_3^{(k)}\|} + \nu_3^{(k)} \mu_3^{(k-1)}$

with $\nu_3^{(k)} = \max \left\{ 0, -1.5 \frac{s_3^{(k)} (\mu_3^{(k-1)})^T}{\|s_3^{(k)}\| \|\mu_3^{(k-1)}\|} \right\}$. Here, $\lambda^{(k)}$, $\varsigma^{(k)}$ and $\xi^{(k)}$ are vectors of the dual variables

at the k -th iteration, $\pi_1^{(k)}$, $\pi_2^{(k)}$, and $\pi_3^{(k)}$ are the updating step sizes in the price updating process.

Note that, for guaranteed convergence, the step sizes $\pi_1^{(k)}$, $\pi_2^{(k)}$, and $\pi_3^{(k)}$ are set to be, respectively, $\frac{\lambda^* - \lambda^{(t)}}{\|\mu_1^{(k)}\|^2}$, $\frac{\varsigma^* - \varsigma^{(t)}}{\|\mu_2^{(k)}\|^2}$, and $\frac{\xi^* - \xi^{(t)}}{\|\mu_3^{(k)}\|^2}$ with λ^* , ς^* , and ξ^* being optimal dual variables. Also, we select the

values of the parameters ν_1 , ν_2 , ν_3 , μ_1 , μ_2 , and μ_3 based on the so-called modified gradient step direction technique proposed in [117] for a faster convergence. Then, the terrestrial users, BSs, MBSs, and the satellite are informed with the updated prices $\lambda^{(k)}$, $\varsigma^{(k)}$ and $\xi^{(k)}$ and will find the optimal solutions of their Lagrangian function maximization problems, $\theta^{(k+1)}$, $\varsigma^{(k+1)}$, $\xi^{(k+1)}$, $\rho^{(k)}$, $\delta^{(k)}$, and $\beta^{(k)}$ with the updated dual variables. If these new optimal solutions still do not satisfy the market clearance constraints, the dual variables will keep being updated. The proposed heavy ball based solution is shown in Algorithm 1. At the beginning of the algorithm, each user, BS, and the satellite will choose an association (i.e. $\theta^{(0)}$, $\varphi^{(0)}$, or $\epsilon^{(0)}$) or allocation scheme (i.e. $\rho^{(0)}$, $\delta^{(0)}$, or $\beta^{(0)}$) for maximizing their profit under randomly chosen initial dual variables (i.e. prices), $\lambda^{(0)}$, $\varsigma^{(0)}$, and $\xi^{(0)}$. Then, these association schemes are sent to the BSs at the radio access links, to MBSs and the satellite at the backhaul links. The BSs, MBSs, and the satellite will update the prices based on the uploaded association schemes and broadcast the price vectors to the users and BSs. As stated next, a Walrasian equilibrium is guaranteed at the convergence of the heavy ball algorithm.

Lemma 2. For a feasible payoff maximization problem (21), the proposed dual decomposition-heavy ball based algorithm will converge to the Walrasian equilibrium, at which the optimal total

payoff is achieved by the system, with diminishing step sizes.

Proof. See Appendix A.3. □

2.3.5 Complexity and Convergence

At each iteration of the proposed algorithm, the decision variables are updated at all the devices, including the users, BSs, MBSs and the satellite. In particular, we assume that the dual optimization problems at each user u , BS n , MBS m and the satellite will converge at iteration $\vartheta_u, \vartheta_n, \vartheta_m$, and ϑ_s , respectively. Then, the complexities of the the decision variable update procedure at each user u , BS n , MBS m and the satellite can be expressed as, respectively, $\mathcal{O}(\vartheta_u)$, $\mathcal{O}(\vartheta_n)$, $\mathcal{O}(\vartheta_m)$, and $\mathcal{O}(\vartheta_s)$. Note that, the linear programming processes at each device are executed at the same time, and, therefore, the whole decision variable update process will have a complexity of $\mathcal{O}(\vartheta^*)$, where ϑ^* is the highest value between $\vartheta_u, \vartheta_n, \vartheta_m$, and ϑ_s . Moreover, the update procedure for dual variables is also executed at devices, including the BSs, MBSs and the satellite, simultaneously. The update procedure at each device has a linear runtime which grows in proportion with the number of dual variable vectors being updated at this device. In particular, dual vectors λ_{nu} with $u \in \mathcal{U}$, are updated at BS n sequentially and yield a complexity of $\mathcal{O}(U)$, dual vectors ς_{nm} , and ς_{nm} with $n \in \mathcal{N}$, are updated sequentially at MBS m and the satellite, respectively, with a complexity of $\mathcal{O}(N)$. In such a case, the whole dual variable update process has a complexity of $\mathcal{O}(U)$, as we are assuming $U > N$. In summary, the complexity of the proposed algorithm is $\mathcal{O}(K^*(U + \vartheta^*))$, assuming that the proposed heavy ball algorithm converges at iteration K^* . This is a reasonable complexity because ϑ^* and K^* are considerably small numbers. As a comparison, the centralized optimization solution of (21) has linear runtime that will grow with the network size, as it must be implemented at a single centralized controller which sequentially allocates the resources to various links. In essence, the complexity of the centralized algorithm is $\mathcal{O}(UN(M + 1))$.

Moreover, the centralized controller of the centralized optimization algorithm will compute and, then, transmit the decision variables to the corresponding devices. In essence, $2NU$ decision variable vector are transmitted to the users and BSs on the radio access links, $2NM + 2N$ decision variable vectors are transmitted to the BSs, MBSs, and the satellite on the backhaul links. However, in a dynamic, multi-tier ISDN system, the locations of DBSs and the satellites vary with time, in which case the centralized algorithm must be executed repeatedly. Assuming that the centralized algorithm is executed K times during the studied period, a total of $2KN(U + M + 1)$ decision variable vectors will be transmitted to the devices using the centralized algorithm. In contrast, the proposed distributed algorithm takes care of the mobility issue in ISDN by dynamically update the price of data service at DBSs and the satellite during the updating procedure. At each iteration of the proposed algorithm, the BSs, MBSs and the satellite will update and broadcast a total of $N + M + 1$ vectors of dual variables to the users and BSs, which, in turn, will update and transmit a total number of $U + N$ vectors of decision variables to the BSs, MBS, and the satellite. The number of vectors being exchanged in the network will be $K^*(U + 2N + M + 1)$ at the convergence of the proposed algorithm. As one can see, the repeated implemented centralized algorithm yields a

Table 2.2: Simulation Parameters [1]

Parameter	Value	Parameter	Value
P_n	20 dBm	P_m	43 dBm
P_S	9.23 dBW	B_N	56 MHz
σ^2	-104 dB	Ω_c	10.5354 dB

much heavier overhead, which directly makes itself not applicable to the dynamic, multi-tier ISDN system.

Also, a centralized approach is rather impractical in an ISDN because of the presence of terrestrial, aerial, and satellite nodes. In such a system, implementing a centralized solution is rather infeasible, as it will require communication across terrestrial and aerial links. In addition, it is not clear where the centralized solution can be implemented. For instance, implementing it at a ground node will make it difficult to control the satellite. Meanwhile, implementing it at a DBS will make it difficult to maintain performance because of the mobility of DBSs and their need to hover. In general, for an ISDN, employing a centralized node is not desirable.

2.4 Simulation Results and Analysis

For our simulations, we consider a scenario with $U = 60$ mobile users, $N = 10$ BSs (including 5 SBSs and 5 DBSs), $M = 1$ macro base station, and 1 LEO satellite constellation. Note that, the height of the DBSs is assumed to be 200 meters. The users, SBSs, DBSs and MBSs are repeatedly deployed at some random locations that are uniformly distributed in an area with a diameter of 1000 meters. Other parameters used in the simulations are listed in Table II. The heavy ball based Walrasian equilibrium results are compared to the centralized optimization based results, the sub-gradient based results, and a random allocation scheme results considering the market clearance. Here, with the centralized optimization algorithm [118], the optimization problem formed in (21) is solved at one MBS and the optimal solutions are transmitted to all of the other nodes (users, BSs, MBSs, and the satellite) in the network. The MATLAB optimization toolbox is used to implement the centralized algorithm. The sub-gradient algorithm represents the primal-dual sub-gradient algorithm in [116]. The settings of the sub-gradient algorithm are similar to the heavy ball algorithm, although any other approach is possible. However, in the sub-gradient algorithm, the dual variables are directly updated in the direction of the current negative sub-gradient. With the random allocation scheme, the time resources are randomly allocated to the users and BSs. 2000 independent runs are used (for averaging) in order to generate the statistical results.

Fig. 2.2 shows the convergence of the proposed heavy ball algorithm. In the results shown in Fig. 2.2, the heavy ball algorithm requires approximately 200 iterations to reach convergence, which is 10% less than the number of iterations required for convergence of the sub-gradient algorithm. This stems from the fact that, the updating direction of the heavy ball algorithm has a smaller

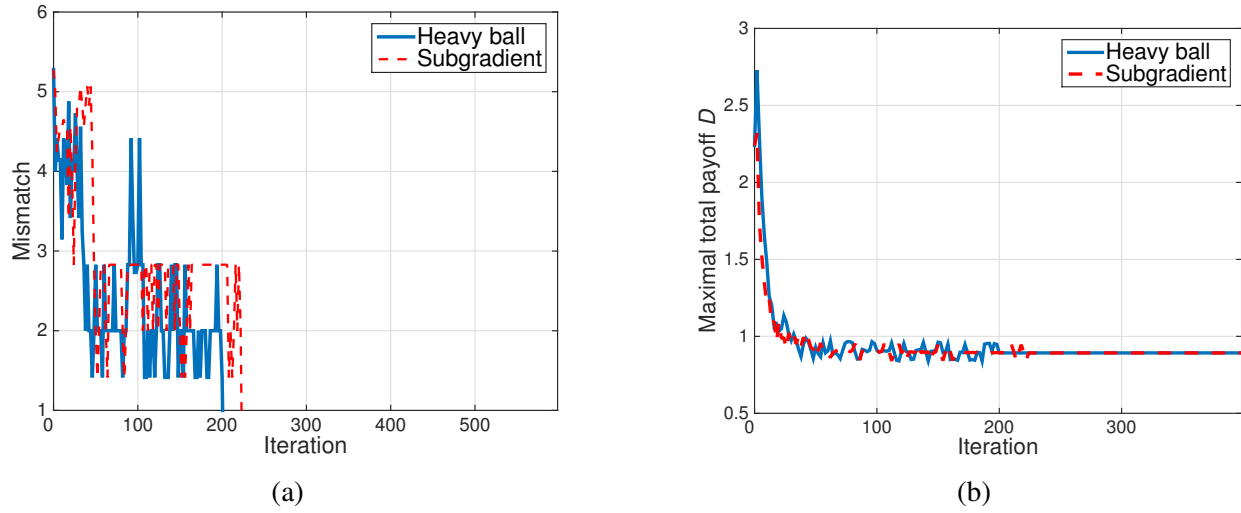


Figure 2.2: Convergence of the heavy ball algorithm.

angle towards the optimal result than the sub-gradient algorithm. From Fig. 2.2 (a), we can see that, the mismatch between the supply and demand of data and backhaul service reaches 0 upon the convergence of the heavy ball algorithm and sub-gradient algorithm where the market clearance constraint is satisfied. Meanwhile, Fig. 2.2 (b) shows that the minimization of D is reached upon the convergence of both algorithms where the total payoff maximization problem (4.10) is solved. Here, the social welfare measures the best payoffs that can be gained by all the ISDN system at each iteration. In such a case, a Walrasian equilibrium scheme is reached at the convergence of the heavy ball algorithm and, also, the sub-gradient algorithm.

Fig. 2.3 shows the convergence of the proposed heavy ball algorithm as the number of DBSs varies. In Fig. 2.3, we can see that as the number of DBS increases, the number of iterations needed for convergence increases. This stems from the fact that, as the number of DBSs increases, the number of communication links increases which requires more iterations to find the optimal user association and resource allocation schemes. Fig. 2.3 also shows that the maximum total payoff decreases as the number of DBSs increases. This is because the data rates at each communication link decrease, with the newly added DBSs introducing significant interference to the ISDN.

In Fig. 2.4, we show how, at a Walrasian equilibrium of the studied ISDN, the price of the data service at radio access links and backhaul service at backhaul links varies with the normalized data rate. Note that, here, the radio access data rates are normalized over the highest data rate on the radio access links, the backhaul data rates are normalized over the highest data rate one can get from backhaul links. As normalized data rate increases from 0.2 to 1, the prices for both a data service and backhaul service increase. This is because the users and BSs would submit higher bids for high-rate data and backhaul service. However, the price only slightly increases as normalized data rate increases from 0.6 to 1. This stems from the fact that, with the increase in its normalized data rate, a link gets the highest “quote” (or price) of its data or backhaul service for guaranteed adoption, henceforth, the users and BSs will stop proposing higher quotes for this link so as to

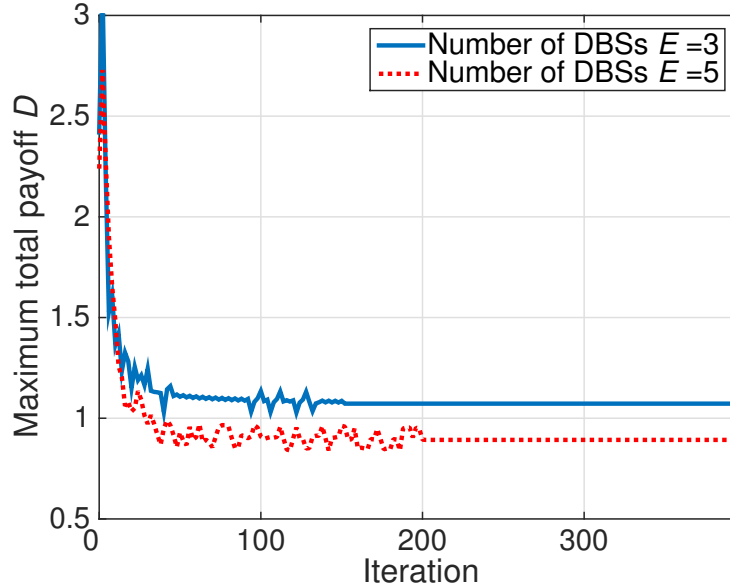


Figure 2.3: Convergence of heavy ball algorithm as the number of DBSs varies.

maintain higher payoffs. We finally note that, the reason why the data rates do not increase from 0 in Fig. 2.4 is that there will be no communication links having a zero data rate. This is because that, at a Walrasian equilibrium, every communication link in the network is allocated non-zero resources, such that they are guaranteed to have a positive data rate.

Fig. 2.5 shows the cumulative distribution function (CDF) in data rates resulting from all the considered schemes. Here, the cumulative distribution function describes the distribution of the data rates at various radio access links and backhaul links in the studied system. More precisely, it allows quantifying the probability with which a communication link in the integrated satellite-drone network will have a data rate less than or equal to certain values (represented by the x-axis). For example, from Fig. 2.5a, we can see that about 30% of the radio access links achieve a data rate lower than 35 Mbps. In Fig. 2.5, we can see that the proposed approaches achieves 2.1 times gains in the number of radio access links that operate over a 40 Mbps rate, and over 3.4 times gains in the number of backhaul links with over 1.6 Gbps data rate, compared to the random allocation scheme. This means that more users and BSs receive higher data rates with the proposed solution. The main reason is that the proposed algorithm encourages communication links to associate with the optimal users and BSs, while fully exploiting the time resources. Meanwhile, Fig. 2.5 shows that the proposed algorithm does not sacrifice the system's communication performance for lower complexity and overhead. In contrast, it still yields the same solution of the centralized algorithm.

In Fig. 2.7, we show how the average data rate at a Walrasian equilibrium varies with the number of DBSs and SBSs. Fig. 2.7 (a) shows that, as the number of DBSs increases from 1 to 6, the average data rates at both radio access links and backhaul links decreases. This stems from the fact the newly added DBSs introduce significant interference to the system. Fig. 2.7 (a) also shows

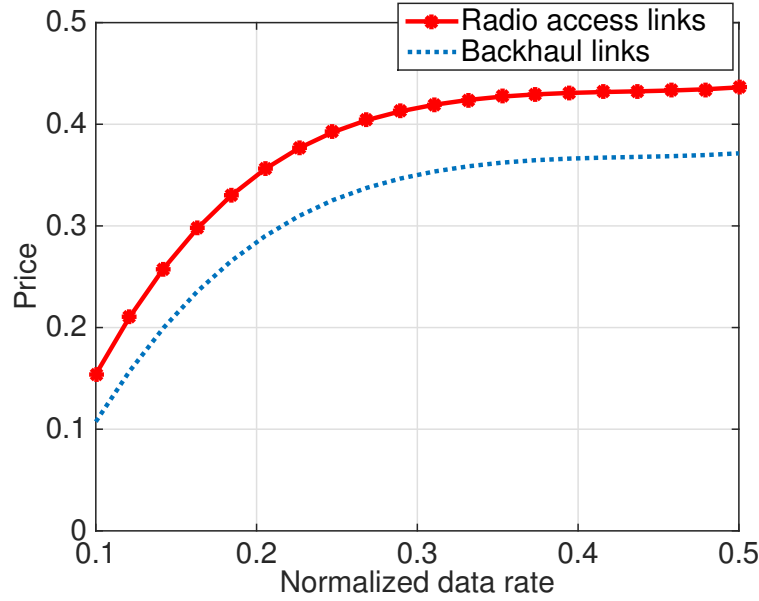


Figure 2.4: Price as the normalized data rate varies.

that, the decrease in the average data at backhaul links is higher than the one at radio access links. This is because the interference introduced by a DBS over the backhaul links is larger than the one it introduces to the radio access links, since the channel gain between a DBS and an MBS or SBS is mostly larger than the channel gain between the DBS and a terrestrial user. Since the transmit power of each DBS is assumed to be equal for all the DBSs, and it will remain constant during the studied period, the received power from a DBS at an MBS or SBS is much larger than the one received at a terrestrial user. Fig. 2.7 (b) shows how the average data rate at a Walrasian equilibrium varies with the number of SBSs. In Fig. 2.7 (b), we can see that, as the number of SBSs increases from 1 to 6, the average data rates at both radio access links and backhaul links decreases. This stems from the fact the newly added SBSs introduce interference to the system. However, the decrease in data rates caused by adding more SBSs is much smaller than the one caused from increasing the number of DBSs. This is due to the fact that DBSs can provide LoS links which will then introduce a significantly larger interference over the backhaul and radio access links.

Fig. 2.8 shows how the amount of data service provided at radio access links varies as the number of DBSs and MBSs increases. Fig. 2.8a shows that, as the number of DBS increases, the amount of data service provided at radio access links increases, with the increase of available time resources in the ISDN. From Fig. 2.8a, we can also see that, as the number of DBSs increases, the amount of data service provided at radio access links increases at a lower speed. This is because a larger number of DBSs yields lower data rates. Fig. 2.8a also shows that the amount of data service provided at the radio access links stops increasing as the number DBSs increases from 3 to 6 when the satellite is not deployed in the system. This is because, as the number of DBSs increases, the only deployed MBS is not capable of serving all the BSs, which leads to some users not being

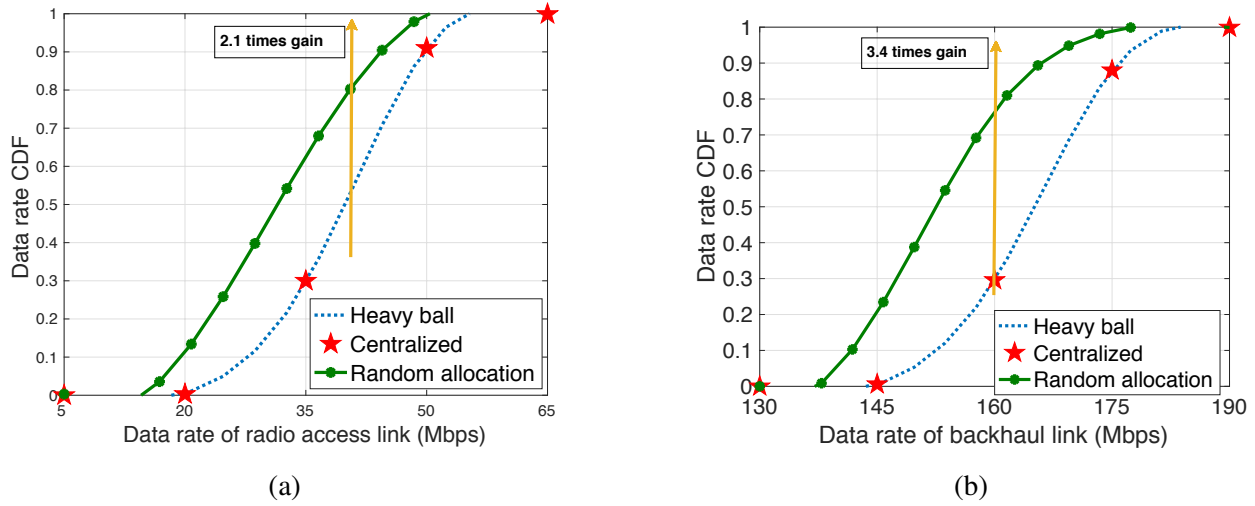


Figure 2.5: Transmitted data at radio access links as the number of DBSs and MBSs varies.

served. Fig. 2.8b shows that, as the number of MBSs increases, the amount of data service provided at the radio access links increases. This is because time resources that are available for the radio access links increases with the number of BSs in the scenario with multiple MBSs. Fig. 2.8b also shows that the amount of data service provided at radio access links increases at a lower rate, as the number of DBSs increases, as the increased interference introduced by the the larger number of MBSs leads to lower data rates on radio access links.

Fig. 2.9 shows how the average data rate at a Walrasian equilibrium varies with the altitude of DBSs. In Fig. 2.9, we can see that, as the altitude of DBSs increases, the average data rates at backhaul links decreases. This stems from the fact an increased DBS altitude leads to larger distance, and thus, worse path loss among the DBSs and MBSs. From Fig. 2.9, we can also see that the average data rate at radio access link increases when the altitude of DBSs increases from 140 to 200 meters, and decreases when altitude of DBSs is larger than 200 meters. This is due to the fact that, as the altitude of DBSs increases, the distance among the DBSs and users increases, the path loss among the all DBSs and users get worse. This means that, at one DBS-user link, both the received power and interference decrease with the increase of the DBS altitude. Firstly, the decrease on the interference is relatively small, in which case the data rate at the DBS-user link increases. However, the reduction on the user's received power increases much faster, the data rates at this user will eventually decrease as the altitude of DBSs increases.

2.5 Conclusion

In this chapter, we have introduced a novel joint resource allocation and user association problem in an ISDN system in which satellite backhaul links share time resources with terrestrial backhaul links and terrestrial radio access links. We have formulated this problem using a competitive

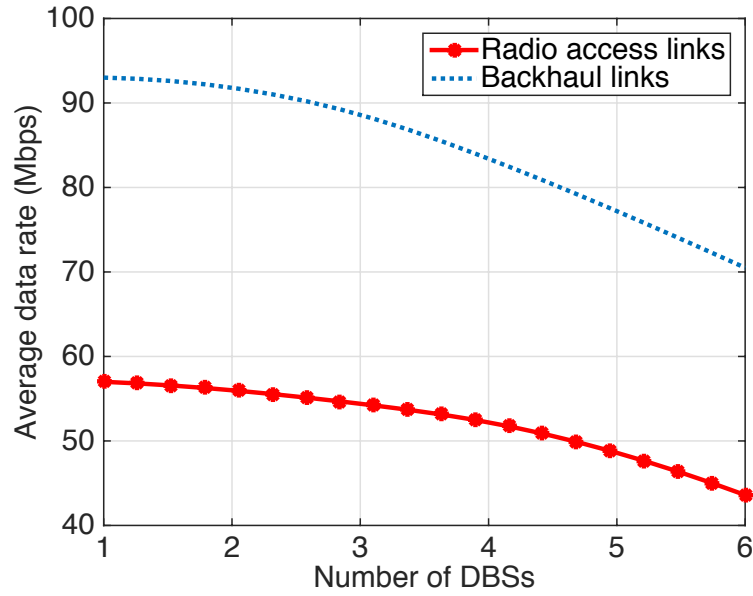


Figure 2.6: Average data rate as the number of DBSs varies.

market in which the data rate of the communication links is optimized distributively with a heavy ball algorithm. Simulation results have shown that, by considering the Walrasian equilibrium in the market, the proposed algorithm can yield over two-fold gain in terms of the number of radio access links with over 40 Mbps rates, and over three-fold gain in terms of the number of backhaul links with over 1.6 Gbps rates. The heavy ball algorithm also yields a 10% improvement in the convergence speed, compared to the sub-gradient algorithm. Future work could include the multi-antenna scenarios with the consideration of precoding technique.

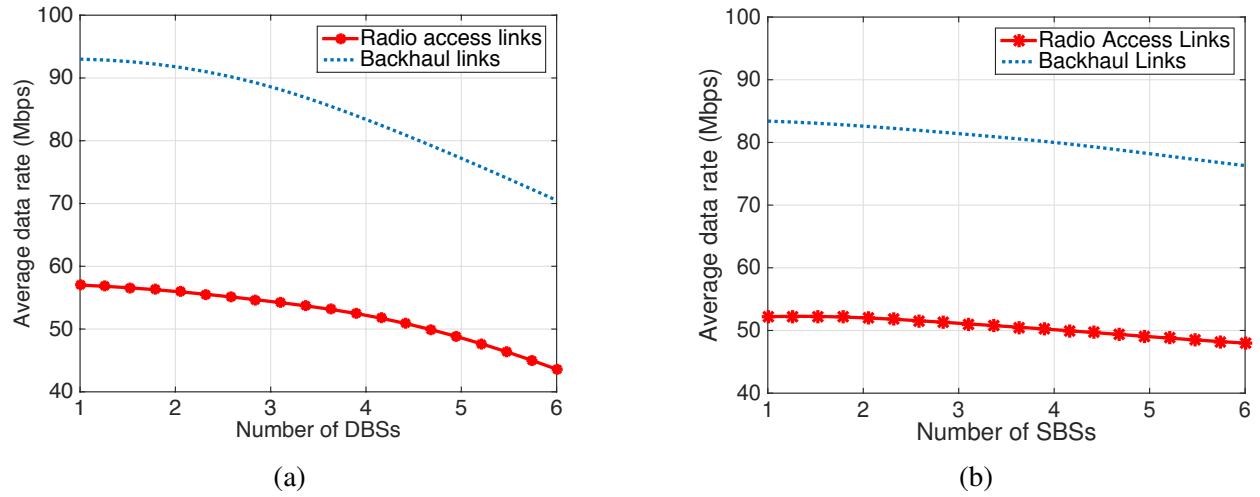


Figure 2.7: Average data rate as the number of DBSs varies.

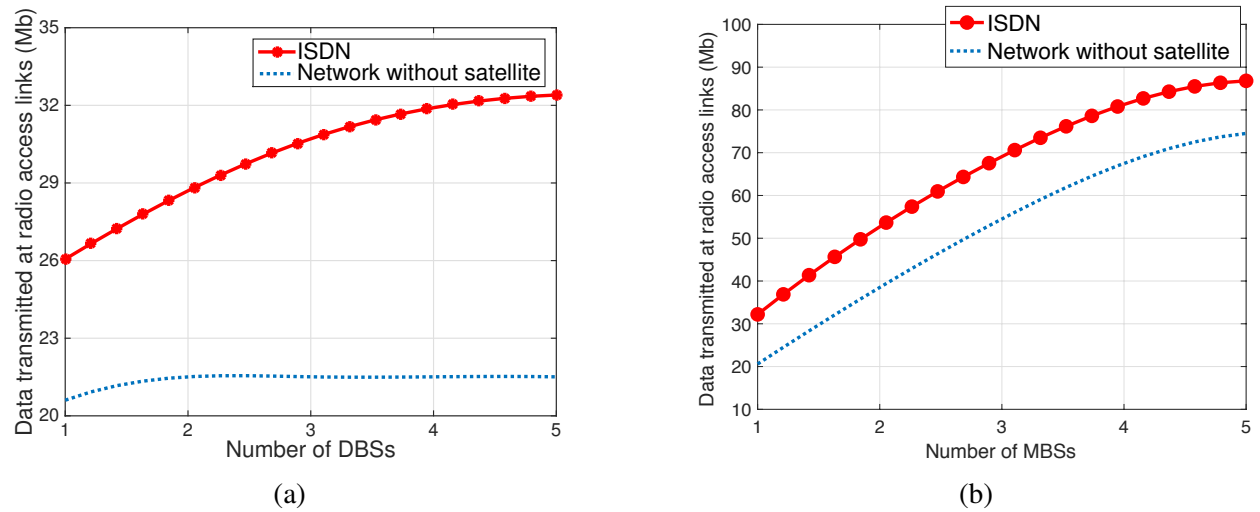


Figure 2.8: Transmitted data at radio access links as the number of DBSs and MBSs varies.

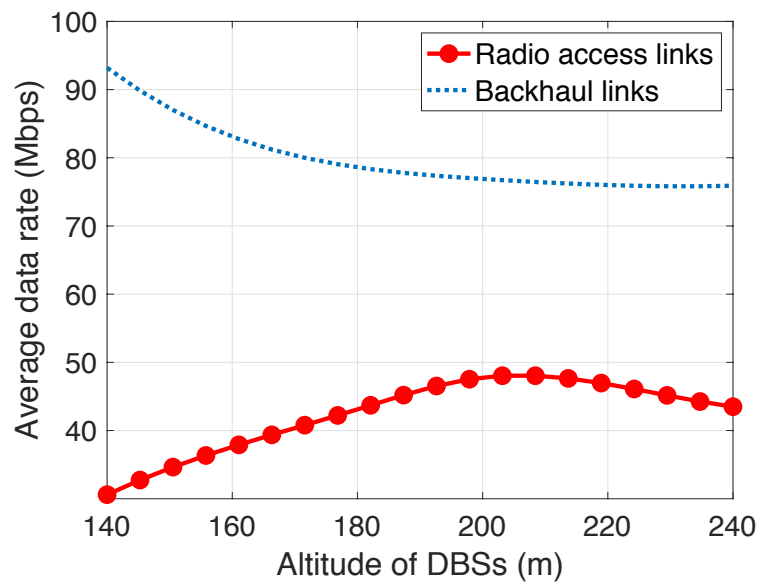


Figure 2.9: Average data rate as the altitude of DBSs varies.

Algorithm 1 Proposed algorithm for ISDN total payoff maximization problem.

Require: The working time limits of each BS, data service request of each user, QoS requirements at each user and BS.

Ensure: Initialize dual variables $\lambda^{(0)}$, $\varsigma^{(0)}$ and $\xi^{(0)}$, which are the initial normalized prices.

- 1: Set step $k = 1$.
 - 2: **for** User $u = 1 : U$ **do**
 - 3: Solve optimization problem (2.24) with dual variable $\lambda^{(k-1)}$ and find the optimal solution $\theta^{(k-1)}$.
 - 4: Send scheme $\theta^{(k-1)}$ to BSs.
 - 5: **end for**
 - 6: **for** BS $n = 1 : N$ **do**
 - 7: Solve optimization problem (2.25) with dual variables $\lambda^{(k-1)}$, $\varsigma^{(k-1)}$ and $\xi^{(k-1)}$ and find the optimal solution $\rho^{(k-1)}$, $\varphi^{(k-1)}$ and $\epsilon^{(k-1)}$.
 - 8: Compute market clearance violation $s_1^{(k)}$, update price $\lambda^{(k)}$.
 - 9: Send schemes $\varphi^{(k-1)}$ and $\epsilon^{(k-1)}$ to MBSs and the satellite, respectively.
 - 10: **end for**
 - 11: **for** MBS $m = 1 : M$ **do**
 - 12: Solve optimization problem (2.27) with dual variable $\varsigma^{(k-1)}$ and find the optimal solution $\delta^{(k-1)}$.
 - 13: Compute market clearance violation $s_2^{(k)}$, update price $\varsigma^{(k)}$.
 - 14: **end for**
 - 15: **for** The satellite **do**
 - 16: Solve optimization problem (2.28) with dual variable $\xi^{(k-1)}$ and find the optimal solution $\beta^{(k-1)}$.
 - 17: Compute market clearance violation $s_3^{(k)}$, update price $\xi^{(k)}$.
 - 18: **end for**
 - 19: Set step $k = k + 1$.
 - 20: **repeat**
 - 21: Step 2 until Step 19.
 - 22: **until** Convergence
 - 23: **return** Optimal allocation vectors $\theta^* = \rho^*$, $\varphi^* = \delta^*$ and $\epsilon^* = \beta^*$.
-

Chapter 3

Competitive Market for Joint Access and Backhaul Resource Allocation in Satellite-Drone Networks

In Chapter 2, the cooperation among NTN and terrestrial networks within downlink service scenario has been investigated with a distributed solution. In this chapter, the selective usage of NTN and terrestrial service on the uplink is studied. In particular, a light weight solution with fast convergence performance is studied for high data rate service scheduling within the integrated system.

3.1 Background, Related Works, and Contributions

The existing literature such as in [119–121] has studied a number of problems related to resource allocation for drone-based systems with backhaul considerations. The work in [119] investigates the problem of dynamic link rerouting in a UAV-based relay network. In [120], a joint caching and resource allocation is investigated for a network of cache-enabled UAVs. In [121], the authors investigate the problem of radio resource allocation in a UAV assistant network, while employing a NOMA scheme on the wireless backhaul transmission. Despite their promising results, these existing works [119–121] do not consider the implementation of an *integrated space-air-ground network (ISDN)*, in which satellite systems provide backhaul connectivity for UAVs. The existing literature such as [105] has studied a number of problems related to ISDNs. The authors in [105] analyze the Doppler effect, the pointing error effect, and the atmospheric turbulence effect on the communication performance of a UAV-to-satellite optical communication system. However, the works in [105] do not consider any spectrum sharing problem in an ISDN system. Indeed, sharing the stringent spectrum resources among satellite backhaul links, terrestrial backhaul links, and the radio access links has a direct impact on the achievable ISDN data rates. Hence, the resource allocation problem over all communication links must be jointly studied within the context of an

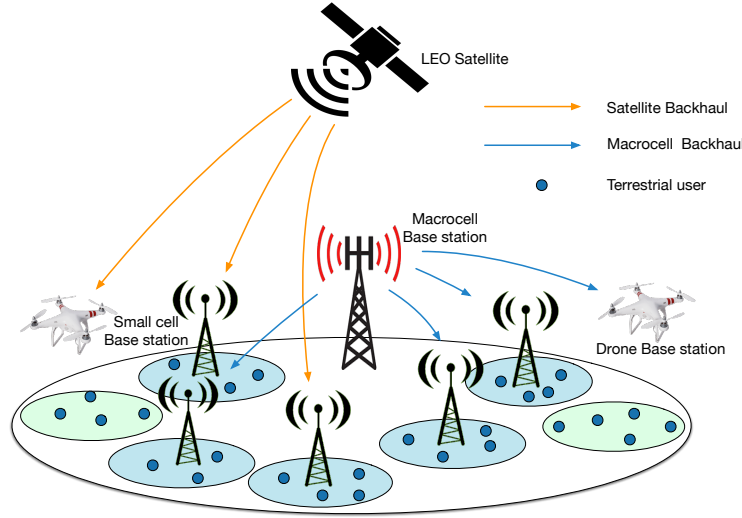


Figure 3.1: Integrated satellite-drone network topology .

ISDN.

The main contribution of this chapter is a novel framework for jointly managing resources across radio access links, satellite backhaul links, and terrestrial backhaul links, while maximizing the data rates in ISDNs. Unlike previous market-based solutions such as [122] and [123] that use market-based approach to consider resource allocation for only two communication links and do not jointly consider resource allocation over backhaul links, this joint backhaul and access resource management problem is formulated in a competitive market in which user association outcomes and wireless resources (sub-channels) are viewed as goods that must be exchanged among the wireless users acting as consumers who seek to maximize their profits represented by their data rates. The data rates of satellite backhaul links are first optimized, based on fixed sub-channel allocation and user allocation schemes at the remaining communication links. The problem of sub-channel allocation and user association at terrestrial backhaul and radio access links is solved as a maximum weighted bipartite matching (MWBM) problem with the Hungarian method. A Walrasian equilibrium of the studied problem is reached when the proposed algorithm converges. Simulation results show that over two-fold and three-fold gains in terms of the number of downlink service links with 35 Mbps rates and the number of backhaul links with 1 Gbps rates, respectively, can be achieved using our approach.

3.2 System Model and problem formulation

Consider an ISDN that consists of a set $\mathcal{U}$ of U terrestrial users, N_D DBSs, N_S SBSs, a set $\mathcal{M}$ of M MBSs providing terrestrial backhaul links, and a LEO satellite constellation providing a satellite backhaul system. In this network, SBSs serve the users in the downlink, while the DBSs are used to supplement the ground network by providing service to the terrestrial users that are not satisfied

with the service provided by SBSs. The backhaul connections of DBSs and SBSs are provided by either the LEO satellite or MBSs. Hereinafter, we represent the set of DBSs and SBSs with $\mathcal{N}$, and both the DBSs and SBSs in the network are referred to as BSs. Note that, in this model, as done in [111], the downlink interference caused by a satellite on the terrestrial links is considered to be negligible due to the satellite's limited transmit power.

3.2.1 Terrestrial Links Analysis

In this subsection, we introduce the transmission links from BSs to users and backhaul links from MBSs to BSs, which are referred to as terrestrial links hereinafter. In this network, BS serves its associated users using the Ka band, as done in [109]. We assume that, at each time slot, the Ka band is divided into K equally-sized frequency sub-channels with bandwidth B^N . Here, the link between BS n and user u is denoted by $I_{n,u}^0$. The link between BS n and MBS m is denoted by $I_{n,m}^1$. The set of *communication links* between BSs and users is denoted as $\mathcal{S}^0$, while the set of *backhaul links* between MBSs and BSs is denoted by $\mathcal{S}^1$. The set of K sub-channels available for each BS n is $\mathcal{K}$, where $k \in \mathcal{K}$ indicates one sub-channel.

Let $\mathbf{x} = [x_{1,1,1}^0, \dots, x_{N,U,K}^0, x_{1,1,1}^1, \dots, x_{N,M,K}^1]$ be the sub-channel allocation vector with $x_{n,u,k}^0 \in \{0, 1\}$ and $x_{n,m,k}^1 \in \{0, 1\}$ indicating whether sub-channel $k \in \mathcal{K}$ is assigned to link $I_{n,u}^0$ and $I_{n,m}^1$, respectively. $x_{n,u,k}^0 = 1$ and $x_{n,m,k}^1 = 1$ indicate that sub-channel k is assigned to link $I_{n,u}^0$ and $I_{n,m}^1$, respectively. Let $\mathbf{y} = [y_{1,1}^0, \dots, y_{N,U}^0, y_{1,1}^1, \dots, y_{N,M}^1]$ be the user association vector with $y_{n,u}^0 \in \{0, 1\}$ being the association status between user u and BS n , and $y_{n,m}^1 \in \{0, 1\}$ being the association status between terrestrial MBS m and BS n . Here, $y_{n,u}^0 = 1$ and $y_{n,m}^1 = 1$ indicate that BS n associates with user u , and BS n associates with MBS m , respectively. Let $\rho_{n,u,k}^0$ and $\rho_{n,m,k}^1$ be an allocation scheme, considering the user association and sub-channel allocation at link $I_{n,u}^0$ and $I_{n,m}^1$, respectively, where $\rho_{n,u,k}^0 = x_{n,u,k}^0 y_{n,u}^0$ and $\rho_{n,m,k}^1 = x_{n,m,k}^1 y_{n,m}^1$. As such, $\rho_{n,u,k}^0 = 1$ indicates that user u is associate with BS n via sub-channel k . $\boldsymbol{\rho}_N = [\rho_{1,1,1}^0, \dots, \rho_{N,U,K}^0, \rho_{1,1,1}^1, \dots, \rho_{N,M,K}^1]$ is defined as an allocation vector for terrestrial links.

The interference at link $I_{n,u}^0$ working on sub-channel k is:

$$N_{n,u,k}(\boldsymbol{\rho}_N) = \sum_{i \in \mathcal{S}, i \neq I_{n,u}^0} \rho_{n_i, u_i, k} O_{n,u}^i P_{n_i} G(\theta_{n,u}^i) h(n_i, u), \quad (3.1)$$

where u_i and n_i represent the receiver (user or BS) and transmitter (BS or MBS) in link i . Here, if link i is a BS-user link, $\rho_{n_i, u_i, k} = \rho_{n_i, u_i, k}^0$, if link i is an MBS-BS backhaul link, $\rho_{n_i, u_i, k} = \rho_{n_i, u_i, k}^1$. $O_{n,u}^i = b_{k(i)}^k / B^N$ is the compensation factor defined as the portion of the interfering sub-channel $k(i)$ that overlaps with the victim sub-channel bandwidth k . Here, $k(i)$ represents the sub-channel allocated to terrestrial link i , in other words, $x(n, u, k(I_{n,u}^0)) = 1$. $b_{k(i)}^k$ is the bandwidth of interfering sub-channel $k(i)$ that overlaps with the bandwidth of victim sub-channel k . P_{n_i} is the transmit power of the transmitter (BS or MBS) at link i . Note that, if $n_i \in \mathcal{N}$, $P_{n_i} = P_{n_i}^{\text{BS}}$, if $n_i \in \mathcal{M}$, $P_{n_i} = P_{n_i}^{\text{MBS}}$, where $P_{n_i}^{\text{BS}}$ is the transmit power of BS n_i , $P_{n_i}^{\text{MBS}}$ is the transmit power of MBS n_i . $G(\theta_{n,u}^i)$ is the transmit gain of the terrestrial antenna at an offset angle $\theta_{n,u}^i$, which is

defined as:

$$G(\theta_{n,u}^i) = \begin{cases} Q, & \text{if } \theta_{n,u}^i \leq \frac{\theta}{2}, \\ Q', & \text{otherwise,} \end{cases} \quad (3.2)$$

where Q and Q' are the antenna gains of the main lobe and the sidelobe, respectively. $\frac{\theta}{2}$ is the half power beamwidth. $\theta_{n,u}^i$ is the offset angle (from the boresight direction) of the link i 's transmit antenna in the direction of link $I_{n,u}^0$'s receiver antenna. $h(n_i, u) = 10^{-0.1(L_{FS}(d_0) + 10\mu \log(d_u^{n_i}) + \chi_\delta)}$ is the path loss from transmitter n_i to user u , considering shadow fading. $L_{FS}(d_0) = 20 \log\left(\frac{4\pi d_0 f_{k(i)}}{v}\right)$ is the free space path loss, with v being the light speed, d_0 being the free-space reference distance, and $f_{k(i)}$ being the carrier frequency of sub-channel $k(i)$. $d_u^{n_i}$ represents the distance between the transmitter n_i to user u . μ represents the path loss exponent for communication link i , and is defined as :

$$\mu = \begin{cases} \mu_{\text{LoS}}, & \text{if } I_{n_i,u}^T \text{ is a LoS link,} \\ \mu_{\text{NLoS}}, & \text{otherwise,} \end{cases} \quad (3.3)$$

where μ_{LoS} and μ_{NLoS} represent path loss exponents for the LoS and NLoS links, respectively. Here, $I_{n_i,u}^T$ is an interference link from transmitter n_i to user u . χ_δ is Gaussian random variable in dB with zero mean, where δ represents the standard deviation for link i in dB, and is defined as follow:

$$\delta = \begin{cases} \delta_{\text{LoS}}, & \text{if } I_{n_i,u}^T \text{ link is a LoS link,} \\ \delta_{\text{NLoS}}, & \text{otherwise,} \end{cases} \quad (3.4)$$

where δ_{LoS} and δ_{NLoS} represent the standard deviation for LoS and NLoS links in dB, respectively.

The SINR of link $I_{n,u}^0$ that uses sub-channel k is computed as $\gamma_{n,u,k}(\rho_N) = \frac{P_n^{\text{BS}} h(n,u) G(\theta_{n,u}^{n,u})}{N_{n,u,k}(\rho_N) + \sigma^2}$, where σ^2 is the thermal noise power. As done in [124], we only consider the LoS transmission links over DBS-user links. The data rate at link $I_{n,u}^0$ is given by:

$$c_{n,u}(\rho_N) = \sum_{k \in \mathcal{K}} \rho_{n,u,k}^0 B^N \log(1 + \gamma_{n,u}(\rho_N)). \quad (3.5)$$

The interference at link $I_{n,m}^1$ that uses sub-channel k is:

$$N_{n,m,k}(\rho_N) = \sum_{i \in \mathcal{I}, i \neq I_{n,m}^1} \rho_{n_i,u_i,k} O_{n,m}^i P_{n_i} h(n_i, n) G(\theta_i^{n,m}), \quad (3.6)$$

where $O_{n,m}^i$ represents the compensation factor for interference between link i and link $I_{n,m}^1$. Thus, the SINR at link $I_{n,m}^1$ is $\gamma_{n,m}(\rho_N) = \frac{P_M^{\text{MBS}} h(n,m) G(\theta_{n,m}^{n,m})}{N_{n,m}(\rho_N) + \sigma^2}$, where P_M^{MBS} represents the transmit power of each MBS m , which is assumed to be equal for all MBSs. The data rate at link $I_{n,m}^1$ is:

$$c_{n,m}(\rho_N) = \sum_{k \in \mathcal{K}} \rho_{n,m,k}^1 B^N \log(1 + \gamma_{n,m}(\rho_N)). \quad (3.7)$$

3.2.2 Satellite Communication Links Analysis

In the Ka band, a LEO satellite provides supplemental backhaul connection to BSs. We assume that the LEO satellite separates Ka band into L equally sized frequency sub-channels with bandwidth B^S . The link between the satellite and BS n is $I_{s,n}^2$, while the set of L frequency sub-channels available for satellite backhaul links is $\mathcal{L}$. Let $\mathbf{x}^2 = [x_{1,1}^2, \dots, x_{N,L}^2]$ be the sub-channel allocation vector, where $x_{n,l}^2 \in \{0, 1\}$ represents the sub-channel assignment status at link $I_{s,n}^2$. $x_{n,l}^2 = 1$ indicates that sub-channel $l \in \mathcal{L}$ is assigned to satellite backhaul link $I_{s,n}^2$. In the BSs association vector $\mathbf{y}^2 = [y_1^2, \dots, y_N^2]$, $y_n^2 \in \{0, 1\}$ indicates the association status between the satellite and BS n , where $y_n^2 = 1$ indicates that BS n is associated with the satellite. Here, $\boldsymbol{\rho}_S = [\rho_{1,1}^2, \dots, \rho_{N,L}^2]$ is defined as a satellite backhaul allocation vector, in which $\rho_{n,l}^2 = x_{n,l}^2 y_n^2$ is one allocation scheme of link $I_{s,n}^2$.

The interference over the transmission link between BS n and a satellite working on sub-channel l can be expressed as:

$$N_{n,l}^S(\boldsymbol{\rho}_N) = \sum_{k \in \mathcal{K}_l} \sum_{i \in \mathcal{I}} \rho_{n_i, u_i, k} O_{s,n}^i P_{n_i} h(n_i, n) G(\theta_i^{s,n}), \quad (3.8)$$

where $\mathcal{K}_l$ is the set of terrestrial sub-channels that partially overlap with the bandwidth of victim satellite sub-channel l .

Similarly, the SINR of the backhaul links between BS n and a satellite on sub-channel l will be $\gamma_{s,n,l}(\boldsymbol{\rho}_N) = \frac{P_S h(s,n) G^S(\theta_{s,n}^{s,n})}{N_{n,l}^S(\boldsymbol{\rho}_N) + N_c + \sigma^2}$, where P_S is the transmit power of the LEO satellite, which is assumed to be equal for all satellites. $G^S(\theta_{s,n}^{s,n})$ is the transmit gain of satellite antenna at offset angle $\theta_{s,n}^{s,n}$. N_c is the co-channel interference due to the use of a multi-beam satellite. Thus, the data rate of link $I_{s,n}^2$ is:

$$c_{s,n}(\boldsymbol{\rho}_N, \boldsymbol{\rho}_S) = \sum_{l \in \mathcal{L}} \rho_{n,l}^2 B^S \log(1 + \gamma_{s,n}(\boldsymbol{\rho}_N)). \quad (3.9)$$

The QoS requirements in the ISDN system is $c_u(\boldsymbol{\rho}_N) = \sum_{n \in \mathcal{N}} c_{n,u}(\boldsymbol{\rho}_N) \geq \theta_u$ and $c_n(\boldsymbol{\rho}_N, \boldsymbol{\rho}_S) = c_{s,n}(\boldsymbol{\rho}_N, \boldsymbol{\rho}_S) + \sum_{m \in \mathcal{M}} c_{n,m}(\boldsymbol{\rho}_N) \geq \theta_n$, where θ_u and θ_n are the minimum rate thresholds required by user u and BS n , respectively.

3.2.3 Problem Formulation

Given this system model, our goal is to find an effective resource allocation and user association scheme for maximizing the system sum-rate. To achieve this goal, we formulate a competitive market to fully maximize the sum-rate by making the most of the spectrum resources, considering the dependence of the communication links' data rate on their adopted allocation scheme. In the considered resource allocation market, the communication links are the consumers, the user

association outcomes and available sub-channels are the goods. The consumers *buy* quantities of goods and also *sell* goods they possess such that they maximize their profit, which is captured by the data rate [123], at the price of interference caused by spectrum sharing. However, the profit optimization objectives of the consumers may be conflicting, as the optimal sub-channel and user association scheme for one consumer may introduce high interference to other consumers in the market. Thus, the consumers' demands may not match the supplies of the sub-channels in the profit maximization process, leading to a loss on the consumers' final profit. To guarantee a full utilization of the spectrum resources, we adopt the notion of *Walrasian equilibrium* [122]:

Definition 2. Given the price of the goods (i.e. the interference of each sub-channel), a *Walrasian equilibrium* the considered ISDN is a pair of vectors $\boldsymbol{\rho}_N^*, \boldsymbol{\rho}_S^*$, such that:

1. For any communication link, the pair $\boldsymbol{\rho}_N^*, \boldsymbol{\rho}_S^*$ is the optimal solution that maximizes their data rate.
2. The quantity of provided sub-channels equals to the demand of sub-channels that are requested by the users, i.e. $\sum_{\substack{n \in \mathcal{N}, \\ l \in \mathcal{L}}} \rho_{n,l}^{2*} = L, \sum_{\substack{u \in \mathcal{U}, \\ k \in \mathcal{K}}} \rho_{n,u,k}^{0*} + \sum_{\substack{m \in \mathcal{M}, \\ k \in \mathcal{K}}} \rho_{n,m,k}^{1*} = K$.

Note that, at the Walrasian equilibrium, the communication links' optimal data rates are guaranteed and the clearance of the market is ensured by a full utilization of frequency resources. The computation of the Walrasian equilibrium in our competitive market is introduced in the next section.

3.3 Proposed Solution

In this section, the problem of finding a Walrasian equilibrium is cast into a social welfare maximization problem where the supplies of frequency resource (i.e. sub-channels) match with their demands. Here, the social welfare is defined as the sum-rates of all communication links. As such, our main goal is to design an allocation scheme to achieve optimal sum-rates across all communication links, considering the clearance of goods. The corresponding sum-rate maximization problem is:

$$\max_{\boldsymbol{\rho}_N, \boldsymbol{\rho}_S} \left(\sum_{u=1}^U c_u(\boldsymbol{\rho}_N) + \sum_{n=1}^{N_D+N_S} c_n(\boldsymbol{\rho}_N, \boldsymbol{\rho}_S) \right), \quad (3.10)$$

$$\text{s. t. } c_u \geq \theta_u, \quad c_n \geq \theta_n, \quad n \in \mathcal{N}, u \in \mathcal{U}, \quad (3.10a)$$

$$\sum_{u \in \mathcal{U}} c_{n,u} \leq \sum_{m \in \mathcal{M}} c_{n,m} + c_{s,n}, \quad n \in \mathcal{N}, \quad (3.10b)$$

$$\sum_{k \in \mathcal{K}} \rho_{n,u,k}^0 = 1, \quad n \in \mathcal{N}, u \in \mathcal{U}, \quad (3.10c)$$

$$\sum_{k \in \mathcal{K}} \rho_{n,m,k}^1 \geq 1, \quad n \in \mathcal{N}, m \in \mathcal{M}, \quad (3.10d)$$

$$\max_{\rho_N} \left(\sum_{n \in \mathcal{N}} \left(\sum_{u \in \mathcal{U}} c_{n,u}(\rho_N) + \sum_{m \in \mathcal{M}} c_{n,m}(\rho_N) \right) + \max_{\rho_S} \sum_{n \in \mathcal{N}} c_{s,n}(\rho_N, \rho_S) \right). \quad (3.12)$$

$$\sum_{\substack{n \in \mathcal{N}, \\ k \in \mathcal{K}}} \rho_{n,u,k}^0 = 1, u \in \mathcal{U}, \quad (3.10e)$$

$$\sum_{\substack{m \in \mathcal{M}, \\ k \in \mathcal{K}}} \rho_{n,m,k}^1 = \sum_{k \in \mathcal{K}} \rho_{n,m,k}^1, n \in \mathcal{N}, \quad (3.10f)$$

$$\sum_{l \in \mathcal{L}} \rho_{n,l}^2 = 1, n \in \mathcal{N}, \quad (3.10g)$$

$$\sum_{\substack{n \in \mathcal{N}, \\ l \in \mathcal{L}}} \rho_{n,l}^2 = L, \quad (3.10h)$$

$$\sum_{\substack{u \in \mathcal{U}, \\ k \in \mathcal{K}}} \rho_{n,u,k}^0 + \sum_{\substack{m \in \mathcal{M}, \\ k \in \mathcal{K}}} \rho_{n,m,k}^0 = K, \quad (3.10i)$$

$$\rho_{n,u,k}^0 \in \{0, 1\}, n \in \mathcal{N}, u \in \mathcal{U}, k \in \mathcal{K}, \quad (3.10j)$$

$$\rho_{n,m,k}^1 \in \{0, 1\}, m \in \mathcal{M}, n \in \mathcal{N}, k \in \mathcal{K}, \quad (3.10k)$$

$$\rho_{n,l}^2 \in \{0, 1\}, n \in \mathcal{N}, l \in \mathcal{L}, \quad (3.10l)$$

$$k(i) \neq k(j), i \in \mathcal{J}_n^B, j \in \mathcal{J}_n^R \quad (3.10m)$$

where the maximization problem in (10) captures profit maximization objective in the market, with (10a) being the QoS requirements. Meanwhile, (10b) indicates that the sum rate of users associated to each BS cannot exceed the capacity of the backhaul of this BS. (10c) shows that each BS-user link can occupy only one sub-channel for data transmission. (10d) indicates that each MBS-BS backhaul link occupy at least one sub-channel. (10e) and (10f) indicate that only one BS-user link is associated to each user, and only one MBS-BS backhaul link is associated to each BS. (10g) indicates that each satellite backhaul link can only be assigned with one sub-channel which should not be shared with other satellite backhaul links, as the satellite links work on single carrier communication mode. (10h) and (10i) represent the clearance of spectrum resources. (10m) indicates that one sub-channel cannot be simultaneously for backhaul links and radio access links of a BS n at the same time. Here, i is one terrestrial link, $\mathcal{J}_n^B$ is a set of backhaul links associated with BS n , and $\mathcal{J}_n^R$ is a set of radio access links provided by BS n .

It is worth noting that the optimization problem in (3.10) is intractable due to the nonlinear coupling between terrestrial links and satellite communication links. This stems from the fact that the sub-channel allocation and user association in the terrestrial links can directly affect the data rates of the satellite backhaul links. Here, based on (6), (9) and (12), we rewrite the optimal problem in (3.10) in the form of (3.12). Thus, a decomposition method can be used in the satellite system and MBSs to allocation sub-channels and user association schemes at the terrestrial links and sub-channels at the satellite backhaul links, separately. First, the data rates of the satellite backhaul links are optimized given the sub-channel allocation and user association schemes for the terrestrial links. The optimal allocation scheme at the terrestrial links is chosen based on the both data rates at

terrestrial links and the corresponding satellite backhaul data rates. The optimal solution of the decomposed problem is the optimal solution of (3.10), as the maximization problem in (12) is equivalent to (3.10).

3.3.1 Satellite Sub-channel Allocation and User Association

The sub-channel allocation and user association problem at the satellite communication links is formulated as

$$\begin{aligned} \max_{\boldsymbol{\rho}_S} \quad & \sum_{n=1}^{N_D+N_S} c_{s,n}(\boldsymbol{\rho}_N, \boldsymbol{\rho}_S), \\ \text{s. t.} \quad & c_n \geq \theta_n, \quad n \in \mathcal{N}, \\ & (10g), (10h), (10l) \end{aligned} \tag{3.11}$$

$$\tag{3.11a}$$

Therefore, the sub-channel allocation and user association problem at the satellite communication links is reduced to finding the optimal sub-channels associated to each satellite backhaul link. The resulting optimal sub-channel allocation vector is denoted as $\boldsymbol{\rho}_S^*$. Precisely speaking, given schemes in $\boldsymbol{\rho}_N$, the sum-rate of the satellite backhaul links is maximized by selecting the sub-channel that provides higher SINR to each satellite link. Therefore, the problem in (3.11) can be optimally solved with the Hungarian method.

3.3.2 Terrestrial Sub-channel Allocation and User Association

Finally, the sum-rate of the users is maximized by selecting the optimal sub-channel allocation and user association scheme in the terrestrial links, while considering the resulting data rates at satellite backhaul links. Note that, the data-rate optimization problem at the terrestrial links is a combinatorial optimization problem. Hence, the optimal strategies for the terrestrial links can be found by exhaustive search, which leads to a prohibitive computational complexity, especially when N , M , and U are large. Thus, the optimization problem in (3.10) is transformed into a bipartite matching problem.

We construct a bipartite graph $\mathcal{A} = (\mathcal{V}_{NK} \times \mathcal{V}_{UM}, \mathcal{E})$, in which $\mathcal{V}_{NK}$ represents the set of sub-channels available to each BS n and $\mathcal{V}_{UM}$ represents the set of terrestrial users and MBSs. Note that, in $\mathcal{V}_{NK}$, sub-channel k of BS i is represented by a vertex $v_{NK}(i, k)$, while, in $\mathcal{V}_{UM}$, a terrestrial user j or MBS j is represented by a vertex $v_{UM}(j)$. Denote $\mathcal{E}$ as the set of edges that connect to the vertices from each set, $\mathcal{V}_{NK}$ and $\mathcal{V}_{UM}$. Let $e(i, j, k) \in \mathcal{E}$ be the edge connecting $v_{NK}(i, k)$ and $v_{UM}(j)$, meaning that sub-channel k of BS i is allocated to user or MBS j . Let matching $\mathcal{T}$ be a subset of edges in $\mathcal{E}$, in which no two edges share a common vertex $v_{NK}(i, k)$, such that each sub-channel k of BS i can only be allocated to one user or MBS j . Nevertheless, in $\mathcal{T}$, all of the edges associated with $v_{UM}(j)$, $j \in \mathcal{U}$, (e.g. the communication links associated with user j), will

Algorithm 2 Proposed algorithm for ISDN sum-rate maximization.

Require: The bipartite graph $\mathcal{A} = (\mathcal{V}_{NK} \times \mathcal{V}_{UM}, \mathcal{E})$, allocation constraints in (3.10).

Ensure: Initialize matching $T_t = \emptyset$, which is the matching adopted at step t .

```

1: for BS  $i = 1 : N$  do
2:   for sub-channel  $k = 1 : K$  do
3:     for terrestrial user or MBS  $j = 1 : U + M$  do
4:       Update the optimal satellite backhaul allocation vector  $\rho_S^*$  based on (3.10) using the Hungarian algorithm.
5:       Find  $e(i, j, k)$  based on the allocation vector corresponding to  $T_{t-1}$ .
6:       Based on the allocation vector  $\rho_N^{i'}$  corresponding to matching  $T_{t-1}$ , calculate  $w(i, j, k)$  of each edge  $e(i, j, k)$ .
7:     end for
8:   end for
9:   Find the optimal terrestrial allocation vector  $\rho_N^*$  using the Hungarian algorithm.
10:  Update  $T_t$ ,  $t = t + 1$ .
11: end for
12: repeat
13:   Step 1 until Step 10.
14: until Convergence to a Walrasian equilibrium
15: return Optimal allocation vectors  $\rho_N^*$  and  $\rho_S^*$ .

```

not share a common vertex $v_{UM}(j)$, such that each terrestrial user j can be matched with only one sub-channel. The remaining sub-channels of BS i are allocated to BS i 's terrestrial backhaul link. Here, the set of all possible matchings $\mathcal{T}$ in the formulated ISDN bipartite graph is denoted as $\mathcal{S}_T$, where constraints in (3.10) are satisfied. Here, $e(i, j, k) \in \mathcal{T}$, indicating that BS i is connected to user (or MBS) j via sub-channel k , is captured by $\rho(i, j, k) = 1$, $e(i, j, k) \notin \mathcal{T}$ is captured as $\rho(i, j, k) = 0$. Here, $\rho(i, j, k)$ represent $\rho_{i,j,k}^1$ if $j \in \mathcal{M}$, represent $\rho_{i,j,k}^0$ if $j \in \mathcal{U}$, while $\rho(i, j, l)$ represents $\rho_{i,l}^2$ if j is the satellite. Thus, an user association and sub-channel allocation scheme can be derived as $x(i, j, k) = \rho(i, j, k)$, and $y(i, j) = 1$ when $\rho(i, j, k) > 0$, otherwise $y(i, j) = 0$. Note that, here, $i \in \mathcal{N}$, $j \in \mathcal{U} \cup \mathcal{M}$. Consequently, each possible sub-allocation and user association scheme in the ISDN system can be represented by a matching $\mathcal{T}$, the set of all possible sub-channel allocation and user association schemes in the formulated bipartite matching problem is included in $\mathcal{S}_T$.

Note that, the weight of edge $e(i, j, k)$ is given by $w(i, j, k)$, which indicates the data rate of the link between BS i and user or MBS j via sub-channel k . More precisely, considering the set of allocation schemes $\rho_N^{i'}$, adopted by $I_{i',j}$, where $i' \neq i$, $i \in \mathcal{N}$, and the corresponding optimal satellite allocation scheme ρ_S^* , the weight of $e(i, j, k)$ is given by $w(i, j, k) = c_{i,j}(\rho(i, j, k), \rho_N^{i'}) + c_{s,n}^{\min}(\rho_S^*, \rho(i, j, k), \rho_N^{i'})$. $c_{s,n}^{\min}(\rho_S^*, \rho(i, j, k), \rho_N^{i'})$ is the minimum satellite backhaul data rate with allocation scheme ρ^{S*} , when there is only one active terrestrial link $I_{i,j}$. Hence, the sum-rate maximization problem in (3.10) can be solved by finding a set of edges $\mathcal{T}^* \in \mathcal{S}_T$ with maximized sum-weight. The formulated bipartite matching problem is called MWBM problem [125], which is solved by Algorithm 1. After obtaining the optimal matching of the bipartite graph, we can get the Walrasian equilibrium in the market. In other words, with the sum-rates over all communication links being optimized and the available sub-channels being fully allocated, Walrasian equilibrium strategies can be reached upon the convergence of the proposed algorithm.

Table 3.1: Simulation Parameters [1]

Parameter	Value	Parameter	Value
P_n	20 dBm	P_M	43 dBm
P_S	9.23 dBW	B^N	56 MHz
σ^2	-104 dB	N_c	10.5354 dB

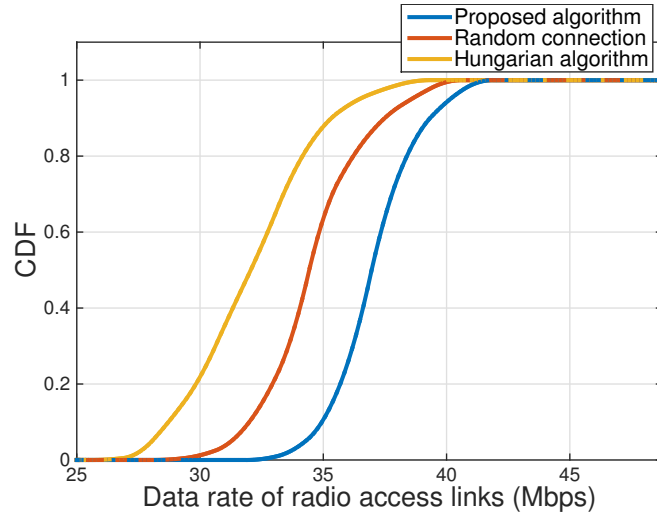


Figure 3.2: Radio access data rate CDF vs. different algorithms.

3.4 Simulation Results and Analysis

For our simulations, we consider a scenario with 130 mobile users, 40 BSs, 1 macro base station, and 1 LEO satellite constellation. Other parameters used in the simulations are listed in Table I. The proposed Walrasian equilibrium results are compared to the random connection scheme results considering the market clearance, noted as random connection algorithm, and a traditional Hungarian algorithm, in which the Walrasian equilibrium is not considered. All statistical results are averaged over a large number of independent runs.

Fig. 1 and Fig. 2 show the CDF in data rates resulting from all the considered schemes. In Fig. 1, we can see that the proposed approaches achieves 2.5 and 6.2 times gains in the number of radio access links that operate at a 35 Mbps rate, compared to the random connection and the traditional Hungarian algorithm, respectively. The proposed approach yields over 3.4 and over 4.4 times gains in the number of backhaul links with 1 Gbps data rate, compared to the random connection and the traditional Hungarian algorithm as shown in Fig. 2. This means that more users and BSs receive higher data rates with the proposed solution. The main reason is that the proposed algorithm encourages communication links to associate with the optimal users and BSs, while fully exploiting the spectrum resources.

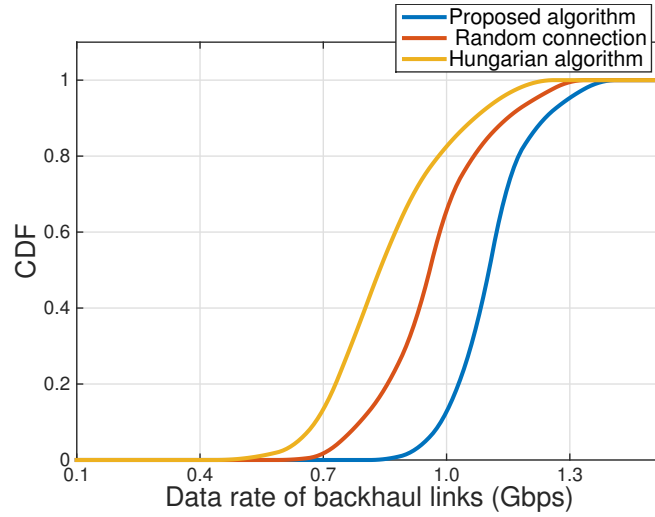


Figure 3.3: Backhaul data rate CDF vs. different algorithms.

3.5 Conclusion

In this chapter, we have introduced a novel joint resource allocation and user association problem in an ISDN system in which satellite backhaul links share spectrum resources with terrestrial backhaul links and terrestrial radio access links. We have formulated this problem using a competitive market in which the sum rate of the communication links is optimized in a sequential manner with a Hungarian based algorithm. Simulation results have shown that, by considering the Walrasian equilibrium in the market, the proposed algorithm can yield over two-fold gain in terms of the number of radio access links with 35 Mbps rates, and over three-fold gain in terms of the number of backhaul links with 1 Gbps rates.

Chapter 4

Distribution Robust-Meta Reinforcement Learning for Dynamic Resilient Contact Plan Design in Cube Satellite Networks

4.1 Background, Related Works, and Contributions

With the optimal integration among NTN and terrestrial networks, we will then investigate the on-demand deployment of NTNs. In particular, as previously discussed in Chapter 1, the LEO cube satellite network within NTNs can provide an endurable, reliable, and accessible data service to users in wireless disadvantaged areas [126, 127]. However, deploying cube satellites for low latency information access is still an important open problem, especially as the satellites have to serve unforeseeable data needs with their limited available on-board resource. In particular, it is challenging for the cube satellites to serve unpredictable, diverse data needs from users around the globe using only limited contact chances in the network. In this chapter, the optimal deployment of cube satellites within NTNs will be investigated with a contact plan design solution that seeks to route data, among the resource constrained satellites, based on the diverse and unforeseeable access needs on the ground.

4.1.1 Related Works

Prior works on satellite communications [80–84] studied a number of problems related to contact plan design in resource constrained cube satellite networks. The work in [80] studies the problem of contact plan design within LEO satellite networks while jointly considering the satellites' on-board energy capacity and their stochastic solar energy infeed. The authors in [81] design contact plan within a resource constrained LEO satellite network to deliver the satellite's observation to fixed ground base station. In [82], the problem of contact selection within a multi-layer satellite-

terrestrial network is investigated with the goal of improving connectivity among the resource constrained satellites. The authors in [83] optimize contact plan design, power allocation, and mission schedule in data relay satellite networks for fair communication performance on satellites. The authors in [84] investigate the suitability of the contact graph routing solution for inter-satellite communication using a Linux testbed. Despite their promising results, these existing works [80–84] do not consider the time-varying, unforeseeable service requests in real world satellite networks. Indeed, the optimization-based solutions in [80–84] are not suitable to design contact plans when the user requests are highly dynamic and unpredictable.

More recently, there has been significant interest in realizing dynamic resilient networking to serve unforeseeable needs within the satellite networks [85–91]. In particular, the work in [85] employs a machine learning method to predict the unforeseeable service needs within the satellite communication networks. The authors in [86] develop a centralized machine learning algorithm that enables real-time estimation on the satellite communication environment, specifically, the ever-changing rain intensity on broadband satellite communication links. Meanwhile, in [87], a deep reinforcement learning (DRL) based intelligent satellite communication configuring solution is deployed on the national aeronautics and space administration (NASA)’s space communication and navigation testbed. Yet when managing operations on every satellite, a centralized solution can cause significant communication overhead, especially for high latency satellite systems. Thus, to intelligently classify the dynamic Internet traffic flows on satellites with low overhead, the authors in [88] develop a distributed hierarchical classification solution using deep learning. The work in [89] employs a distributed extreme learning machine solution to route data among LEO satellites by relying on the forecast of dynamic traffic density. In [90], the authors develop a MARL solution that enables multiple satellites to cooperatively manage their spectrum use. Even though the machine learning based solutions developed in [85–90] are capable of accomplishing certain networking tasks with unknown service needs or unknown communication environment, they are mostly overfitted to the target tasks. In particular, the learning solutions studied in [85–90] are trained to serve specific data needs within specific communication environment. When serving unseen data needs, the learning solution must be retrained within the satellite network, which causes excessive computation costs. As the data needs change constantly in practical applications, the satellite networks must spend a great amount of time and energy on training the machine learning based networking solutions. To reduce such cost for dynamic resilient operations, the notion of meta learning is introduced to generalize the machine learning solutions for different tasks that are expected within a certain distribution. In particular, using the learning schemes generalized by meta training, communication systems can accomplish unseen service tasks that fall within certain distribution, using only a few training examples [91]. However, these meta training mechanisms can still be constrained by preparing the satellites for user needs that fall within specific distribution. Once serving user needs unexpected within this distribution, the satellites still have to spend significant computation time to deliver data services to all the users in need. Since the satellite communication system is deployed to serve users around the globe, whose needs are different, dynamic and change in various patterns, the traditional machine learning algorithms and meta training mechanisms in [85–91] can not be treated as cost effective solutions for the satellite networks.

4.1.2 Contributions

The main contribution of this chapter is a novel dynamic resilient contact plan solution that allows cube satellites to serve diverse user needs that are unpredictable and dynamic. In essence, our key contributions include:

- We consider a practical cube satellite communication system in which the satellites must deliver data service to ground users with dynamic and unpredictable needs. Within this system, the ground gateways will feed the requested data to the satellites. The satellites will, then, distribute the requested data service by routing data to the satellite that is orbiting toward the target users. For effective service coverage, the ground gateways will make plans on how the data will be routed among satellites using the only limited contact chances in the system when feeding these data to satellites. We formulate this contact plan design problem as a dynamic team game, and we solve it using a team optimal strategy which maximizes the fraction of users being served within expected time.
- We propose a novel, distribution-robust meta reinforcement learning (D-RMRL) algorithm that is shown to reach an optimal service coverage of dynamic service needs, with low communication overhead and computation cost. This algorithm reduces the communication overhead in the system by attributing team utility to the gateways, and, thus, reinforcing team beneficial contact plans on each data flow without exchanging the gateway's actions choices and environmental observation. The proposed algorithm also reduces the learning cost in the system with a meta training mechanism that initializes the learning procedure with prior estimation on possible data needs at current service occasion.
- To help the satellites serve all possible users requests, we develop a distribution-robust meta training mechanism. To serve the users whose needs change in different patterns at different service occasions, the proposed D-RMRL algorithm will initiate a meta training procedure at each service occasion for fast convergence of maximal service coverage. Moreover, the distribution-robust meta training mechanism shortens this meta training procedure by starting meta training with existing training experience. In particular, at unseen service occasions, the meta training procedure will start from the estimation of similar service occasions and, thus, can converge faster to the optimal learning initializations within the satellite system.

Simulation results show that proposed distributed D-RMRL algorithm can improve the service coverage and convergence speed by up to 53.3% and 76.3%, compared to traditional multi-agent algorithms. The results also show that using the proposed meta training mechanism, our distributed D-RMRL algorithm can find optimal trajectories in an unseen environment with a 49.1% faster convergence speed, compared to the RL algorithm without meta training capabilities. Moreover, the meta training procedure is shortened by 48% with the proposed meta pre-training scheme. To our best knowledge, *this is the first work that designs on-demand contact plan solutions within resource constrained cube satellite communication system with a low overhead D-RMRL algorithm.*

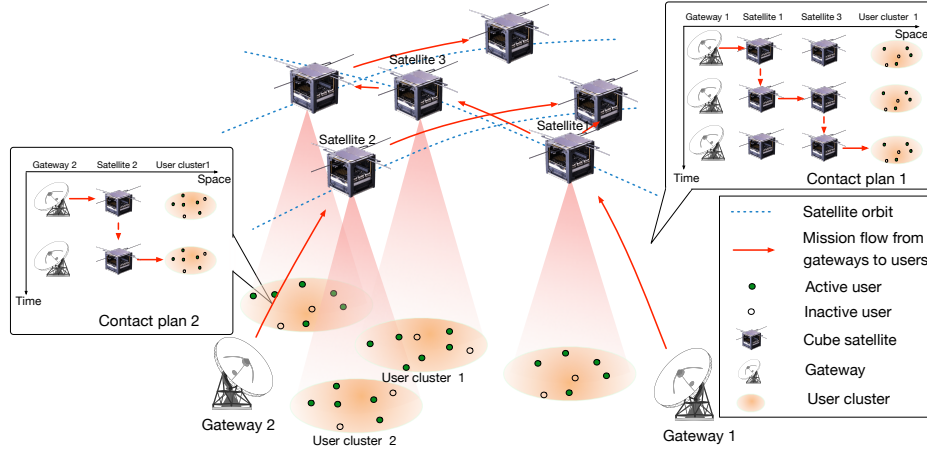


Figure 4.1: Topology of our considered network.

The rest of this chapter is organized as follows. The system model and problem formulation are described in Section II. In Section III, the proposed algorithm is presented. In Section IV, simulation results are analyzed. Finally, conclusions are drawn in Section V.

4.2 System Model and Problem Formulation

Consider a cube satellite network (CSN) consists of a set $\mathcal{L}$ of L LEO cube satellites, and a set $\mathcal{N}$ of N distributed ground gateways. The satellites serve, in the downlink, the radio access requests from a set $\mathcal{U}$ of U terrestrial users, after acquiring all the requested data from the ground gateways. However, while orbiting around the globe, the satellites can only communicate with a ground device (i.e a ground user or a gateway) when they move into the view (i.e. transmission range) of this device. Meanwhile, the ground users are located at different areas each of which can only fall within the service coverage of one cube satellite at a given time, as shown in Fig. 6.1 (each such area is called a *user cluster* with the set of such clusters defined as $\mathcal{C}$). Thus, within the considered system, the satellites will be fed with the requested data by ground gateways when they move into the view of the gateways. After this, the satellites will store the data until this satellite moves into the view of the neighboring satellite that is moving toward the target user clusters, and, then, this neighboring satellite will route the data to this neighbor satellite. All the communication links in our system will use the Ka band (26.5 - 40 GHz), which is a well established millimeter wave (mmW) range suitable for satellite communication, as discussed in [109].

In the considered system, we define a set $\mathcal{V}$ of V contents requested by the ground users. Each content is of equal size ν . We assume user u requests a total of v_u contents in vector $\mathbf{v}_u$ at time slot $t_u \in \Delta$, with $v_u \leq V$. We define $\mathbf{b} = [v_1, v_2, \dots, v_U]$, $\mathbf{t} = [t_1, t_2, \dots, t_U]$, and $\mathbf{v} = [v_1, v_2, \dots, v_U]$ as, respectively, the vector of quantity, occurrence, and content of the users' access requests. We use $\mathbf{z} = [\mathbf{b}, \mathbf{t}, \mathbf{v}]$ to represent *one realization* of the users' access request, and we define $\mathcal{Z}$ as the set of all possible user request realizations in the system. Notice that, the

quantity v_u , occurrence t_u , and composition $\mathbf{v}_u$ of user u 's data requests are assumed to be independent, and they follow unknown distributions $P(\mathcal{Z}|\mathbf{o})$ that vary over different service occasions. A service occasion $\mathbf{o} = [\tau, \mathbf{l}^C]$ is captured by the time of day τ^o and the location of the active clusters $\mathbf{l}^o = [l_i^o]_{i \in \mathcal{C}}$, with l_i^o being the location of cluster $i \in \mathcal{C}$. Here, a possible user request realization that follows user request distribution $P(\mathcal{Z}|\mathbf{o})$ at service occasion $\mathbf{o}$ can be captured by $\mathbf{z}_o = [\mathbf{b}_o, t_o, \mathbf{v}_o]$.

All requested data must be delivered within a time duration that is divided into consecutive time slots in set $\Delta = \{1, \dots, T\}$. Each time slot has a duration of $\tilde{t}$, within which the network topology of the system is considered to be fixed. Using the CSN system, the time required to satisfy the users' data service request depends on several factors, including the *effective contact time* among the satellites and their associated satellites, or gateways, as well as the data rates at the corresponding contact links. Here, the effective contact time between a satellite and its associated devices is the time duration that this satellite travels within the view of this target device and transmits the requested data. The effective contact time is determined by the predictable topology of the ISDN system, as well as the limited contact chances within the system. Next, the network topology and resource limitation is modeled as an extended time-evolving graph. Then, the communication data rates in the system are analyzed.

4.2.1 Network Topology with Resource Limitation

The network topology limited by contact, spectrum and storage resource in the considered system is captured by an extended time-evolving directed graph $\mathcal{G}(\mathcal{W}, \mathcal{E})$. This graph consists of T layers each of which corresponds to a network topology at one time slot $t \in \Delta$. Within this graph, the vertices $\mathcal{W} = \mathcal{L} \cup \mathcal{N} \cup \mathcal{U}$ correspond to the devices in the system including satellites, gateways, and users. $\mathcal{E} = \mathcal{E}_c \cup \mathcal{E}_s$ is the set of edges that connect vertices from $\mathcal{W}$. Here, $\mathcal{E}_c = \{(e_1(t), e_2(t)) | e_1(t) \in \mathcal{N} \cup \mathcal{L}, e_2(t) \in \mathcal{L} \cup \mathcal{U}, e_1(t) \text{ is in the view of } e_2(t), t \in \Delta\}$ captures the edges connecting devices $e_1(t)$ and $e_2(t)$, i.e. at slot t , $e_1(t)$ can transmit data to $e_2(t)$. Meanwhile, the edges in $\mathcal{E}_s = \{(e_1(t), e_1(t)) | t \in \Delta, e_1(t) \in \mathcal{L} \cup \mathcal{N} \cup \mathcal{U}\}$ captures the edges that keep data on device $e_1(t)$, meaning that $e_1(t)$ is still capable of storing data at slot t .

The *mission* in the system is defined as the process of delivering the requested data from the gateways to users. The *mission flow* initiated at gateway n to serve data requests at user u , i.e., the flow of data delivered to user u from gateway n , is modeled as a tuple $f_{n,u} = (t_s(f_{n,u}), t_d(f_{n,u}), v(f_{n,u}))$, with $t_s(f_{n,u})$, and $t_d(f_{n,u})$ being, respectively, the start and end time of the flow. $v(f_{n,u})$ is the volume of the flow in bits, with $\sum_{n \in \mathcal{N}} v(f_{n,u}) = v_u \nu$. The set of all data service mission flows is $\mathcal{F} = \{f_{n,u} | n \in \mathcal{N}, u \in \mathcal{U}\}$.

These data service missions are scheduled by the *contact plans* defined as $x(e_k(t), e_{k+1}(t), f_{n,u})$. Here, $x(e_k(t), e_{k+1}(t), f_{n,u}) = 1$ means that edge $(e_k(t), e_{k+1}(t))$ is assigned to mission flow $f_{n,u}$, otherwise, $x(e_k(t), e_{k+1}(t), f_{n,u}) = 0$. We define a contact plan vector in the system as $\mathbf{x} = [x(e_k(t), e_{k+1}(t), f_{n,u}), \forall (e_k(t), e_{k+1}(t)) \in \mathcal{E}, f_{n,u} \in \mathcal{F}, t \in \Delta, k \in \mathcal{K}]$. Meanwhile, the vector of contact plans that are scheduled by gateway n at time slot t is captured in $\mathbf{x}_{n,t} = [x(e_k(t), e_{k+1}(t), f_{n,u}), \forall (e_k(t), e_{k+1}(t)) \in \mathcal{E}, k \in \mathcal{K}, u \in \mathcal{U}]$, with $\mathcal{K}$ being the set of all

possible vertices on the flow.

The maximum data volume of mission flow $f_{n,u}$ on communication edge $(e_k(t), e_{k+1}(t))$, in time slot t , is given by

$$y(e_k(t), e_{k+1}(t), f_{n,u}) = c_{e_k, e_{k+1}}(t) \tilde{t}, \quad (4.1)$$

where $c_{e_k, e_{k+1}}(t)$ is the data rate on the communication link between devices e_k and e_{k+1} at slot t , and is defined in the next subsection. Meanwhile, the maximum data volume on a storage edge in $\mathcal{E}_e$ is the storage capacity of this device. In particular, the storage capacity of satellite l is denoted as Γ_l , while the gateways are assumed to have a sufficient large storage capacity, i.e. $\Gamma_n \rightarrow \infty$ for all $n \in \mathcal{N}$.

4.2.2 Communication Performance Analysis

In the considered system, a TDMA scheme is adopted to support directional transmissions at the significantly high path loss mmW band, while maintaining low complexity designs for transceivers [128]. Note that, in this model, as done in [111], the interference caused by ground gateways on the inter-satellite communication links, as well as the interference caused by the satellites on the ground users, will be considered negligible due to the limited transmit power of ground gateways and satellites, as well as the long distance between the ground device and satellites.

4.2.2.1 Inter-Satellite Link Channel Model

The path loss (in dB) of the inter-satellite link between transmitter satellite l_1 and receiver satellite l_2 in slot t is given by

$$L_{l_1, l_2}^{\text{IS}}(t) = \left(\frac{p_c}{4\pi d_{l_1, l_2}(t) f_c} \right)^2, \quad (4.2)$$

where f_c is the carrier frequency of the communication link, $d_{l_1, l_2}(t)$ is the distance between satellite l_1 and satellite l_2 in slot t , and p_c is the speed of light. Thus, the signal-to-noise ratio (SNR) of the inter-satellite link between satellite l_1 and satellite l_2 in slot t , will be given by [129]

$$\gamma_{l_1, l_2}^{\text{IS}}(t) = \frac{P_{l_1} G_{t, l_1} G_{r, l_2} L_{l_1, l_2}^{\text{IS}}(t)}{\kappa T_s \left(\frac{E_b}{N_0} \right)_{\text{req}} M}, \quad (4.3)$$

where P_{l_1} is the transmit power of satellite l_1 assumed to be equal for all satellites. G_{t, l_1} , and G_{r, l_2} , are, respectively, the transmit antenna gain of satellite l_1 and the receive antenna gain of satellite l_2 . κ is the Boltzmann's constant (in JK^{-1}) and T_s is the total system noise temperature (in Kelvin). $\left(\frac{E_b}{N_0} \right)_{\text{req}}$ is the required ratio of received energy-per-bit to noise-density, N_0 is the noise power spectral density. M is the link margin. Thus, the data rate at link between satellite l_1 and l_2 will then be $c_{l_1, l_2}(t) = B_S \log(1 + \gamma_{l_1, l_2}^{\text{IS}}(t))$, with B_S being the bandwidth allocated to inter-satellite link between satellite l_1 and satellite l_2 .

4.2.2.2 Satellite-Ground Link Channel Model

The path loss (in dB) of the satellite-ground link between satellite $l \in \mathcal{L}$ and ground user $u \in \mathcal{U}$ in slot t is given by

$$L_{l,u}^{\text{SG}}(t) = L_{l,u}^{\text{F}}(t) L_{l,u}^{\text{N}}(t), \quad (4.4)$$

where $L_{l,u}^{\text{F}}(t) = \left(\frac{p_c}{4\pi d_{l,u}(t) f_c} \right)^2$ is the free space path loss, with $d_{l,u}(t)$ is the distance between satellite l and ground user u in slot t . $L_{l,u}^{\text{N}}(t) = \delta_{l,u}(t) \delta_{l,u}^{\text{R}}(t)$ is the rain attenuation on the link, with $\delta_{l,u}(t)$ being the equivalent effective slant-path length, which is defined as [130]

$$\delta_{l,u}(t) = \begin{cases} \frac{h_R(t) - h_S}{\sin \theta_{l,u}(t)}, & \text{if } \theta_{l,u}(t) \geq 5^\circ, \\ \frac{2(h_R(t) - h_S)(t)}{\left(\sin^2 \theta_{l,u}(t) + \frac{2(h_R(t) - h_S)(t)}{R_e} \right)^{0.5} + \sin \theta_{l,u}(t)}, & \text{otherwise,} \end{cases} \quad (4.5)$$

in which $h_R(t)$ is the rain height at slot t defined by [131]

$$h_R(t) = \begin{cases} 5, & \text{if } \beta_{l,u}(t) \leq 23^\circ, \\ 5 - 0.075 (\beta_{l,u}(t) - 23), & \text{otherwise,} \end{cases} \quad (4.6)$$

with $\beta_{l,u}(t)$ being the difference between the longitude of satellite l and ground user u . h_S is height above mean sea level of the ground station on the link. R_e , and $\theta_{l,u}(t)$ are the effective radius of the Earth (8500 km) and the elevation angle of the satellite-ground link between satellite l and ground user u at slot t , respectively. Then, based on [132], the rain attenuation per kilometer is defined as $\delta_{l,u}^{\text{R}}(t) = \rho R_{l,u}(t)^\eta$, with $R_{l,u}(t)$ (mm/h) being the rainfall intensity on the link between satellite l and ground user u , at slot t , ρ and η being the coefficients as defined in [132]. Meanwhile, the path loss (in dB) of the satellite-ground link between satellite $l \in \mathcal{L}$ and ground gateway $n \in \mathcal{N}$ in slot t , $L_{n,l}^{\text{GS}}(t) = L_{n,l}^{\text{F}}(t) L_{n,l}^{\text{N}}(t)$, is defined in the same way of $L_{l,m}^{\text{GS}}(t)$, with $L_{n,l}^{\text{F}}(t)$ being the free space path loss, and $L_{n,l}^{\text{N}}(t)$ being the rain attenuation on the link.

Thus, the SNR of the satellite-ground link between satellite l and ground user u in slot t , will thus be [133]

$$\gamma_{l,u}^{\text{SG}}(\mathbf{x}, t) = \frac{P_l G_{t,l} G_{r,u} L_{l,u}^{\text{SG}}(t)}{\sigma^2}, \quad (4.7)$$

where $G_{t,l}$, and $G_{r,u}$, are, respectively, the transmitting antenna gain on the satellite l and the receiving antenna gain on ground user u . σ^2 is the noise power on the link. Meanwhile, the SNR of the satellite-ground link between ground gateway n and satellite l in slot t , will be [133]

$$\gamma_{n,l}^{\text{GS}}(\mathbf{x}, t) = \frac{P_n G_{t,n} G_{r,l} L_{n,l}^{\text{GS}}(t)}{\sigma^2}, \quad (4.8)$$

with P_n being the transmit power of gateway n , which is assumed to be equal for all gateways. $G_{t,n}$, and $G_{r,l}$, are, respectively, the transmit antenna gain on gateway n and the receive antenna gain on satellite l . Thus, the data rate at the link between ground gateway n and satellite l in slot t will be $c_{n,l}(t) = B_C \log(1 + \gamma_{n,l}^{\text{GS}}(\mathbf{x}, t))$, with B_C being the bandwidth allocated to the satellite-ground link between ground gateway n and satellite l .

4.2.3 Utility Function Model

Given that the goal of the CSN system is to effectively serve all the access requests from ground users, the utility function is defined to capture the portion of users being served within expected time. In particular, a data service mission is referred as successfully scheduled if all of the requested data is delivered to the target users before the scheduled end time, that is $t_d(f_{n,u}) \leq T$, $\forall f_{n,u} \in \mathcal{F}$. The utility achieved by a gateway $n \in \mathcal{N}$ is, thus, defined as a binary variable $\mu_{n,u} \in \{0, 1\}$, which denotes whether the mission flow $f_{n,u}$ is successfully scheduled or not, with $\mu_{n,u} = 1$ indicating that $f_{n,u}$ is successfully scheduled. Thus, the completion of mission $f_{n,u}$ can be given by

$$\mu_{n,u}(\mathbf{x}_{n,t} | \mathbf{x}_{n,t-1}, \mathbf{o}) = \mathbb{1}_{\{y_u^R(\mathbf{x}_{n,t}) \geq v_{u,\mathbf{o}} \nu, \text{ and, } y_u^R(\mathbf{x}_{n,t-1}) < v_{u,\mathbf{o}} \nu\}}. \quad (4.9)$$

where $y_e^R(\mathbf{x}_{n,t}) = \sum_{t'=1}^t \sum_{k \in \mathcal{K}} y(e_k(t'), e_{k+1}(t'), f_{n,u}) x(e_k(t'), e_{k+1}(t'), f_{n,u}) \mathbb{1}_{e_{k+1}(t')=u}$ is the amount of data received by user u until slot t , using contact plan captured by $\mathbf{x}_{n,t}$. $v_{u,\mathbf{o}}$ is quantity of data request at user u within a realization that is sampled at service occasion $\mathbf{o}$. $\mathbb{1}_\chi = 1$ when χ is true, otherwise, $\mathbb{1}_\chi = 0$. Here, $\mathbb{1}_{\{y_u^R(\mathbf{x}_{n,t}) \geq v_{u,\mathbf{o}} \nu, \text{ and, } y_u^R(\mathbf{x}_{n,t-1}) < v_{u,\mathbf{o}} \nu\}}$ represents that the contents requested by user u have been delivered to this user at slot t , with the contact plan designed by gateway n .

4.2.4 Problem Formulation

In our system model, the gateways make plans on how the request data will be routed toward the target users using the limited contact chances in $\mathcal{G}$, while feeding the data to satellites. The gateways can predict the available contact chances among satellites in the system, as the network topology is predictable. Then, the satellites can distribute the requested data as instructed by the gateways. However, the contact plan decisions made by one gateway will directly affect the ones made by other gateways, as none of the satellites is willing to transmit the same contents to the satellites, or deliver data to the users that are already served. Thus, the gateway will, cooperatively, serve partial of the user requests to maximize service coverage using the limited contact chances and spectrum resources in the system. In order to capture the correlated decision making processes on the gateways, we formulate a *dynamic team game* (DTG) [134] defined by the tuple $\langle \mathcal{N}, \mathcal{A}, \mathcal{S}, G, \boldsymbol{\pi}, T \rangle$, where

- $\mathcal{N}$ is the set of gateways that act as the players in the game.
- $\mathcal{A}$ is the set of available actions for the gateways, which is the set of devices that can be contacted. The action chosen by gateway n at step t of the game is denoted as $\mathbf{a}_{n,t} = [a_{n,1,t}, a_{n,2,t}, \dots, a_{n,U,t}]$, with $a_{n,u,t}$ being the device that is chose to store or transmit the data within mission flow $f_{n,u}$, i.e. $x(e(t), a_{n,u,t}, f_{n,u}) = 1$. The set of all gateways' actions at step t is denoted in $\mathbf{a}_t = [\mathbf{a}_{1,t}, \mathbf{a}_{2,t}, \dots, \mathbf{a}_{N,t}]$.
- $\mathcal{S} = \{\mathbf{s}_0, \mathbf{s}_1, \dots, \mathbf{s}_T\}$ is the set of states at the gateways, with $\mathbf{s}_t = [\mathbf{s}_{1,t}, \mathbf{s}_{2,t}, \dots, \mathbf{s}_{N,t}]$

being the set of all the states of the gateways at step t . The state of gateway n at step t , i.e. $\mathbf{s}_{n,t}$, is the locations of the mission flows in the system, thus, $\mathbf{s}_{n,t} = \mathbf{a}_{n,t-1}$.

- $G(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) = \sum_{n \in \mathcal{N}} \sum_{u \in \mathcal{U}} \mu_{n,u}(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$ is the utility function of the gateways measuring the effective mission completion achieved by all of the gateways in the system at service occasion $\mathbf{o}$, and is known as the *team reward*.
- $\boldsymbol{\pi} = [\pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})]_{n \in \mathcal{N}, t \in \Delta}$ is the vector of strategies of all gateways, where $\pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$ is the probability that gateway n selects contact plan $\mathbf{a}_{n,t} \in \mathcal{A}$ at state $\mathbf{s}_{n,t}$, and service occasion $\mathbf{o}$. Here, the probabilistic strategies are implemented by gateways choosing different contact plans with certain frequencies. When serving user requests, the gateways will sample one possible contact plans from strategy $\boldsymbol{\pi}$.
- T is the terminal step of the game before which all the requested data must reach the target user. Here, a data service mission can terminate before step T , since the satellites may deliver all the requested data to the target users before step T .

In the formulated game, the gateways select contact plans to deliver data to ground users, based on their strategy $\boldsymbol{\pi}$. The team reward is measured as the feedback to the gateways' action (i.e. contact plan) selections, as every gateway in the system seeks to maximize the overall service coverage. In particular, given that the primary objective of the CSN system is to serve all user requests within expected time using the limited contact chances, the utility of the gateways in the TG is defined as the overall mission completion achieved by all gateways. The goal of the gateways is to find the strategies that lead to the maximal team utility. To solve this game, we use the concept of *team optimal strategy* [134]. The concept of a team optimal strategy $\boldsymbol{\pi}^*$ is defined as

Definition 3. A *team optimal strategy* in the considered team game is the strategy $\boldsymbol{\pi}^*$ that satisfy the following condition

$$\overline{G}(\boldsymbol{\pi}^* | \mathbf{o}) \geq \overline{G}(\boldsymbol{\pi} | \mathbf{o}), \forall \boldsymbol{\pi} \in \Psi, \quad (4.10)$$

where $\overline{G}(\boldsymbol{\pi} | \mathbf{o}) = \sum_{n=1}^N \sum_{\mathbf{a}_{n,t} \in \mathcal{A}} (\sum_{u \in \mathcal{U}} \mu_{n,u}(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})) \pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$ is the expected final utility with strategy $\boldsymbol{\pi}$, Ψ is the strategy space of the gateways.

Note that, a team optimal strategy is a Nash equilibrium of the dynamic team game, as every agent in the system chooses strategies that maximize its own utility and will not achieve any improvement with any other strategies. Meanwhile, the team optimal strategy will guarantee a maximum expected mission completion in the CSN system, and is the optimal solution of the following problem

$$\max \overline{G}(\boldsymbol{\pi} | \mathbf{o}), \quad (4.11)$$

$$\text{s. t. } \sum_{u \in \mathcal{U}} \sum_{n \in \mathcal{N}} \varepsilon_e(\mathbf{a}_{n,u,t}, \mathbf{s}_{n,u,t}) \leq \Gamma_e, \forall e \in \mathcal{L} \cup \mathcal{N}, t \in \Delta \quad (4.11a)$$

$$\sum_{t \in \Delta} \sum_{n \in \mathcal{N}} y_n^T(\mathbf{a}_{n,u,t}, \mathbf{s}_{n,u,t}) = v_u \nu, \forall f_{n,u} \in \mathcal{F}, \quad (4.11b)$$

$$\varepsilon_e(a_{n,u,t}, s_{n,u,t}) - \varepsilon_e(a_{n,u,t-1}, s_{n,u,t-1}) + y_e^R(a_{n,u,t}, s_{n,u,t}) - y_e^T(a_{n,u,t}, s_{n,u,t}) = 0, \forall n \in \mathcal{N}, u \in \mathcal{U}, e \in \mathcal{L} \cup \mathcal{N}, \quad (4.11c)$$

$$\varepsilon_u(a_{n,u,t}, s_{n,u,t}) - \sum_{t'=1}^t y_u^R(a_{n,u,t'}, s_{n,u,t'}) = 0, \forall n \in \mathcal{N}, u \in \mathcal{U}, \quad (4.11d)$$

$$\sum_{n \in \mathcal{N}} \sum_{u \in \mathcal{U}} \sum_{e' \in \mathcal{W}} (\mathbb{1}_{s_{n,u,t}=e, a_{n,u,t}=e'} + \mathbb{1}_{s_{n,u,t}=e', a_{n,u,t}=e}) \leq 1, e(t) \in \mathcal{W}, \quad (4.11e)$$

where $\varepsilon_e(a_{n,u,t}, s_{n,u,t}) = \sum_{t'=0}^t [y_e^R(a_{n,u,t'}, s_{n,u,t'}) - y_e^T(a_{n,u,t'}, s_{n,u,t'})]$ is the amount of data on mission flow $f_{n,u}$ that is stored at device e , in time slot t . Here, $y_e^R(a_{n,u,t}, s_{n,u,t}) = y(s_{n,u,t}, a_{n,u,t}, f_{n,u}) \mathbb{1}_{a_{n,u,t}=e}$, and $y_e^T(a_{n,u,t}, s_{n,u,t}) = y(s_{n,u,t}, a_{n,u,t}, f_{n,u}) \mathbb{1}_{s_{n,u,t}=e}$ are, respectively, the amount of data on flow $f_{n,u}$ that device e received and transmitted at slot t . Here, (6.5a) means that, within the considered time duration, the amount of data stored on a device should not exceed this device's storage capacity. (6.5b)-(6.5d) are constraints on the balance of mission flows over the communication links in the system. In particular, (6.5b) indicates that the requested data of mission $f_{n,u}$ must be cooperatively delivered by the gateways. (6.5c) captures that the fact, at time slot t , the amount of data transmitted and stored by a device is the sum of the amount of data received by this device within this slot, and the amount of data stored at this device before slot t . (6.5d) indicates that, at time slot t , the data stored at a ground user is the sum of data this user has received from the BSs in the system. Meanwhile, (6.5e) captures limitations on feasible communication contacts and implies that a device can only either receive, or transmit on one communication link in the system.

However, we notice that the contact plan design problem in (6.5) cannot be solved with traditional optimization solutions, such as branch and bound or linear programming. This stems from the fact that the objective function in (6.5) is unpredictable with the user requests being dynamic and change in unknown patterns. The RL algorithms can help the CSN system serve unknown user needs within unknown environments. In particular, an RL approach can help the gateways learn to plan mission flows in the unknown environment using trial and error. However, traditional RL algorithms, such as Q learning, policy gradient, and echo state networks [135–137], are also not suitable for solving the considered problem in (6.5) for multiple reasons. First, using traditional RL algorithms, the gateways must know each other's strategies, environmental observation, and rewards for effective cooperation, which leads to a high overhead in terms of computation and communication. Meanwhile, the trial and error scheme used in RL can introduce excessive learning cost on the gateways. Thus, to solve the time sensitive contact plan design problem (6.5) for dynamic, unpredictable data access needs, we propose a distributed distribution-robust meta reinforcement learning solution with pre-training capability.

The proposed distributed D-RMRL algorithm allows the gateways to learn contact plan design solutions with a trial and error based RL scheme, and reduces excessive learning cost with value decomposition, meta learning, and pre-training techniques. In particular, the proposed solution decomposes problem (6.5) using the value decomposition assumption, such that the team optimal strategy is learned at each gateway by relying on the gateways' individual strategies and observation on the environment. It starts a learning procedure shortened by meta trained learning initializations that are updated to be close to the all team optimal strategies at current service occa-

sion. Then, the meta training procedure is shortened by a pre-training scheme, with which it starts with meta training experience from similar service occasions. The proposed distributed D-RMRL algorithm is introduced in the next section.

4.3 Proposed Distribution Robust Meta Reinforcement Learning Algorithm

We now introduce the distributed D-RMRL algorithm, which merges the concept of value decomposition [138], model agnostic meta-learning [139], pre-training [139], with the actor-critic RL framework. The value decomposition assumption helps the actor-critic RL algorithm find the team optimal strategy for problem (6.5) in a distributed way at each gateway, with reduced communication and computation overhead. The meta training scheme further reduces the learning cost on gateways with a generalized learning procedure that starts with primary estimations on all the possible user requests at current service occasion. The pre-training scheme further shortens the meta training procedure with a parameter transfer technique. Next, we first introduce the components of the proposed distributed D-RMRL algorithm, followed by an explanation on how the distributed D-RMRL algorithm learns to solve problem (6.5) at each gateway in the system, using the value decomposition. We, then, explain how the distributed D-RMRL is meta-trained for easier convergence at different data request realizations. Finally, we introduce how the pre-training scheme improves the out-of-distribution performance of the proposed distributed D-RMRL.

4.3.1 Distribution-Robust Meta Reinforcement Learning Components

The proposed distributed D-RMRL algorithm consists of six components, including the players, actions, states and rewards $r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o})$ (i.e. utility functions $r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) = G(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o})$) of the DTG, a value function $V_{\theta_c}(\mathbf{s}_t)$ defined by a deep neural network parametrized by θ_c , as well as a strategy function $\theta_{a,n}$ defined by a deep neural network parametrized by $\theta_{a,n}$. Here, value function $V_{\theta_c}(\mathbf{s}_t)$ estimates the future rewards that can be expected by the gateways at different states. In particular, this value function takes gateways' state as input and outputs estimated future reward $\sum_{n \in \mathcal{N}} \sum_{\kappa=t}^T \gamma^{t-\kappa+1} r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o})$, with γ being the discounted factor. Policy function $\pi_{\theta_{a,n}}$ takes gateway n ' state as input and outputs strategy $\pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$, that defines the probability that gateway n chooses to transmit data to (or storage data at) devices in $\mathbf{a}_{n,t}$, at state $\mathbf{s}_{n,t}$ of service occasion $\mathbf{o}$.

Using distributed D-RMRL, the gateways learn to behave within an unknown wireless environment from their *experiences* on routing and delivering data with contact plans sampled from strategy π . In particular, at service occasion $\mathbf{o}$, the gateways will keep selecting actions at different states to serve ground users, until all the user requests are satisfied or the terminal step of DTG is reached. The gateways can evaluate their action choices with team utility $r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o})$, and then, record

its experience defined in vector $\varepsilon_n = [\mathbf{a}_{n,1}, \mathbf{s}_{n,1}, r(\mathbf{a}_1 | \mathbf{s}_1, \mathbf{o}), \dots, \mathbf{a}_{n,t}, \mathbf{s}_{n,t}, r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o})]$. The gateways will update the value function toward accurate reward estimation, and update the policy functions toward maximal overall on-time data delivery with the following procedure, after collecting one experience. Using such a mini-batch training procedure, the D-RMRL algorithm reduces the variance on the system's learning performance caused by the action sampling, and the need on storing a big dataset in the system.

4.3.2 Reinforcement Learning for Strategy Optimization

Using the distributed D-RMRL algorithm, the gateways aim at solving problem (6.5) using a value function with minimal temporal difference (TD) error on achievable reward estimation defined in [140]

$$A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) = Nr(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) + \gamma V_{\theta_c}(\mathbf{s}_{t+1}) - V_{\theta_c}(\mathbf{s}_t) \quad (4.12)$$

where TD error $A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o})$ is called the *team advantage* of the gateways at state $\mathbf{s}_t$ with action $\mathbf{a}_t$. This function measures the difference between the gateways' achieved reward and estimated reward, and reveals how actions in $\mathbf{a}_t$ are better than other action selections at current state $\mathbf{s}_t$. Moreover, to solve (6.5), the gateways also seek to achieve policies that maximize the expected rewards estimated by $\sum_{\mathbf{a}_t \in \mathcal{A}} A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) \prod_{n=1}^N \pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$. Thus, the gateways will improve value function $V_{\theta_c}(\mathbf{s}_t)$ toward error free estimation by updating θ_c in the opposite direction of the gradient of team advantage $A^2(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o})$, and optimize policy functions $\pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$ by updating $\theta_{a,n}$ in the direction of the gradient of expected reward $\sum_{\mathbf{a}_t \in \mathcal{A}} A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) \prod_{n=1}^N \pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$.

However, deviation on the team advantage can cause significant communication and computation overhead on the gateways, this is because that 1) the derivative of the accumulated policies $\prod_{n=1}^N \pi_n(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o})$ has high computation complexity, and 2) the derivative requires the gateways to know the global states, the actions taken by all gateways, and the achieved team advantages at every step t . Therefore, we introduce the value decomposition assumption given by

$$V_{\theta_c}(\mathbf{s}_t) = \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t}), \quad (4.13)$$

in which $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t})$, parametrized by $\theta_{c,n}$, is the individual value function attributed to gateway n that only reinforces team beneficial actions on this gateway. In particular, using this assumption, team advantage $A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o})$ can, then, be expressed as

$$\begin{aligned} A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) &= Nr(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_c}(\mathbf{s}_{n,t+1}) - \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t}) \\ &= \sum_{n \in \mathcal{N}} \tilde{A}_n(\mathbf{a}_{n,t}, \mathbf{s}_{n,t}, \mathbf{o}). \end{aligned} \quad (4.14)$$

Here, $\tilde{A}_n(\mathbf{a}_{n,t}, \mathbf{s}_{n,t}, \mathbf{o}) = r(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) + \gamma \tilde{V}_{\theta_c}(\mathbf{s}_{n,t+1}) - \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t})$ is the individual advantage, or the local estimation error at gateway n . Thus, the estimated overall reward (i.e. team value

Algorithm 3 Distributed D-RMRL algorithm for contact plan design at one realization of user access request at one service occasion.

Require: Service occasion $\mathbf{o}$, user locations, time constraints.

Ensure: Initialize value functions $\tilde{V}_{\theta_{c,n}^{(1)}} = \tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}$, and policy functions $\pi_{\theta_{a,n}^{(1)}} = \pi_{\theta_{a,n}^{(1)}, \mathbf{o}}$, for $n \in \mathcal{N}$.

- 1: **for** distributed D-RMRL training epoch $i = 1 : I$ **do**
 - 2: The gateways carry out an experience by generating a sequence of actions that are randomly selected based on current policy functions.
 - 3: Calculate advantage A based on (5.8).
 - 4: **for** Each gateway $n = 1 : N$ **do**
 - 5: Update individual value function with (5.11).
 - 6: Calculate individual advantages $\tilde{A}_n(\mathbf{a}_{n,t}, \mathbf{s}_{n,t}, \mathbf{o})$.
 - 7: Update policy function with (5.11).
 - 8: **end for**
 - 9: **end for**
-

$V_{\theta_c}(\mathbf{s}_t)$), and estimation error (i.e. team advantage $A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o})$) are distributed to each gateway thus allowing the gateways to update their individual value functions $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t})$ using

$$\begin{aligned}
 \theta_{c,n}^{(i+1)} &= \theta_{c,n}^{(i)} - \alpha_c^{(i)} \nabla_{\theta_{c,n}} \sum_{t=1}^T A^2(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) \\
 &= \theta_{c,n}^{(i)} - \alpha_c^{(i)} \nabla_{\theta_{c,n}} \sum_{t=1}^T \left(Nr(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_c}(\mathbf{s}_{n,t+1}) - \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t}) \right)^2 \\
 &= \theta_{c,n}^{(i)} + 2\alpha_c^{(i)} \sum_{t=1}^T A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) \nabla_{\theta_{c,n}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t}), \tag{4.15}
 \end{aligned}$$

with $\theta_{c,n}^{(i)}$ being the value function parameter of gateway n and $\alpha_c^{(i)}$ being the value update step size, at iteration i of the distributed D-RMRL training. Thus, using the value decomposition assumption in (5.9), the derivative of team advantage for the update of value functions is simplified by deviating the local individual value functions relying only on the gateways' local states $\mathbf{s}_n$. Also, note that, $Nr(\mathbf{a}_t | \mathbf{s}_t, \mathbf{o}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_c}(\mathbf{s}_{n,t+1})$ is a supervised term whose partial derivative will not be counted in the gradient update step defined in (5.11). Then, based on the policy gradient theorem [137], the update on gateway n 's policy function parameters is given by

$$\theta_{a,n}^{(i+1)} = \theta_{a,n}^{(i)} + \alpha_a^{(i)} \sum_{t=1}^T \tilde{A}_n(\mathbf{a}_{n,t}, \mathbf{s}_{n,t}, \mathbf{o}) \nabla_{\theta_{a,n}} \log \pi_{\theta_{a,n}}(\mathbf{a}_{n,t} | \mathbf{s}_{n,t}, \mathbf{o}), \tag{4.16}$$

where $\theta_{a,n}^{(i)}$ is the policy function parameter at gateway n and $\alpha_a^{(i)}$ is the policy update step size, at the i -th iteration of distributed D-RMRL. Thus, the deviation on the expected reward for the update of policy functions is simplified by relying only on the gateways' local states $\mathbf{s}_{n,t}$, actions $\mathbf{a}_{n,t}$.

As summarized in Algorithm 3, to service a user request realization at current service occasion $\mathbf{o}$, the gateways will start by delivering data with contact plans sampled from meta trained initial

policy functions $\pi_{\theta_{a,n}^{(1)}, \mathbf{o}}$ and value functions $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}$. Then, the gateways exchange their achieved service coverage (i.e. achieved reward) and estimated reward (i.e. values of $\tilde{V}_{\theta_{c,n}^{(1)}}(\mathbf{s}_{n,t})$) for calculating the team utility, and team advantage $A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o})$. Note that the proposed algorithm uses a mini-batch training mechanism, with which the gateways update policies and reward estimations with a whole experience ε_n for reduced variance on sampling, and cost on dataset storage. Thus, using their collected experience on serving user data requests and achieved advantage values, the gateways will update their value and policy parameters, independently, with (5.11) and (5.11). Such updates will be repeated until the convergence to a team optimal strategy is reached. Such initializations are trained to be close to the optimal policies and reward estimations at all possible user request realizations in $P(\mathcal{Z}|\mathbf{o})$ using stochastic gradient descent as presented in the following subsection.

4.3.3 Meta Trained Learning Initialization for Task Robust Performance

As previously mentioned, the goal of meta training is to prepare the satellite network for serving all possible user requests at one service occasion. In particular, the meta training procedure finds the learning initializations, i.e. initial policy functions $\pi_{\theta_{a,n}^{(1)}, \mathbf{o}}$ and value functions $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}$, that are already close to the optimal strategies and value estimations within all possible user request realizations at current service occasion. While solving the studied problem, the satellite system will sample user request realizations and implement meta training using its experience on serving these user requests. In essence, the goal of meta training is formulated as

$$\min_{\theta_c^{(1)}, \theta_a^{(1)}} \sum_{\mathbf{z}_j \sim P(\mathcal{Z}|\mathbf{o})} \sum_{n=1}^N L_{c,n}(\tilde{V}_{\theta_{c,n}^{(1)}}, \mathbf{z}_j) + L_{a,n}(\pi_{\theta_{a,n}^{(1)}}, \mathbf{z}_j), \quad (4.17)$$

where $L_{c,n}(\tilde{V}_{\theta_{c,n}^{(1)}}, \mathbf{z}_j)$ captures the distance between agent n ' initial values and optimal values, $L_{a,n}(\pi_{\theta_{a,n}^{(1)}}, \mathbf{z}_j)$ captures the distance between agent n ' initial policies and optimal policies, for user data request realization $\mathbf{z}_j \sim P(\mathcal{Z}|\mathbf{o})$. $\theta_c^{(1)} = [\theta_{c,1}^{(1)}, \theta_{c,2}^{(1)}, \dots, \theta_{c,N}^{(1)}]$ and $\theta_a^{(1)} = [\theta_{a,1}^{(1)}, \theta_{a,2}^{(1)}, \dots, \theta_{a,N}^{(1)}]$ are, respectively, the vector of initial value function parameters and policy function parameters chosen at all gateways. In the considered CSN system, the gateways will collect experiences on learning to serve different user requests at current service occasion, and they keep updating their learning initializations by reducing loss function with stochastic gradient descent (SGD).

To reduce the cost on computing $L_{c,n}(\tilde{V}_{\theta_{c,n}^{(1)}}, \mathbf{z}_j)$ and $L_{a,n}(\pi_{\theta_{a,n}^{(1)}}, \mathbf{z}_j)$ that is caused by the search of optimal policies and values, a function approximation is implemented. In particular, the distance between the agents' initial values and optimal values, $L_{c,n}(\tilde{V}_{\theta_{c,n}^{(1)}}, \mathbf{z}_j)$, is approximated by the evaluation on the update direction of $\theta_{c,n}$ as defined in

$$\tilde{L}_{c,n}(\tilde{V}_{\theta_{c,n}^{(2)}}, \mathbf{z}_j) = \sum_{t=1}^T \left(r(\mathbf{a}_{n,t}^{(\varepsilon'_{n,j})} | \mathbf{s}_{n,t}^{(\varepsilon'_{n,j})}, \mathbf{o}) + \gamma \tilde{V}_{\theta_{c,n,j}^{(2)}}(\mathbf{s}_{n,t+1}^{(\varepsilon'_{n,j})}) - \tilde{V}_{\theta_{c,n,j}^{(2)}}(\mathbf{s}_{n,t}^{(\varepsilon'_{n,j})}) \right)^2, \quad (4.18)$$

in which $\theta_{c,n,j}^{(2)} = \theta_{c,n}^{(1)} + \alpha_c \nabla_{\theta_{c,n}^{(1)}} \sum_{t=1}^T A^2 \left(\mathbf{a}_t^{(\epsilon_{n,j})}, \mathbf{s}_t^{(\epsilon_{n,j})}, \mathbf{o} \right)$ captures the updated value function parameters on gateway n . In particular, the value functions $\tilde{V}_{\theta_{c,n}^{(1)}}$ and policy function $\pi_{\theta_{a,n}^{(1)}}$ are updated by relying on the experience, $\epsilon_{n,j}$, which the satellite system collects on serving user requests in realization $\mathbf{z}_j$, using $\tilde{V}_{\theta_{c,n}^{(1)}}$ and $\pi_{\theta_{a,n}^{(1)}}$. Then the updated value functions $\tilde{V}_{\theta_{c,n}^{(2)}}$ are evaluated with their errors on estimating future rewards, i.e. $\tilde{L}_{c,n} \left(\tilde{V}_{\theta_{c,n}^{(2)}} \left(\mathbf{z}_j \right) \right)$, which are calculated based on the experience $\epsilon'_{n,j}$ the satellite system collects on serving user requests in realization $\mathbf{z}_j$, using $\tilde{V}_{\theta_{c,n}^{(2)}}$ and $\pi_{\theta_{a,n}^{(2)}}$. Here, $\mathbf{a}_{n,t}^{(\epsilon'_{n,j})}$, and $\mathbf{s}_{n,t}^{(\epsilon'_{n,j})}$ are, respectively, the t -th action and state of gateway n within experience $\epsilon'_{n,j}$. Moreover, the distance between the agents' initial policies and optimal policies, $L_{a,n} \left(\pi_{\theta_{a,n}^{(1)}}, \mathbf{z}_j \right)$, is approximated by the update direction of $\theta_{a,n}^{(1)}$ evaluated in

$$\begin{aligned} & \tilde{L}_{a,n} \left(\pi_{\theta_{a,n,j}^{(2)}}, \mathbf{z}_j \right) \\ &= \left(r \left(\mathbf{a}_t^{(\epsilon'_{n,j})} \middle| \mathbf{s}_t^{(\epsilon'_{n,j})}, \mathbf{o} \right) + \gamma \tilde{V}_{\theta_{c,n,j}'} \left(\mathbf{s}_{n,t+1}^{(\epsilon'_{n,j})} \right) - \tilde{V}_{\theta_{c,n,j}'} \left(\mathbf{s}_{n,t}^{(\epsilon'_{n,j})} \right) \right) \log \pi_{\theta_{a,n,j}^{(2)}} \left(\mathbf{a}_{n,t}^{(\epsilon'_{n,j})} \middle| \mathbf{s}_{n,t}^{(\epsilon'_{n,j})}, \mathbf{o} \right), \end{aligned} \quad (4.19)$$

in which $\theta_{a,n,j}^{(2)} = \theta_{a,n}^{(1)} + \alpha_a \tilde{A}_n \left(\mathbf{a}_{n,t}^{(\epsilon_{n,j})}, \mathbf{s}_{n,t}^{(\epsilon_{n,j})} \right) \nabla_{\theta_{a,n}^{(1)}} \log \pi_{\theta_{a,n}^{(1)}} \left(\mathbf{a}_n^{(\epsilon_{n,j})} \middle| \mathbf{s}_{n,t}^{(\epsilon_{n,j})}, \mathbf{o} \right)$ captures the updated policy function parameters on gateway n . In essence, the approximated losses evaluate the policy and value update directions with the estimation performance achieved by updated policy and value functions within the target user request realization. These loss functions are, then, iteratively minimized using SGD.

As summarized in Algorithm 4, at each meta training iteration, J different user request realizations are sampled from distribution $P(\mathcal{Z} | \mathbf{o})$ at service occasion $\mathbf{o}$. Within each one of the sampled realization, $\mathbf{z}_j$, the satellite system serves ground users with contact plans that are sampled using the policy and value functions on the gateways. Then, the satellite system updates the policy and value functions using (5.11) and (5.11). After this, using loss functions in (5.12) and (5.13), the performance of the updated policy and value functions are evaluated based on the satellite system's experience $\epsilon'_{n,j}$ on serving user requests in $\mathbf{z}_j$ using these updated functions. The policy and value function parameters are, then, updated in the opposite direction of the gradient of loss functions calculated within all J sampled user request realizations. In the next iteration of meta training, the satellite system collects experiences on serving J newly sampled user request realizations with the updated policy and value functions, based on which the system update on the policy and value functions. Such update on the learning initializations will be repeated until a convergence is reached on the gateways.

In essence, the meta training procedure seeks to find the optimal learning initializations, i.e. $\pi_{\theta_{a,n,\mathbf{o}}^{(1)}}$, and $V_{\theta_{c,n,\mathbf{o}}^{(1)}}$, from which the proposed distributed D-RMRL solution starts its shortened searching of optimal strategies in every possible user requests at service occasion $\mathbf{o}$. These learning initializations are updated to be close to the optimal policies and values for all possible user request realizations at service occasion $\mathbf{o}$, such that it is easier for the distributed D-RMRL so-

Algorithm 4 Meta training for optimal learning initializations of D-RMRL at one service occasion.

Require: Service occasion $\mathcal{O}$, user locations, time constraints.

Ensure: Initialize value functions $\tilde{V}_{c,n}^{(1,1)}$, and policy functions $\pi_{\theta_{a,n}^{(1,1)}}$, for $n \in \mathcal{N}$.

```

1: for Meta training epoch  $i = 1 : I$  do
2:   for Environment sampling epoch  $j = 1 : J$  do
3:     Sample an user request realization  $\mathbf{z}_j \sim \mathcal{Z}$ .
4:     Carry out an experience by sampling actions with current policy functions, i.e.  $\pi_{\theta_{a,n}^{(1,i)}}$ , at each gateways
        $n \in \mathcal{N}$ .
5:     Calculate team advantage  $A$  based on (5.8).
6:     for Each gateway  $n = 1 : N$  do
7:       Update individual value and policy functions using (5.11) and (5.11).
8:       Carry out an experience by sampling actions with the updated policy functions, i.e.  $\pi_{\theta_{a,n,j}^{(2)}}$ .
9:       Calculate loss  $\tilde{L}_{c,n}(\tilde{V}_{c,n,j}^{(2)}, \mathbf{z}_j)$  and  $\tilde{L}_{a,n}(\pi_{\theta_{a,n,j}^{(2)}}, \mathbf{z}_j)$  with (5.12) and (5.13).
10:    end for
11:  end for
12:  for Each gateway  $n = 1 : N$  do
13:    Update initial value parameters  $\theta_{c,n}^{(1,i+1)} = \theta_{c,n}^{(1,i)} - \beta_c^{(i)} \nabla_{\theta_{c,n}^{(1,i)}} \sum_{j=1}^J \tilde{L}_{c,n}(\tilde{V}_{c,n,j}^{(2)}, \mathbf{z}_j)$ , and policy param-
       eters  $\theta_{a,n}^{(1,i+1)} = \theta_{a,n}^{(1,i)} - \beta_a^{(i)} \nabla_{\theta_{a,n}^{(1,i)}} \sum_{j=1}^J \tilde{L}_{a,n}(\pi_{\theta_{a,n,j}^{(2)}}, \mathbf{z}_j)$ .
14:  end for
15: end for
16: return Optimal initial policy functions  $\pi_{\theta_{a,n,\mathcal{O}}^{(1)}} = \pi_{\theta_{a,n}^{(1,*)}}$  and initial value functions  $V_{\theta_{c,n,\mathcal{O}}^{(1)}} = V_{\theta_{c,n}^{(1,*)}}$ , for
     $n \in \mathcal{N}$ .

```

lution to find the team optimal strategies at this service occasion. However, this meta training is executed for one specific service occasion. To further reduce the cost and improve the convergence performance on finding optimal learning initializations at different service occasions, we propose a jump-start optimization scheme. This scheme seeks to initialize a meta training procedure with learning experience at treated service occasions, using a parameter transfer approach.

4.3.4 Pre-training for Distribution Robust Meta Learning Performance

With the goal of improving meta training speed of D-RMRL, the proposed pre-training procedure is formulated as pump-start optimization problem, which is formulated as ¹

$$\min_{\theta_{c,n}^p, \theta_a^p} \sum_{n=1}^N L_{c,n}^p(\tilde{V}_{c,n}^p, \tilde{V}_{c,n}^{(1),\mathcal{O}}) + L_{a,n}^p(\pi_{\theta_{a,n}^p}, \pi_{\theta_{a,n}^{(1),\mathcal{O}}}), \quad (4.20)$$

$$\text{s. t. } \tilde{V}_{c,n}^p \in [\tilde{V}_{c,n}^{(1),\mathcal{O}}]_{\mathcal{O} \in \mathcal{O}^M}, \quad \pi_{\theta_{a,n}^p} \in [\pi_{\theta_{a,n}^{(1),\mathcal{O}}}]_{\mathcal{O} \in \mathcal{O}^M}, \quad (4.20a)$$

¹Here, the pump-start of meta training includes the policy and value functions from which the meta training procedure starts with.

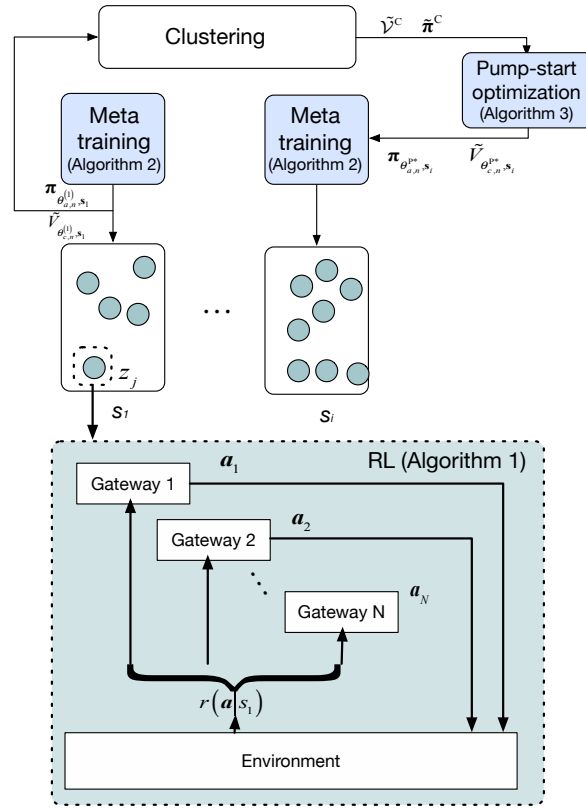


Figure 4.2: Illustration of the proposed T-RMRL algorithm.

where $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}$, and $\pi_{\theta_{a,n}^{(1)}, \mathbf{o}}$ are, respectively, the optimal initial value and policy functions at service occasion $\mathbf{o}$. $\tilde{V}_{\theta_{c,n}^p}$, $\pi_{\theta_{a,n}^p}$ are, respectively, the pump-start value and policy functions. $L_{c,n}^p(\tilde{V}_{\theta_{c,n}^p}, \tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}})$, and $L_{a,n}^p(\pi_{\theta_{a,n}^p}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}})$ measures the distance between the pump-starts and optimal learning initializations at service occasions $\mathbf{o}$. $\mathcal{O}^M$ is the set of service occasions at which the meta training in Algorithm 4 is implemented, such that the optimal initializations at these service occasions are calculated. In essence, the pump start optimization procedure seeks to start the meta training procedure at unseen service occasions with learning experiences from treated service occasions. It chooses the learning experience that is closest to the optimal functions at current service occasion, for improved convergence performance.

To reduce the complexity on measuring loss functions $L_{c,n}^p(\tilde{V}_{\theta_{c,n}^p}, \tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}})$ and $L_{a,n}^p(\pi_{\theta_{a,n}^p}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}})$ caused by computing the optimal initializations, an approximation on the loss functions is implemented. In particular, $L_{a,n}^p(\pi_{\theta_{a,n}^p}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}})$ is approximated by

$$\tilde{L}_{a,n}^p(\pi_{\theta_{a,n}^p}, \mathbf{o}) = \sum_{j=1}^{J^p} \left[r\left(\mathbf{a}_t^{(\varepsilon'_{n,j})} | \mathbf{s}_t, \mathbf{o}\right) - \tilde{V}_{\theta_{c,n,j}^{p'}}(\mathbf{o}_n) \right] \log \pi_{\theta_{a,n,j}^{p'}}\left(\mathbf{a}_n^{(\varepsilon'_{n,j})} | \mathbf{s}_{n,t}, \mathbf{o}\right), \quad (4.21)$$

where $\theta_{a,n,j}^{p'} = \theta_{a,n}^{(1)} + \alpha_a \tilde{A}_n\left(\mathbf{a}_n^{(\varepsilon_{n,j})}, \mathbf{o}_n\right) \nabla_{\pi_{\theta_{a,n}^p}} \log \pi_{\theta_{a,n}^p}\left(\mathbf{a}_n^{(\varepsilon_{n,j})} | \mathbf{s}_{n,t}, \mathbf{o}\right)$ is the updated policy function parameters of gateway n , based on experience $\varepsilon_{n,j}$ sampled from task $z_j \sim P(\mathcal{Z} | \mathbf{o})$. $\mathbf{a}_n^{(\varepsilon'_{n,j})}$ is the action of gateway n within experience $\varepsilon'_{n,j}$, which is sampled from task $z_j \sim P(\mathcal{Z} | \mathbf{o})$ with the updated policy. Meanwhile, $L_{c,n}^p(\tilde{V}_{\theta_{c,n}^p}, \tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}})$ is approximated as

$$\tilde{L}_{c,n}^p(\tilde{V}_{\theta_{c,n}^p}, \mathbf{o}) = \sum_{j=1}^{J^p} \left(r\left(\mathbf{a}_t^{(\varepsilon'_{n,j})} | \mathbf{s}_t, \mathbf{o}\right) - \tilde{V}_{\theta_{c,n,j}^{p'}}(\mathbf{s}_{n,t}) \right)^2, \quad (4.22)$$

in which $\theta_{c,n,j}^{p'} = \theta_{c,n}^p + \alpha_c \nabla_{\theta_{c,n}^p} A^2(\mathbf{a}_n^{(\varepsilon_{n,j})}, \mathbf{o})$ is the updated value function parameters of gateway n , based on experience $\varepsilon_{n,j}$ sampled from task $z_j \sim P(\mathcal{Z} | \mathbf{o})$. J^p is the number of tasks sampled for the pump-start performance evaluation at service occasion $\mathbf{o}$. Then, the bump-starts $\tilde{V}_{\theta_{c,n}^p} \in [\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}]_{\mathbf{o} \in \mathcal{O}^M}$, $\pi_{\theta_{a,n}^p} \in [\tilde{\pi}_{\theta_{a,n}^{(1)}, \mathbf{o}}]_{\mathbf{o} \in \mathcal{O}^M}$ resulting with minimum value of $\sum_{n=1}^N \tilde{L}_{c,n}^p(\tilde{V}_{\theta_{c,n}^p}, \mathbf{o}) + \tilde{L}_{a,n}^p(\pi_{\theta_{a,n}^p}, \mathbf{o})$ will be selected as the initial policy and value functions for the meta training procedure.

The proposed distributed D-RMRL with pre-training capability is summarized in Algorithm 5. At the beginning of the algorithm, the gateways will randomly start a meta training procedure, using Algorithm 4, for optimized learning initializations at service occasion $\mathbf{o}$. At unseen service occasions, the gateways initialize a shortened meta training procedure with the meta trained policy and value functions. In particular, the gateways will collect experience on serving users with the available pump-start policy and value functions, then update the functions, and evaluate the updates

Algorithm 5 Pre-trained D-RMRL algorithm for trajectory design at one realization of user access request at different service occasions.

Require: User locations, time constraints.

Ensure: Meta train learning initializations $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}}$ at service occasion $\mathbf{o}$, with random chosen pump-starts.

```

1: for Pre-training iteration  $i = 1 : I$  do
2:   Sample service occasion  $\mathbf{o}_i$ .
3:   for Each gateway  $n = 1 : N$  do
4:     for Environment sampling epoch  $j = 1 : J^P$  do
5:       Sample an user request realization  $\mathbf{z}_j \sim \mathcal{Z}$ .
6:       for Each group  $k = 1 : K$  do
7:         Gateway  $n$  serves users and collect experience using the  $k$ -th available pump-start policy function  $\pi_{\theta_{a,n}^p, k}$ , and value function  $\tilde{V}_{\theta_{c,n}^p, k}$ .
8:         Update pump-start policy and value functions using (5.11) and (5.11).
9:         Collect experience using updated pump-start policy function  $\pi_{\theta_{a,n}^{p'}, k}$ , and value function  $\tilde{V}_{\theta_{c,n}^{p'}, k}$ .
10:      end for
11:    end for
12:    Evaluate updated pump-start policy and value functions using (4.21) and (4.22).
13:    Gateway  $n$  selects the the bump-starts  $\tilde{V}_{\theta_{c,n}^{p*}}, \pi_{\theta_{a,n}^{p*}}$  resulting with minimum value of  $\sum_{n=1}^N \tilde{L}_{c,n}^p(\tilde{V}_{\theta_{c,n}^p, k}, \mathbf{o}_i) + \tilde{L}_{a,n}^p(\pi_{\theta_{a,n}^p, k}, \mathbf{o}_i)$ .
14:    Meta training on learning initializations  $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}_i}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}_i}$ , starting from pump-starts  $\tilde{V}_{\theta_{c,n}^{p*}, \mathbf{o}_i}$ , and  $\pi_{\theta_{a,n}^{p*}, \mathbf{o}_i}$ , at service occasion  $\mathbf{o}_i$ .
15:    Using Algorithm 3, gateway  $n$  learns to deliver data with learning initializations  $\tilde{V}_{\theta_{c,n}^{(1)}, \mathbf{o}_i}, \pi_{\theta_{a,n}^{(1)}, \mathbf{o}_i}$ , at service occasion  $\mathbf{o}_i$ .
16:    Set  $\mathbf{o}_i$  into  $\mathcal{S}^M$ .
17:  end for
18: end for

```

with loss functions in (4.22) and (4.21). Finally, the gateways will use the pump-start policy and value functions with minimum loss values to initiate the meta training procedure at current service occasion. The meta training procedure at different service occasions will be initiated with optimal pump-start policy and value functions, such that, the gateways can easily converge to the optimal serve coverage to every user requests at all different service occasions.

4.3.5 Convergence and complexity Analysis

Next, we first show that the proposed D-RMRL algorithm is guaranteed to converge to a team optimal strategy of problem (6.5).

Lemma 3. The proposed D-RMRL algorithm is guaranteed to converge to an optimal solution of problem (6.5), if the following conditions are satisfied [91]:

1. Value function $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,t})$ converges to a local minimum with $\nabla_{\theta_{c,n}} A^2(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) = 0$, and advantage function $A(\mathbf{a}_t, \mathbf{s}_t, \mathbf{o}) \neq 0$.

2. $\max_{a_{n,k} \in \mathcal{C}, o_{n,k} \in \mathcal{O}} \left| \frac{\partial \pi_{\theta_{a,n}}(a_{n,k} | o_{n,k})}{\partial \theta_i \partial \theta'_j} \right| < \infty$, for any elements θ_i and θ_j in the policy function parameter vector $\theta_a = [\theta_{a,1}, \theta_{a,2}, \dots, \theta_{a,N}]$.
3. $\lim_{i \rightarrow \infty} \alpha_c^{(i)} = 0$, $\sum_{i=1}^{\infty} \alpha_c^{(i)} = \infty$, $\lim_{i \rightarrow \infty} \alpha_a^{(i)} = 0$ and $\sum_{i=1}^{\infty} \alpha_a^{(i)} = \infty$.

In summary, using Lemma 1, we can see that the proposed D-RMRL algorithm is guaranteed to reach a team optimal solution of problem (6.5) by updating the policy and value functions locally at each gateway in the system.

Meanwhile, one step of policy and value function update yield complexity of $\mathcal{O}(n_c C)$ and $\mathcal{O}(n_a C)$, respectively. Here, n_c , and n_a are, respectively, the number of elements in policy function $\theta_{c,n}$, and value function $\theta_{a,n}$. The number of elements are considered to be small, given the low dimensional structure of the policy and value functions. C is the time complexity on calculating the gradient of each element in the policy and value functions. Thus, the complexity of the proposed D-RMRL algorithm is $\mathcal{O}(v(n_c + n_a)C)$, with v being the function update step at which the proposed algorithm converges. This complexity is considered to be low, since the number of policy and value function elements, i.e. n_c , and n_a , are small and are similar to that of a single agent RL algorithm. The proposed distribution-robust meta training mechanism further reduces this complexity with faster convergence, i.e. smaller v . In particular, the distribution-robust meta training mechanism initializes the RL solution with policy and value functions that are close to optimal policies and values for all possible user requests at current service occasion. The complexity of the distribution-robust meta training procedure is $\mathcal{O}(2(n_c + n_a)C(J^P + v_m J))$, with $\mathcal{O}(2(n_c + n_a)C)$ being the complexity of one step of meta training on one training sample (i.e. one service experience collected on the gateways). Here, J^P , and $v_m J$ are, respectively, the number of service experiences used for pump start optimization and meta training at one service occasion. Here, v_m is the number of iterations that the meta training procedure needs to reach convergence, which is minimized by the pump-start optimization procedure. This complexity is reasonable since meta training procedure uses the satellite system's experiences on serving ground user's daily request as training samples, and only introduce a low computation cost on gateways that is serving an unknown user request. In particular, the computation complexity of the proposed D-RMRL algorithm implemented on the gateways for serving one user request realization is $\mathcal{O}\left(2(n_c + n_a)C\left(\frac{J^P}{v_m J} + 1\right) + v(n_c + n_a)C\right)$.

Moreover, in regards to the implementation of the proposed algorithm, one can see that the algorithm is carried out by the gateways in the satellite communication system. The gateways will feed data to satellite constellation, and, at the same time, plan data routing in the system using the proposed algorithm. Using the proposed algorithm, the gateways will exchange information including achieved rewards, and estimated rewards, to achieve high overall service coverage. Thus, a total of $2N^2v$ pieces of information will be transmitted within the system at the convergence of the proposed algorithm. However, with a centralized RL solution, $(\nu_s + \nu_a + 1)N^2v$ pieces of information including the ν_s dimensional states, ν_a dimensional actions, and rewards will be exchanged among the gateways. Thus, the proposed algorithm is considered to have a lighter communication

Table 4.1: Simulation Parameters [2]

Parameter	Value	Parameter	Value
P_n	20 W	P_s	20 W
B_C	62.4 MHz	B_S	100 MHz
$\rho,$	0.2291	η	0.9129

overhead which makes itself more suitable for the dynamic satellite communication system.

4.4 Simulation Results and Analysis

For our simulations, we consider a scenario with five gateways serving $U = 500$ mobile users with the help of a LEO cube satellite constellation and $M = 20$ ground IAB BSs. The rest main simulation parameters are listed in Table I. Within the studied Monte Carlo experiment, the users are repeatedly deployed at random locations that are uniformly distributed in every cluster at every independent run. Meanwhile, user access request realizations at different service occasion are sampled and analyzed using the proposed solution. In particular, the quantity, content, and occurrence of user requests are assumed to follow Gaussian distribution with standard deviation, and mean sampled from a uniform distribution for every service occasion. The results of the proposed D-RMRL algorithm are compared with the ones with independent actor critic (IAC) algorithm [102] that updates identical policy and value functions at all agents, the value decomposition network (VDN) based RL defined in section III.B, and the MRL defined in section III.C. The IAC algorithm updates the policy and value functions locally at each gateways as in the proposed solution, but toward the maximal rewards of each individual gateway. Thus, the comparison between IAC and the proposed solution can justify how the value decomposition assumption improves the performance of the proposed algorithm. The results of the proposed algorithm are further compared to the ones from the VDN based RL without meta training capability, the MRL solution, as well as oracle results. Using the VDN solution, the gateways start with randomly chosen policy and value functions, while using the MRL solution, the gateways start with policy and value functions chosen by randomly initialized meta training mechanism. The VDN, and MRL solution learn contact plan designs within exactly the same environments sampled by the proposed D-RMRL algorithm. Thus, the comparison among the proposed D-RMRL, VDN, and MRL can justify how the proposed distribution-robust meta training mechanism shortens the meta training and RL procedure. The oracle results present the gateways' performance with optimal strategies in the considered environment. All statistical results are averaged over a large number of independent runs.

In Fig. 4.3, we show the convergence of all the considered RL algorithm. In this figure, we can see that the MRL algorithm proposed in section III.C requires approximately 800 iterations to reach convergence, which improves the convergence speed by up to 53.3% compared to the RL algorithm without meta training capability. This stems from the fact that, with the meta training initialization, the RL algorithm can start from policies that are close to the team optimal strategies in the target service task. Meanwhile, Fig. 4.3 also shows that the RL algorithm proposed in section III.B yields

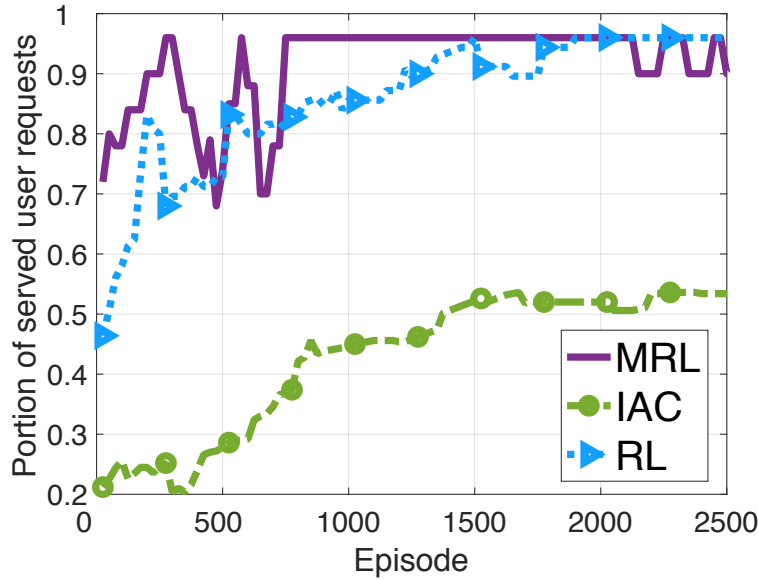


Figure 4.3: Convergence of all considered RL algorithms.

a 76.3% higher final service coverage than IAC, as it reinforces team beneficial strategies on each gateways with the decomposed team utility. Yet, by updating identical policy and value functions on the gateways, the IAC algorithm can only find strategies that maximize the gateways' individual utilities. Moreover, from Fig. 4.3, we also see that the proposed RL algorithm converges in similar speed compared to the IAC algorithm, since the gateways will update their own policy and value functions with their local experience within both scheme. In summery, the RL algorithm proposed in section III.B can improve the effective service coverage achieved by the satellite system with a low time complexity that is comparable to single agent RL algorithms.

In Fig. 4.4, we show the convergence of the proposed distribution-robust meta training scheme. From this figure, we can see that, with the proposed pump-start optimization procedure, the distribution-robust meta training scheme in Algorithm 5 converges at approximately the 400-th iteration, which is 48% and 64% faster than, respectively, the original meta training scheme, and pre-trained meta training scheme. This gain stems from the fact that the pump-start optimization initializes the meta training procedure with estimation of similar user access request from former service occasions. By starting the meta training procedure from the optimized pump start, the gateways can save significant computation cost on meta training.

Fig. 4.5 shows the convergence of the proposed D-RMRL method as the number of unexpected user requests varies. Here, the unexpected user requests are requests the satellites seek to but never learn to serve, and do not exist in the distribution estimated by the meta training procedure. Fig. 4.5 shows that, as the number of unexpected user requests increases, the number of iterations needed for the MRL algorithm to converge increases because of the improper estimation on the upcoming service tasks. From Fig. 4.5, we also observe that, as the number of unexpected user

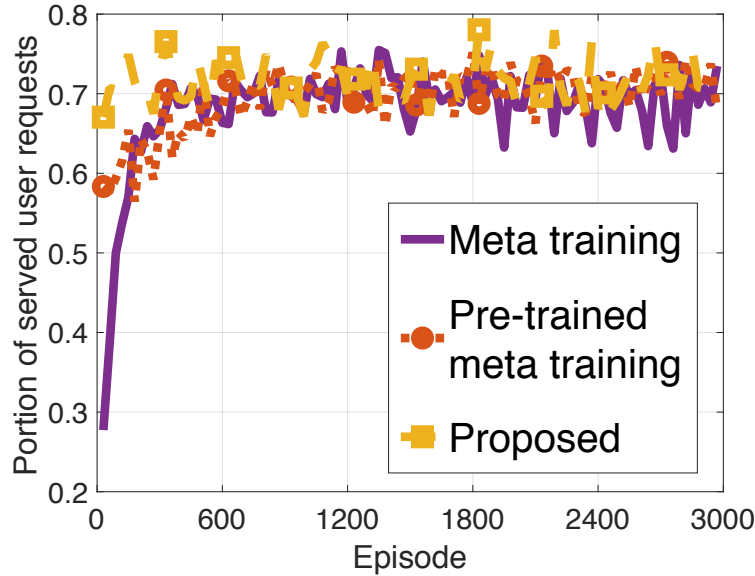


Figure 4.4: Convergence of all considered meta training scheme.

requests increases, the number of iterations needed for convergence at RL and D-RMRL barely change. This is because the complexity of considered problem does not change the unexpected user request. Randomly chosen learning initialization will keep the RL algorithm converges in a relatively low speed. While, by preparing the satellite system for possible user requests in all service occasion, the D-RMRL algorithm can converge much faster, i.e. up to 49.1% and 48% faster than the RL, and MRL algorithm, respectively.

4.5 Conclusion

In this chapter, we have studied the problem of contact plan design for on-demand data delivery with a cube satellite network. Using this network, the ground users' dynamic and unforeseeable needs are served by the orbiting cube satellites, which achieve data from distributed ground gateways, and relay data to users using IAB ground base stations. While feeding data to the network, the gateways designate the devices that will carry and distribute data using the limited contact chances in the time varying cube satellite network. We have formulated the studied problem in a dynamic game setting and have proposed a distributed D-RMRL algorithm to solve this problem. The proposed distributed D-RMRL algorithm makes the gateways independently optimize their strategies toward the maximal overall service coverage, by sharing only their achieved and estimated reward to each other. To reduce the excessive computation cost of this machine learning based solution in unseen and unexpected tasks, the proposed distributed D-RMRL algorithm use a pre-trained meta training mechanism to optimize the initializations of the RL and meta-training

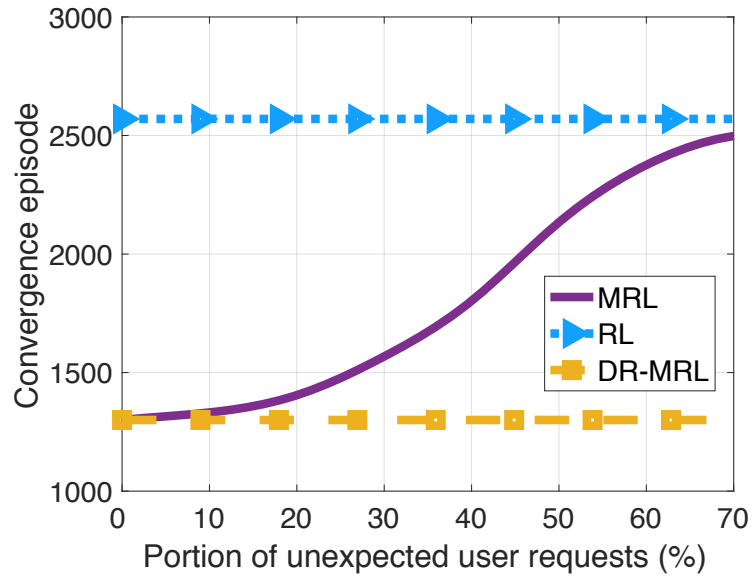


Figure 4.5: Convergence of all considered RL algorithms as the number of unexpected user requests varies.

procedure. Simulation results show that the proposed distributed D-RMRL algorithm with meta training mechanism outperforms the traditional RL algorithms.

Chapter 5

Meta-Reinforcement Learning for Trajectory Design in Wireless UAV Networks

5.1 Background, Related Works, and Contributions

As we discussed in Chapter 1, compared to the cube satellites, the drone driven communication platforms within NTN can provide a more cost-effective, flexible approach to boost the coverage and capacity of future wireless networks [141–143]. However, effectively deploying a group of DBSs for providing timely on demand wireless connectivity to ground users in dynamic wireless environments is still an important open problem. In particular, designing trajectories for a group of independent DBSs is challenging particularly when the DBSs only have limited information on the wireless requests of the ground users, which are often highly unpredictable and dynamic. In this chapter, we study the optimal deployment of DBSs within NTNs with a trajectory design solution that coordinates the DBSs with prior estimation of user needs.

5.1.1 Related Works

The existing literature in [92–95, 97] studied a number of problems related to trajectory design for drone-based wireless networks. The work in [92] studies the drone trajectory optimization problem by jointly considering both the drone’s communication throughput and its energy consumption. The authors in [93] design the trajectory of a solar-powered DBS to enhance its wireless communication performance. In [97], the problem of trajectory design and user association in a multi-drone communication system is solved with a block coordinate descent solution. The authors in [94] propose a dynamic trajectory control algorithm to improve the communication performance of the DBSs. The authors in [95] optimize time allocation, reflection coefficient adjustment, and

DBS trajectory in backscatter communication networks. Despite their promising results, these existing works [92–95, 97] do not consider practical DBS-assisted wireless networks in which the ground user requests for wireless service follow unpredictable patterns. Indeed, the optimization based solutions in [92–95, 97] are not suitable to design DBS trajectories when the user requests are unknown and unforeseeable.

Recently, there has been significant interest in realizing intelligent control of drones in dynamic networking environments, using machine learning tools for trajectory design [69–74, 102, 103, 144]. In [69], the problem of real-time dynamic maneuver design at a data collecting DBS is studied and solved using RL. The work in [70] employs two powerful deep neural networks to intelligently guide an energy-limited DBS under environmental dynamics. In [144], the authors develop an RL algorithm that enables a drone to act as a data-collection relay with the goal of maximize information freshness. However, the work in [144] is restricted to the case of a single drone. For intelligently controlling multi-drone systems, the authors in [71] design interference-aware paths for a group of cellular-connected drones by using a MARL solution that relies on a deep echo state network (ESN) architecture. The work in [72] proposes a deep MARL architecture that directs drones within a continuous set of locations to accomplish time-sensitive sensing tasks. Meanwhile, in [73], the authors develop a MARL framework to allow drones dynamically manage resources according to their local observations. The authors in [74] propose an ESN based learning architecture to predict the users' mobility patterns and, then, realize an optimal deployment of a group of DBSs. The works in [102] and [103] propose distributed multi-agent algorithms that allow a group of agents to update their individual strategies considering the team benefits. However, most of the existing MARL solutions such as those in [71], and [74, 102, 103] require DBSs to share their states and actions while searching for the optimal strategies. These traditional RL solutions have high complexity, as they solve multi-agent problems by updating strategies based on the entire set of agents' actions and strategies whose dimension increases exponentially with the number of agents. Meanwhile, the MARL solutions in [72] and [73] allow the agents to search strategies independently based on their own actions and states. However, using these RL solutions, the DBSs cannot optimize the sum utilities of all DBSs, and thus, cannot maximize the overall coverage of the ground users, since the DBSs are optimizing their individual utilities. In addition, traditional RL solutions such as those used in [71–74, 102, 103] cannot efficiently adapt the trajectories of the DBSs to unseen environments, as they are often overfitted to their training tasks. This is because the hyper-parameters, exploration strategies, and initializations of traditional RL algorithms are manually adjusted for fitting the training tasks. Once the agent faces an unseen task, manually adjusted RL algorithms may not converge to the optimal solution and, even if they do, the convergence speed will be very slow. As a result, the traditional RL algorithms in [69–74, 102, 103, 144] cannot effectively find optimal DBS trajectories in unseen environments. Finally, we note that in [145], we studied the problem of trajectory design for a single DBS operating in a dynamic environment using meta-learning. However, this prior work relies on a simple algorithm that cannot be scaled to larger networks.

5.1.2 Contributions

The main contribution of this chapter is a novel distributed framework for designing the trajectories of a group of cooperative DBSs in unpredictable, dynamic environments. To our best knowledge, *this is the first work that designs trajectories for a team of DBSs in unpredictable, dynamic environments using a multi-agent meta reinforcement learning solution*. In brief, our key contributions include:

- We consider a practical drone-aided wireless system in which a team of DBSs cooperatively navigate in an area, under strict energy constraints and limited information on surrounding environments, with the goal of providing uplink wireless connectivity to ground users. The DBSs can provide on-demand coverage to the ground users while adapting their trajectories to those users' unpredictable access requests. We formulate this trajectory design problem as an optimization framework whose structure is shown to be non-convex. To solve this problem, the “myopic” DBSs, which have only limited access to the information of ground users, seek to find trajectories that maximize the expected portion of served users – a wireless coverage metric that we use as the team utility of the group of DBSs.
- To solve the formulated trajectory optimization problem, we propose a novel, distributed, value decomposition reinforcement learning (VD-RL) algorithm. This algorithm is shown to reach a local optimal solution of the studied non-convex problem without requiring the DBSs to share their actions, states, or strategies. This makes the proposed VD-RL algorithm much less complex than traditional distributed MARL algorithms (e.g., those in [71], and [74, 102, 103]), as the DBSs can update their strategies based on their own low-dimensional actions and states. The proposed VD-RL algorithm allows the DBSs to independently select strategies that maximize the team utility by decomposing and attributing this team utility to each DBS. Thus, with the proposed VD-RL algorithm, the DBSs can find a local optimal solution, which yields a much higher team utility, compared to the one that can be achieved by existing MARL solutions in [72] and [73].
- To improve the convergence speed of VD-RL for unseen environments, we propose a meta training mechanism that uses an optimization based solution to meta train the VD-RL algorithm. In particular, the proposed meta training mechanism seeks to find proper policy and value function initializations that estimate all possible user request patterns for the VD-RL algorithm. Using this meta-learning mechanism, the VD-RL solution can quickly converge to a team optimal strategy when faced with an unseen task. The proposed optimization based meta training mechanism has a lower complexity compared to the meta training solutions in [146] and [147], because it does not require additional neural networks for meta-learning. Using the proposed approach, the DBSs can be meta-trained independently with their own actions and states. In particular, the meta-learning approach helps DBSs to cope with various tasks instead of a single sampled task as in [148], by using a meta-trained initialization.

Simulation results show that proposed VD-RL algorithm can improve the service coverage and

convergence speed by up to 53.2% and 30.6%, compared to traditional multi-agent algorithms. The results also show that using the proposed meta training mechanism, our VD-RL algorithm can find optimal trajectories in an unseen environment with a 53.8% faster convergence speed, and a 5.6% better coverage, compared to the VD-RL algorithm without meta-learning.

The rest of this chapter is organized as follows. The system model and problem formulation are described in Section II. In Section III, the proposed algorithm is developed and discussed. In Section IV, simulation results are presented. Finally, conclusions are drawn in Section V.

The rest of this chapter is organized as follows. The system model and problem formulation are described in Section 5.2. In Section 5.3, the proposed algorithms are developed and discussed. In Section 5.4, numerical simulation results are presented. Finally, conclusions are drawn in Section 5.5.

5.2 System model

Consider a geographical area within which a set $\mathcal{U}$ of U randomly deployed terrestrial users request uplink data service. A set $\mathcal{N}$ of N fixed-wing DBSs are dispatched to satisfy the uplink access requests of those ground users, as shown in Fig. 6.1. In the considered area, an user that requests data service is called an *active user*, otherwise, it will be an *inactive user*. We assume the users to be separated into different groups, each of which is called a *cluster*. Note that, a “cluster” is defined as an area that falls within the service coverage of only one DBS. We also assume that, at any given time, each cluster will be served by a single DBS. The set of such clusters is denoted as $\mathcal{C}$. The DBSs will travel across the clusters in a steady straight-and-level flight (SLF), and hover over each cluster with a steady circular flight (SCF), at a constant speed V_s [92]. Each DBS $n \in \mathcal{N}$ flies at its own constant altitude ¹ H_n to avoid collision with other DBSs. All the DBSs must return to their original location O within a time period T for battery charging. Moreover, all DBSs are assumed to have the same battery capacity. The trajectory that records DBS n ’s movement within the time period T is represented by a vector $\boldsymbol{\xi}_n = [\xi_{n,1}, \xi_{n,2}, \dots, \xi_{n,K}]^\top$, with $\xi_{n,k} \in \mathcal{C} \cup \{O\}$ being the k -th cluster that DBS n serves, or the initial location O that DBS n returns to after serving the users, and K being the maximum number of locations each DBS can fly across under its energy constraints. In other words, each DBS n flies across no more than K different locations, each of which is called one *step* on the trajectory of DBS n , the set of such steps is denoted as $\mathcal{K}$. For example, $\boldsymbol{\xi}_n = [\xi_{n,1}, O, \dots, O]$, with $\xi_{n,1} \in \mathcal{C}$, implies that DBS n serves cluster $\xi_{n,1}$ and then flies back to the initial location O , while $\boldsymbol{\xi}_n = [\xi_{n,1}, \xi_{n,2}, \dots, \xi_{n,K-1}, O]$, with $\xi_{n,k} \in \mathcal{C}$, implies that DBS n serves clusters $\xi_{n,1}, \xi_{n,2}, \dots, \xi_{n,K-1}$, and then flies back to the initial location O .

¹Here, the differences in the altitudes between the DBSs are assumed to be considerably small when compared to the DBSs’ dedicated altitudes H_n . Thus, the altitude differences only have limited impact on the DBSs’ communication performance, including data rates, as well as service coverage.

5.2.1 Communication Performance Analysis

In our network, the users adopt an orthogonal frequency division multiple access (OFDMA) technique and transmit data over a set of uplink *resource blocks* (RBs) [149]. Each dispatched DBS will arbitrarily allocate one RB to each one of its associated users within a cluster. We assume that each DBS can keep serving its associated users within a d_r -meter radius over each cluster as shown in Fig. 6.1. The area within this range is called a *service area*. Also, user u is assumed to request a total of b_u data (in bits) at time epoch t_u . The DBSs must satisfy these requests within their battery capacities. Let $\mathbf{b} = [b_1, \dots, b_U]$ and $\mathbf{t} = [t_1, \dots, t_U]$ be, respectively, the vector of quantity and occurrence of the users' access request in the network. The quantity b_u and active time t_u are assumed to be independent random variables that follow unknown distributions. Hereinafter, we denote $\mathbf{z} = [\mathbf{b}, \mathbf{t}]$ as one *realization* of the users' access request within duration T . In this model, the deployed DBSs are *myopic*, that is, they only know the access quantities and active time of the users that they are currently serving. The path loss (in dB) of the LoS and non-line-of-sight (NLoS) air-to-ground communication links between DBS n to user u are given by the popular air-to-ground model in [2]

$$\begin{aligned} h_{u,n}^{\text{LoS}} &= 20 \log \left(\frac{4\pi f_c d_{u,n}}{c} \right) + \zeta_{u,n}^{\text{LoS}}, \\ h_{u,n}^{\text{NLoS}} &= 20 \log \left(\frac{4\pi f_c d_{u,n}}{c} \right) + \zeta_{u,n}^{\text{NLoS}}, \end{aligned} \quad (5.1)$$

where f_c is the carrier frequency of the communication link between DBS n and user u , $d_{u,n}$ is the distance between user u and DBS n , and c is the speed of light. $\zeta_{u,n}^{\text{LoS}}$ and $\zeta_{u,n}^{\text{NLoS}}$ are, respectively, the additional path losses at the LoS and NLoS air-to-ground links between DBS n to user u . The value of $\zeta_{u,n}^{\text{LoS}}$ and $\zeta_{u,n}^{\text{NLoS}}$ follow Gaussian distributions with different parameters $(\mu_{\text{LoS}}, \delta_{\text{LoS}}^2)$ and $(\mu_{\text{NLoS}}, \delta_{\text{NLoS}}^2)$, respectively. Note that the path loss values between DBS n and user u are considered to be stable with DBS n 's movement, as the distance between DBS n and user u will only experience small changes when DBS n flies within the service area. The SNR at the LoS and NLoS links between DBS n and user u , will thus be

$$\begin{aligned} \gamma_{u,n}^{\text{LoS}} &= \frac{P}{N_0 B 10^{\frac{h_{u,n}^{\text{LoS}}}{20}}}, \\ \gamma_{u,n}^{\text{NLoS}} &= \frac{P}{N_0 B 10^{\frac{h_{u,n}^{\text{NLoS}}}{20}}}, \end{aligned} \quad (5.2)$$

where P represents the transmit power of user u , which is assumed to be equal for all users. N_0 is the noise power spectral density, and B is the RB bandwidth (equal for all RBs). In order to avoid LoS interference to ground links and because the number of DBSs in real systems will be small, we assume that each DBS will use its own dedicated frequency band. The data rate at the link between DBS n and user u will then be $c_{u,n} = \beta_{u,n}^{\text{LoS}} B \log(1 + \gamma_{u,n}^{\text{LoS}}) + \beta_{u,n}^{\text{NLoS}} B \log(1 + \gamma_{u,n}^{\text{NLoS}})$, where $\beta_{u,n}^{\text{LoS}} = [1 + \varphi \exp(-\phi \frac{180}{\pi} \theta_{n,u} + \varphi \phi)]^{-1}$ is the probability of having a LoS link between DBS n and user u , $\beta_{u,n}^{\text{NLoS}} = 1 - \beta_{u,n}^{\text{LoS}}$ is the probability of having a NLoS link between DBS n and

user u . Here, φ and ϕ are constant values that depend on the studied communication environments, while $\theta_{n,u}$ is the elevation angle between DBS n and user u .

5.2.2 Utility Function Model

In the studied scenario, the goal of the dispatched DBSs is to cover all access requests from ground users. In such a case, the utility of each DBS is defined as the *successful service rate*, which captures the fraction of users being served by a given DBS in a given time period. Note that, when DBS n arrives at a cluster, it will only serve the user requests that were not served by the DBSs that arrived at this cluster before DBS n . For the special case in which multiple DBSs arrive at a cluster at the same time, only one DBS will serve the entire cluster, and the remaining DBSs will directly proceed toward other clusters. Note that once a DBS finds out (with their limited situational awareness) that another DBS is hovering on the service area of a cluster, it knows that this cluster is being served, and it will leave this cluster. Thus, the successful service rate achieved at DBS n by serving cluster $\xi_{n,k}$ is given by

$$\mu_{n,k}(\xi) = \frac{\sum_{u \in \mathcal{U}} \mathbb{1}_{\{u \in \mathcal{U}_{n,k}, T - \tau_n^* \leq t_u \leq T - \tau_{n,k}\}}}{\sum_{u \in \mathcal{U}} \mathbb{1}_{\{0 \leq t_u \leq T\}}}, \quad (5.3)$$

where $\xi = [\xi_1, \xi_2, \dots, \xi_N]$ is the matrix of trajectories of the DBSs, $\mathcal{U}_{n,k}$ is the set of active users in cluster $\xi_{n,k}$. $\tau_{n,k}$ is the time duration that DBS n is allowed to keep flying with its remaining energy level, after successfully serving cluster $\xi_{n,k}$ ². Note that, hereinafter, the time duration $\tau_{n,k}$ is called the *available service time* of DBS n at step k . In (5.3), $\mathbb{1}_{\{x\}} = 1$ when x is true, otherwise, $\mathbb{1}_{\{x\}} = 0$. Here, $\sum_{u \in \mathcal{U}} \mathbb{1}_{\{0 \leq t_u \leq T\}}$ is the number of active users within the studied time duration, and $\sum_{u \in \mathcal{U}} \mathbb{1}_{\{u \in \mathcal{U}_{n,k}, T - \tau_n^* \leq t_u \leq T - \tau_{n,k}\}}$ is the number of active users served by DBS n in cluster $\xi_{n,k}$. $\tau_n^* = \min_{n' \in \mathcal{N}_n} \tau_{n',k'}$ is the available service time of the last DBS that arrived at cluster $\xi_{n,k}$ before DBS n . $T - \tau_n^* \leq t_u \leq T - \tau_{n,k}$ represents that user u requests data before DBS n 's arrival, and have not served by any other DBSs, as shown in Fig. 5.1. Here, $\mathcal{N}_n = \{n' | n' \in \mathcal{N} \setminus n, \xi_{n',k'} = \xi_{n,k}, \tau_{n',k'} \geq \tau_{n,k}\}$ is the set of DBSs that arrive at cluster $\xi_{n,k}$ before DBS n , where $\xi_{n',k'} = \xi_{n,k}$ implies that cluster $\xi_{n,k}$ is also the k' -th cluster served by DBS n' . Moreover, DBS n 's available service time upon its arrival at cluster $\xi_{n,k}$ on trajectory ξ_n is given by $\tau_{n,k} = T - \sum_{\kappa=0}^{k-1} \frac{d_{n,\kappa,\kappa+1}}{V} - \sum_{\kappa=1}^{k-1} D_{n,\kappa}^*$, with $d_{n,\kappa,\kappa+1}$ being the distance between cluster $\xi_{n,\kappa}$ and $\xi_{n,\kappa+1}$. Here, $\frac{d_{n,\kappa,\kappa+1}}{V}$ is the time needed by DBS n to travel in an SLF from cluster $\xi_{n,\kappa}$ to cluster $\xi_{n,\kappa+1}$. $D_{n,\kappa}^*$ is the time needed by DBS n for hovering while serving the users in cluster $\xi_{n,\kappa}$, and this hovering time is given by

$$D_{n,\kappa}^* = \begin{cases} \max_{u \in \mathcal{U}_{n,\kappa}^*} D_{u,n} - \frac{2d_r}{V}, & \text{if } \max_{u \in \mathcal{U}_{n,\kappa}^*} D_{u,n} - \frac{2d_r}{V} > 0, \\ 0, & \text{otherwise,} \end{cases} \quad (5.4)$$

²The utility at a given DBS is determined by the trajectories of all DBSs since the number of user requests in the course of this DBS n 's hovering over a target cluster is determined by the number of active users served by the DBSs previously arrived at this cluster.

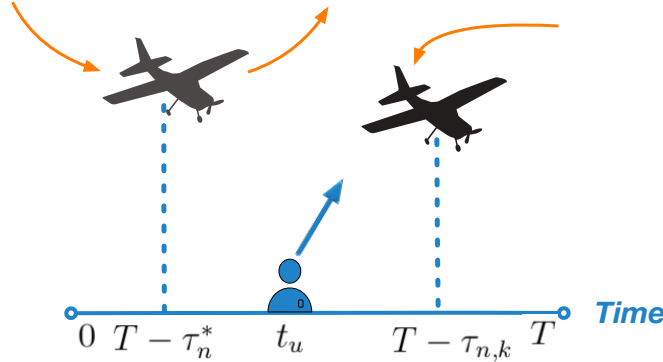


Figure 5.1: Snapshot of one user access request. DBSs only connect with active users that have not been served by other DBSs.

where $D_{u,n} = \frac{b_u}{c_{u,n}}$ is the transmission delay of user u , when being served by DBS n . $\max_{u \in \mathcal{U}_{n,\kappa}} D_{u,n}$ is the time that DBS n used to serve all the users in cluster $\xi_{n,\kappa}$. $D_{n,\kappa}^*$ is the time DBS n consumes for SCF hovering, while $\frac{2d_r}{V}$ is the time DBS n consumes its SLF travel within the service area. We define $\mathcal{U}_{n,\kappa}^* = \{u | u \in \mathcal{U}_{n,\kappa}, T - \tau_n^* \leq t_u \leq T - \tau_{n,\kappa}\}$ as the set of active users that can be served by DBS n in cluster $\xi_{n,\kappa}$.

5.2.3 Problem Formulation

In our system model, a group of energy-constrained and myopic DBSs fly independently to cover all the access requests from ground users. The successful service rate achieved by the DBSs is defined as a *team utility*, and is given by

$$G(\xi) = \sum_{k=1}^K \sum_{n \in \mathcal{N}} \mu_{n,k}(\xi). \quad (5.5)$$

The goal of the DBSs is to find optimal trajectories that maximize the expected team utility. Next, we first define expected team utility and then we introduce the optimization problem. Let $\pi_n(\xi | \xi_{n,k}, \tau_{n,k})$ be the strategy of DBS n , defined as the probability that DBS n moves toward cluster $\xi \in \mathcal{C} \cup \{O\}$, after successfully serving cluster $\xi_{n,k}$ with available service time $\tau_{n,k}$, and let $\pi = [\pi_n(\xi | \xi_{n,k}, \tau_{n,k})]_{n \in \mathcal{N}, k \in \mathcal{K}}$ be the vector of strategies of all DBSs. Then the expected team utility is defined as

$$\overline{G}(\pi) = \sum_{\xi \in \mathcal{C}} G(\xi) \prod_{n=1}^N \prod_{k=1}^K \pi_n(\xi | \xi_{n,k}, \tau_{n,k}), \quad (5.6)$$

where $\mathcal{C}$ is the set of all possible trajectories of the DBSs. As such, the trajectory design problem can be formulated as

$$\max_{\pi} \overline{G}(\pi), \quad (5.7)$$

$$\text{s. t. } \sum_{\xi \in \mathcal{C}} \prod_{n=1}^N \prod_{k=1}^K \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) = 1, \quad (5.7a)$$

$$\sum_{\xi \in \mathcal{C} \cup \{O\}} \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) = 1, \forall n \in \mathcal{N}, \xi \in \mathcal{C}, k \in \mathcal{K}, \quad (5.7b)$$

$$0 \leq \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) \leq 1, \forall n \in \mathcal{N}, \xi \in \mathcal{C}, k \in \mathcal{K}, \quad (5.7c)$$

Here, (6.5a) means that the DBSs must choose trajectories from $\mathcal{C}$. (6.5b) indicates that, within the considered time duration, each DBS n must choose to serve one cluster in $\mathcal{C}$ or return to the origin. The optimal solution of problem (6.5) guarantees a maximum expected team utility at the DBSs, and it is called a *team optimal strategy*. Here, we note that the use of traditional optimization algorithms, such as branch and bound or nonlinear programming, is not suitable to solve (6.5), as problem (6.5) is non-convex, and the successful service rate $\mu_{n,k}(\xi)$ achieved at each DBS is unpredictable with the values of $\mathbf{b}$ and $\mathbf{t}$ following unknown distributions. Moreover, traditional machine learning algorithms such as Q learning, policy gradient, and ESN [135–137] that have been previously used to solve complex optimization problem are also not suitable to solve the problem (6.5). This is because those algorithms must be manually adjusted to solve their training tasks, and they cannot find optimal strategies for the DBSs in unseen environments. To solve the non-convex problem (6.5) formulated in dynamic, unpredictable environments, we propose a distributed meta-trained VD-RL algorithm that finds team optimal strategies by meta training a distributed VD-RL solution equipped with a primary estimation on unseen environments. The distributed VD-RL algorithm updates the DBSs' strategies and estimation on the strategy outcomes at each individual DBS, based on only this DBS's actions and states. The meta training procedure finds an initialization of the policy and value functions for the VD-RL solution. This initialization is close to all optimal policy and value functions at all possible environments and, hence, it enables the VD-RL solution quickly converge in the dynamic environments and reduces the VD-RL solution's cost on time, energy, and the drone hardware. The proposed meta-trained VD-RL algorithm is introduced in the next section.

5.3 Proposed solutions

We now introduce a distributed meta-trained VD-RL algorithm, that merges the concept of value decomposition network [138], model agnostic meta-learning [139], with the policy gradient (PG) framework. The traditional PG algorithm can find the optimal trajectory for a single DBS. However, PG cannot find a team optimal strategy for a group of DBSs, as it will direct all the DBSs to one trajectory. In order to design trajectories for multiple DBSs, an algorithm that can reach a team optimal strategy for the DBSs must be proposed. Moreover, to prepare the DBSs for unseen environments, the proposed algorithm must not be overfitted to its training tasks, and it should converge to a team optimal strategy when the team utility function in (5.5) changes. To address these challenges, we propose a meta-trained VD-RL that coordinates a group of DBSs in various environments. Next, we first introduce the proposed VD-RL algorithm which solves the non-convex problem (6.5). Then, we explain how to use meta-learning approaches to meta train this VD-RL algorithm in order to allow it to cope with various environments.

5.3.1 Value Decomposition based Reinforcement Learning Algorithm

Next, we first introduce the components of the VD-RL algorithm. Then, we explain how the VD-RL algorithm can decompose problem (6.5) to problems that can be solved at each individual DBS. Finally, we introduce the procedure of using the proposed VD-RL algorithm to solve the non-convex problem in (6.5).

5.3.1.1 Value decomposition based reinforcement learning components

The proposed VD-RL algorithm consists of seven components

- *Agents*: The agents in VD-RL are the DBSs in set $\mathcal{N}$.
- *States*: Each agent has a state that consists of both its location, represented by the cluster it currently serves i.e. $\xi_{n,k}$, and its energy level, which is captured by the time $\tau_{n,k}$ that this DBS still has in order to return to the origin. Thus, the state of DBS n at step k is given by $\mathbf{s}_{n,k} = [\xi_{n,k}, \tau_{n,k}]$. The set of states at all DBSs is $\mathcal{S} = \{\mathbf{S}_0, \mathbf{S}_1, \dots, \mathbf{S}_K\}$, with $\mathbf{S}_k = [\mathbf{s}_{1,k}, \mathbf{s}_{2,k}, \dots, \mathbf{s}_{N,k}]$ being the matrix of the DBSs' states at step k .
- *Actions*: The action of each agent is the cluster that it seeks to serve, or the origin location that it will return to after serving several clusters. The action chosen by DBS n at step k is given by $a_{n,k} \in \mathcal{C} \cup \{O\}$, while the vector of all DBSs' actions at step k is $\mathbf{a}_k = [a_{1,k}, a_{2,k}, \dots, a_{N,k}]$.
- *Strategy*: The strategy of each DBS is defined as the probability of choosing a given action $a_{n,k} \in \mathcal{C} \cup \{O\}$ at a given state $\mathbf{s}_{n,k}$ and is denoted by $\pi_n(a_{n,k} | \mathbf{s}_{n,k})$. $\boldsymbol{\pi} = [\pi_n(a_{n,k} | \mathbf{s}_{n,k})]_{n \in \mathcal{N}, k \in \mathcal{K}}$ is the vector of strategies of all DBSs.
- *Policy function*: We define a policy function $\boldsymbol{\pi}_{\theta_{a,n}}$ that is a deep neural network parametrized by $\theta_{a,n}$ and is used to generate the strategy of DBS n . This policy function $\boldsymbol{\pi}_{\theta_{a,n}}(a_{n,k} | \mathbf{s}_{n,k})$ takes DBS n 's state as an input and outputs a strategy for DBS n at this state.
- *Reward*: The reward of each DBS measures the benefit of a selected action. In particular, aiming at maximizing coverage in the considered area, the reward of each DBS is defined as the successful service rate achieved by all the DBSs, which is given by $r(\mathbf{a}_k | \mathcal{S}) = \sum_{n \in \mathcal{N}} \mu_{n,k}(\boldsymbol{\xi})$. Here, different from traditional RL algorithms in which each agent only maximizes its own achievable utility, our proposed VD-RL algorithm enables each DBS to maximize the utility of all DBSs, which is also called the team stage reward³.

³Note that, DBS n 's successful service rate $\mu_{n,k}$ defined in (5.3), depends on DBS n 's current action $a_{n,k}$, state $\mathbf{s}_{n,k}$, and the actions taken by other DBSs before DBS n 's arrival at cluster $\xi_{n,k}$. As the DBSs operate asynchronously in the wireless environment as shown in Fig. 5.2, one DBS may reach step $k' \neq k$ on its trajectory at the time DBS n arrives at its step k . In other words, at time epoch $T - \tau_{n,k}$, the DBSs will reach different steps of their trajectories. The set of DBSs' states at time epoch $T - \tau_{n,k}$ is denoted as $\mathcal{B}_{T-\tau_{n,k}}$. In this case, the team stage reward resulting

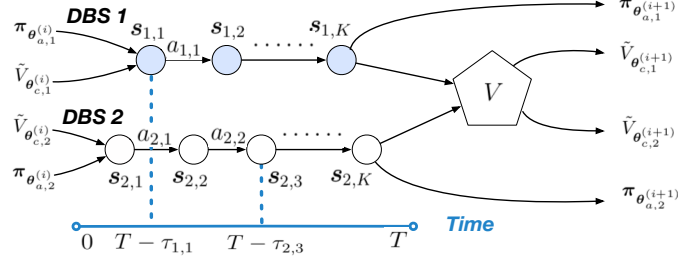


Figure 5.2: Illustration of one iteration of the VD-RL algorithm defined in Algorithm 6.

- *Value function:* We define a value function $V(\mathbf{S}_k)$ that is a deep neural network used to estimate the DBSs' achievable future rewards at every state $\mathbf{S}_k$. In particular, within the proposed VD-RL algorithm, the goal of the DBSs is to find a team optimal strategy that maximizes the expected team utility in (5.5). To this end, the DBSs must consider the current and also future rewards that can be achieved at every state. Thus, at every state $\mathbf{S}_k$, the DBSs seeks to maximize a discounted future reward $\sum_{n \in \mathcal{N}} \sum_{\kappa=k}^K \gamma^{k-\kappa+1} r(\mathbf{a}_\kappa | \mathcal{S})$, which is estimated by value function $V(\mathbf{S}_k)$, with γ being the discounted factor⁴.

In VD-RL, the DBSs interact with the unknown wireless environment by selecting actions in $\mathcal{C} \cup \{O\}$ based on the strategy generated by policy function $\pi_{\theta_{a,n}}$. In particular, the DBSs select actions step by step and receive the team stage rewards as feedback. An *experience* defined in vector $e_n = [s_{n,1}, a_{n,1}, r(\mathbf{a}_1 | \mathcal{S}), \dots, s_{n,K}, a_{n,K}, r(\mathbf{a}_K | \mathcal{S})]$ is collected by each DBS $n \in \mathcal{N}$. The DBSs then update the value function with their collected experiences and update their strategies to maximize future rewards estimated by $V(\mathbf{S}_k)$. In particular, the DBSs will update their estimation on future rewards after they return to the origin. At the origin, each DBS n only needs to know the number of active users served by all the DBSs from their last experience to update its estimation on future rewards, it does not have to know the actions taken by other DBSs. However, value function $V(\mathbf{S}_k)$ still depends on the states of all DBSs. Hence, each DBS may not be able to train its value function individually. Therefore, we introduce a value decomposition method using which each DBS can train its value function based on its local states.

from action $\mathbf{a}_k$ should be $r(\mathbf{a}_k | \bigcup_{n \in \mathcal{N}} \mathcal{B}_{T-\tau_{n,k}})$. Hereinafter, we denote it as $r(\mathbf{a}_k | \mathbf{S})$ for simplicity. It is worthy noting that, within the proposed VD-RL algorithm, the DBSs only record the number of served active users at each step on their trajectories.

⁴The discounting factor γ determines the scale of these steps. In particular, a discount factor close to 1 provides a long-term goal that accumulates rewards far into the future, while a discount factor close to 0 provides a short-sighted goal with which the DBSs only focus on immediate rewards. In fact, a proper scaling/discounting of future rewards can help a learning algorithm converge faster to the optimal solutions. If one uses the exact future rewards, the DBSs will only focus on their long-term goal and, thus, they would directly head to the cluster with most active users without stopping at other clusters. This would lead to a lower overall successful service rate (which can be seen, for example, in the numerical results in our conference paper [145]).

5.3.1.2 Value decomposition

The VD-RL algorithm needs to find a value function accurately estimate future rewards, and, thus, seeks to minimize the TD error metric that is defined as follows [140]:

$$A(\mathbf{a}_k, \mathbf{S}_k) = Nr(\mathbf{a}_k | \mathbf{S}) + \gamma V(\mathbf{S}_{k+1}) - V(\mathbf{S}_k) \quad (5.8)$$

where TD error $A(\mathbf{a}_k, \mathbf{S}_k)$ is called the *team advantage* of the DBSs at state $\mathbf{S}_k$ with action $\mathbf{a}_k$. The team advantage measures the difference between the DBSs' future reward estimated by $Nr(\mathbf{a}_k | \mathbf{S}) + \gamma V(\mathbf{S}_{k+1})$, and the one estimated by $V(\mathbf{S}_k)$. In essence, the team advantage keeps improving the value function $V(\mathbf{S}_k)$'s estimation accuracy using new sampled experiences. Moreover, the team advantage also reveals how actions in $\mathbf{a}_k$ are better than other action selections by showing the difference between the future reward reached by actions in $\mathbf{a}_k$ and the estimated future reward $V(\mathbf{S}_k)$. However, to enable the DBSs to update their value functions individually, one must divide the value of function $V(\mathbf{S}_k)$ among all DBSs. The value allocated to each DBS must reinforce team beneficial actions, and weaken all other actions at each DBS. Our VD-RL algorithm implements this value decomposition across the DBSs. In particular, the proposed VD-RL algorithm decomposes value $V(\mathbf{S}_k)$ under the assumption that this value is the sum of the values attributed to each DBS

$$V(\mathbf{S}_k) = \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k}), \quad (5.9)$$

where $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k})$ is the individual value function parametrized by $\theta_{c,n}$ at DBS n . Thus, the team advantage $A(\mathbf{a}_k, \mathbf{S}_k)$ be expressed as

$$\begin{aligned} A(\mathbf{a}_k, \mathbf{S}_k) = \\ Nr(\mathbf{a}_k | \mathbf{S}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k+1}) - \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k}). \end{aligned} \quad (5.10)$$

Based on the assumption (5.9), the DBSs can update their policy and value functions independently.

5.3.1.3 Value decomposition based reinforcement learning solution

When searching for the team optimal strategy that solves problem (6.5), the VD-RL algorithm needs to find optimal value functions that accurately estimate future rewards at every state, as well as the optimal policy functions that can always yield the actions leading to the highest future rewards at each DBS. In particular, the update of the individual value function $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k})$ of each

DBS n is given by

$$\begin{aligned}
\theta_{c,n}^{(i+1)} &= \theta_{c,n}^{(i)} - \alpha_c^{(i)} \nabla_{\theta_{c,n}^{(i)}} \sum_{k=1}^K A^2(\mathbf{a}_k, \mathbf{S}_k) \\
&= \theta_{c,n}^{(i)} - \alpha_c^{(i)} \nabla_{\theta_{c,n}^{(i)}} \sum_{k=1}^K \left(N r(\mathbf{a}_k | \mathbf{S}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}^{(i)}}(\mathbf{s}_{n,k+1}) \right. \\
&\quad \left. - \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}^{(i)}}(\mathbf{s}_{n,k}) \right)^2 \\
&= \theta_{c,n}^{(i)} + 2\alpha_c^{(i)} \sum_{k=1}^K A(\mathbf{a}_k, \mathbf{S}_k) \nabla_{\theta_{c,n}^{(i)}} \tilde{V}_{\theta_{c,n}^{(i)}}(\mathbf{s}_{n,k}),
\end{aligned}$$

where $\theta_{c,n}^{(i)}$ is the value function parameter of DBS n and $\alpha_c^{(i)}$ is the value function parameter update step size, at the i -th VD-RL training iteration ⁵.

After precisely estimating the future team stage rewards using $V(\mathbf{S}_k)$, and attributing the value to each one of the DBS, the policy function of each DBS can be updated. Let $\tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) = r(\mathbf{a}_k | \mathbf{S}) + \gamma \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k+1}) - \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k})$ be the individual advantage at BS n at state $\mathbf{s}_{n,k}$ with action $a_{n,k} \in \mathbf{a}_k$. Hence, we have $A(\mathbf{a}_k, \mathbf{S}_k) = \sum_{n \in \mathcal{N}} \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k})$. Based on the the policy gradient theorem [137], the update on DBS n 's policy function parameters is given by

$$\begin{aligned}
\theta_{a,n}^{(i+1)} &= \\
&\theta_{a,n}^{(i)} + \alpha_a^{(i)} \sum_{k=1}^K \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) \nabla_{\theta_{a,n}^{(i)}} \log \pi_{\theta_{a,n}^{(i)}}(a_{n,k} | \mathbf{s}_{n,k}), \tag{5.11}
\end{aligned}$$

where $\theta_{a,n}^{(i)}$ is the policy function parameter at DBS n and $\alpha_a^{(i)}$ is the policy function parameter update step size, at the i -th VD-RL training iteration.

The proposed VD-RL solution is summarized in Algorithm 6. At the beginning of the algorithm, each DBS serves its users with a randomly initialized value function $V_{\theta_{c,n}^{(1)}}$ and policy function $\pi_{\theta_{a,n}^{(1)}}$. The successful service rates that the DBSs achieved with the selected policy functions are calculated and recorded by the DBSs upon their return to the origin. The DBSs update their value and policy function parameters based on (5.11) and (5.11), using the recorded successful service rates. Then, the DBSs serve the users with the updated policy functions. Note that the proposed VD-RL algorithm deploys a mini-batch training mechanism. In particular, using the proposed algorithm, the DBS updates its policy over a whole experience in order to reduce the variance caused by the action sampling, as well as the need to store a big dataset. Thus, while flying over the target area, the DBSs will, step by step, select clusters to serve by sampling actions from their

⁵Note that, here, $N r(\mathbf{a}_k | \mathbf{S}) + \gamma \sum_{n \in \mathcal{N}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k+1})$ acts as the supervised terms (targets) in the training procedure of the value functions, and, thus, do not take derivations. Also, the limited situational awareness only allows the DBSs to define their individual value functions based on their local states (otherwise, significant coordination and overhead would be needed which is impractical for a drone network). The proposed mathematical induction and mini-batch training procedure allows the DBSs update their individual value functions based on their local states.

Algorithm 6 VD-RL algorithm for trajectory design at one realization of user access request.

Require: User locations, time constraints.

Ensure: Initialize value functions $\tilde{V}_{\theta_{c,n}^{(1)}}$, and policy functions $\pi_{\theta_{a,n}^{(1)}}$, for $n \in \mathcal{N}$.

```

1: for VD-RL training epoch  $i = 1 : I$  do
2:   The DBSs carry out an experience by generating a sequence of actions that are randomly selected based on
   current policy functions.
3:   Calculate advantage  $A$  based on (5.8).
4:   for Each DBS  $n = 1 : N$  do
5:     Update individual value function with (5.11).
6:     Calculate individual advantages  $\tilde{A}_n(a_{n,k}, s_{n,k})$ .
7:     Update policy function with (5.11).
8:   end for
9: end for

```

policy functions. Upon their return to the origin, the DBSs will exchange their achieved rewards, and update their policy and value functions by accumulating these rewards. Also, for a given DBS, carrying out an experience means that this DBS will serve one realization of user requests, and then, record its actions, states, and achieved rewards from this service. The recorded actions, states, and rewards constitute the experience of a given DBS.

5.3.1.4 Convergence and complexity analysis

Next, we show that the proposed VD-RL algorithm is guaranteed to converge to a local team optimal strategy.

Proposition 2. The proposed distributed VD-RL algorithm is guaranteed to converge to a local optimal solution of problem (6.5), if the following conditions are satisfied:

1. Value function $\tilde{V}_{\theta_{c,n}}(s_{n,k})$ converges to a local minimum with $\nabla_{\theta_{c,n}} A^2(a_k, S_k) = 0$, and advantage function $A(a_k, S_k) \neq 0$.
2.
$$\max_{a_{n,k} \in \mathcal{C}, s_{n,k} \in \mathcal{S}} \left| \frac{\partial \prod_{n \in \mathcal{N}} \pi_{\theta_{a,n}}(a_{n,k} | s_{n,k})}{\partial \theta_i \partial \theta'_j} \right| < \infty$$
, for any elements θ_i and θ_j in the policy function parameter vector $\theta_a = [\theta_{a,1}, \theta_{a,2}, \dots, \theta_{a,N}]$.
3. $\lim_{i \rightarrow \infty} \alpha_c^{(i)} = 0$, $\sum_{i=1}^{\infty} \alpha_c^{(i)} = \infty$, $\lim_{i \rightarrow \infty} \alpha_a^{(i)} = 0$ and $\sum_{i=1}^{\infty} \alpha_a^{(i)} = \infty$.

Proof. See Appendix B.1. □

From Proposition 1, we can see that the proposed VD-RL algorithm is guaranteed to converge to a local optimal solution of problem (6.5), by decomposing value $V(S_k)$ to each DBSs.

With regards to the complexity of the proposed VD-RL algorithm, one can see that only the successful service rate and estimated future reward, i.e. individual value, reached by each DBS at each step on their trajectories will be shared and transmitted among the DBSs. Thus, the VD-RL algorithm reduces the dimensionality of the problem by updating the policy and value functions at each DBS based only on this DBS's actions and states. Note that, the studied multi-agent problem is much more complex than a single agent problem, as the exponential growth on the action and state spaces in multi-agent problem causes a curse of dimensionality⁶. Nonetheless, our approach allows to reduce this multi-agent dimensionality to that of a single-agent problem. In essence, the complexity of the VD-RL solution is $\mathcal{O}(v(n_c + n_a)C)$, where v is the iteration when the proposed VD-RL algorithm converged, while n_c and n_a are the number of elements in $\theta_{c,n}$ and $\theta_{a,n}$. C is the time complexity of calculating the gradient of each element in $\theta_{c,n}$ and $\theta_{a,n}$. This complexity is considerably low since, as already mentioned, it is similar to that of a single agent policy gradient algorithm.

In summary, using the VD-RL algorithm, the DBSs update their own policy and value functions locally and can finally reach a local optimal solution of problem (6.5). However, when the environment changes, the local optimal strategy reached by the VD-RL algorithm will no longer be the strategy that maximizes the expected coverage. In this case, the whole procedure in Algorithm 6 must be performed again to solve problem (6.5). As the studied wireless environment is highly dynamic, the VD-RL algorithm in Algorithm 6 must be repeatedly performed to search a team optimal strategy in every environment. This increases the complexity of finding a VD-RL solution to a time complexity $\mathcal{O}(v(n_c + n_a)CX)$, when the user access requests change X times. In order to reduce this time complexity, we use the framework of meta-learning [139]. In particular, we propose a meta training procedure that can supplement the proposed algorithm by finding an initialization of the policy and value functions for the VD-RL solution. The meta-trained initialization are close to optimal policy and value functions at all possible environments which enables the VD-RL algorithm to quickly converge in dynamic environments.

5.3.2 Meta Training Procedure

Next, we introduce our meta-learning approach that meta trains a VD-RL solution using MAML [139]. The proposed meta training method seeks a VD-RL solution capable of quickly reaching team optimal strategies in various environments. This meta training procedure prepares the VD-RL solution with a set of well established initial policy and value functions with suitable estimation on a distribution of user request realizations, i.e. $p(\mathcal{Z})$. These initial policy and value functions are close to the optimal ones that provide team optimal strategies to each of the user request realizations in $\mathcal{Z}$. During the meta training procedure, a realization z_j is first sampled from $p(\mathcal{Z})$. Then, the DBSs collect experiences $e_{n,j}$ using their initial policy functions $\pi_{\theta_{a,n}^{(1)}}$, with the consideration that ground users are requesting access based on z_j . The policy and value functions at the DBSs are

⁶Curse of dimensionality means that the neural networks' estimation error increases when they deal with high dimensional data [150]

updated using Algorithm 6, based on the collected experiences $e_{n,j}$. The updated policy and value functions at DBS n within z_j are denoted as $\tilde{V}_{\theta'_{c,n,j}}$, and $\pi_{\theta'_{a,n,j}}$, respectively. The updated policy and value functions are then tested on new experiences $e'_{n,j}$ with actions sampled using $\pi_{\theta'_{a,n,j}}$, at current realization z_j . The feedback from such tests at DBS n will be the values of loss functions that measure the distance from the updated policy and value functions to the optimal policy and value functions that produce team optimal strategies. These are given by

$$\begin{aligned} \tilde{L}_{c,n}(\tilde{V}_{\theta'_{c,n,j}}, z_j) &= \sum_{k=1}^K \left(r \left(a_k^{(e'_{n,j})} \middle| s_k^{(e'_{n,j})} \right) \right. \\ &\quad \left. + \gamma \tilde{V}_{\theta'_{c,n,j}} \left(s_{n,k+1}^{(e'_{n,j})} \right) - \tilde{V}_{\theta'_{c,n,j}} \left(s_{n,k}^{(e'_{n,j})} \right) \right)^2, \end{aligned} \quad (5.12)$$

$$\begin{aligned} \tilde{L}_{a,n}(\pi_{\theta'_{a,n,j}}, z_j) &= \sum_{k=1}^K \left(r \left(a_k^{(e'_{n,j})} \middle| s_k^{(e'_{n,j})} \right) + \gamma \tilde{V}_{\theta'_{c,n,j}} \left(s_{n,k+1}^{(e'_{n,j})} \right) \right. \\ &\quad \left. - \tilde{V}_{\theta'_{c,n,j}} \left(s_{n,k}^{(e'_{n,j})} \right) \right) \log \pi_{\theta'_{a,n,j}} \left(a_{n,k}^{(e'_{n,j})} \middle| s_{n,k}^{(e'_{n,j})} \right), \end{aligned} \quad (5.13)$$

where $a_k^{(e'_{n,j})}$ and $s_k^{(e'_{n,j})}$ are, respectively, the k -th action and state of DBS n within experience $e'_{n,j}$. By minimizing the loss functions in (5.12) and (5.13), the proposed meta training method updates the policy and value function parameters toward the parameters of optimal policy and value functions in various user access request realizations sampled from $p(\mathcal{Z})$. More concretely, the objective of the meta training procedure is

$$\min_{\theta_c, \theta_a} \sum_{z_j \sim p(\mathcal{Z})} \sum_{n=1}^N \tilde{L}_{c,n}(\tilde{V}_{\theta'_{c,n,j}}, z_j) + \tilde{L}_{a,n}(\pi_{\theta'_{a,n,j}}, z_j), \quad (5.14)$$

where $\theta_c = [\theta_{c,1}, \theta_{c,2}, \dots, \theta_{c,N}]$ is the vector of value function parameters at all DBSs. $\theta'_{c,n,j} = \theta_{c,n} - \alpha_c \nabla_{\theta_{c,n}} \sum_{k=1}^K A^2(a_k^{(e_{n,j})}, s_k^{(e_{n,j})})$ is the updated value function parameters of DBS n , based on experience $e_{n,j}$. $\theta'_{a,n,j} = \theta_{a,n} + \alpha_a \sum_{k=1}^K \tilde{A}_n(a_{n,k}^{(e_{n,j})}, s_{n,k}^{(e_{n,j})}) \nabla_{\theta_{a,n}} \log \pi_{\theta_{a,n}}(a_{n,k}^{(e_{n,j})} | s_{n,k}^{(e_{n,j})})$ is the updated policy function parameter of DBS n , based on experience $e_{n,j}$. Here, we can see that the optimization variables in (5.14) are initial policy and value function parameters θ_c and θ_a . That is, the meta-learning approach seeks to find the optimal initial function parameters of the VD-RL algorithm so as to minimize the estimation errors captured in (5.12) and (5.13), for different environments that can be sampled from $p(\mathcal{Z})$. This parameter initialization, i.e. the meta-trained policy and value functions, will make it easier for the VD-RL solution to find the team optimal strategies in unseen environments.

To solve problem (5.14), a standard stochastic gradient descent method is applied. Fig. 5.3 shows one iteration for solving problem (5.14). For this example, at each meta training iteration, the meta training method performs one step stochastic gradient descent on (5.14) over the initial policy and value function parameters $\theta_{c,n}$ and $\theta_{a,n}$ based on sampled experiences in J different environments.

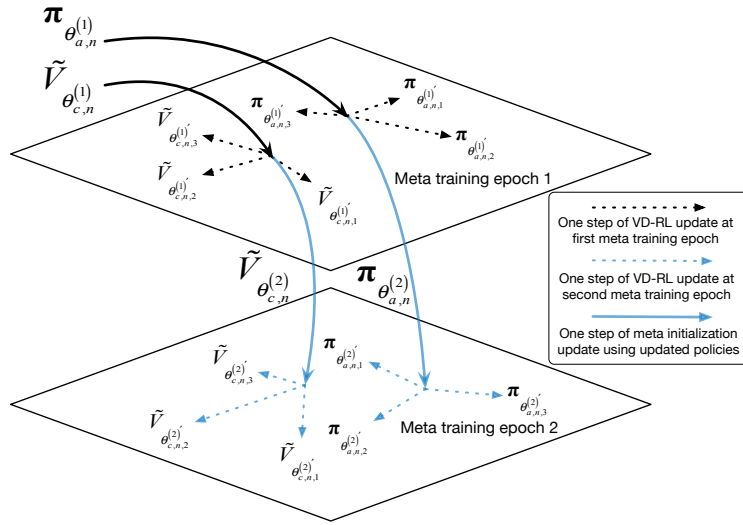


Figure 5.3: Illustration of the proposed meta-training method at DBS n with $J = 3$. Here, 3 different user requests are drawn from $p(\mathcal{Z})$ at each training episode. $\tilde{V}_{\theta_{c,n,1}^{(1)'}}$ and $\pi_{\theta_{a,n,1}^{(1)'}}$ are the updated policy and value functions in the first sampled environment,

The full procedure of using meta training method with VD-RL is summarized in Algorithm 7.⁷ In this meta training procedure, the DBSs collect their experiences from serving J different user requests drawn from $p(\mathcal{Z})$, with initial value functions $\tilde{V}_{\theta_{c,n}^{(1)}}$ and policy functions $\pi_{\theta_{a,n}^{(1)}}$. After serving one sampled user request j , the DBSs execute one step VD-RL update and get new value functions $\tilde{V}_{\theta_{c,n,j}^{(1)'}}$ and policy functions $\pi_{\theta_{a,n,j}^{(1)'}}$. The updated policy and value functions are evaluated by the loss functions defined in (5.12) and (5.13), based on experiences collected with strategy $\pi_{\theta_{a,n,j}^{(1)'}}$. The meta training procedure in Algorithm 7 seeks to finding a set of initial policy and value functions that are close to the optimal policy and value functions at every user request realization. Equipped with this initialization, the VD-RL can now find a local optimal solution of problem (6.5) in unseen environment using a small number of update steps in Algorithm 6. Thus, the meta-training method minimizes the losses that are collected from the sampled user request realizations, by updating the value and policy parameters, i.e. $\theta_{c,n}^{(1)}, \theta_{a,n}^{(1)}$, in the opposite direction of the gradient of the losses. After one step of this update, the value and policy parameters, i.e. $\theta_{c,n}^{(2)}, \theta_{a,n}^{(2)}$, will be closer to the optimal policy and value functions in the sampled environment. Subsequently, the DBSs will start serving user requests with value functions $\tilde{V}_{\theta_{c,n}^{(2)}}$, and policy functions $\pi_{\theta_{a,n}^{(2)}}$. Then, the DBSs collect their experience from serving different user requests, based on which they can keep updating the policy and value functions, as shown in Fig. 5.3, until convergence⁸.

⁷Note that, as the meta training procedure seeks to learn a user request distribution $p(\mathcal{Z})$ in Algorithm 7, all the user requests are sampled from the same distribution $p(\mathcal{Z})$.

⁸Here, we need to point out that, a machine learning model reaches convergence when its training loss falls within an error range around the final value. Thus, we consider that the meta training model has converged when additional training will not improve the model, at which point the meta training procedure is completed.

Using our proposed meta training procedure, the DBSs finally find a VD-RL solution with optimal initializations of the policy and value functions, i.e. $\tilde{V}_{\theta_{c,n}^*}$, and $\pi_{\theta_{a,n}^*}$. The complexity of the meta training procedure is $\mathcal{O}(2v_m(n_c + n_a)CJ)$, with v_m being the number of iterations that the meta training method needs for convergence. This complexity is considerably low, as the proposed meta training mechanism does not need any additional neural networks in the training procedure. Also, since the DBSs can collect experiences on serving ground user's daily requests, the proposed meta training procedure's complexity will not introduce additional costs in terms of time, energy, or DBS hardware.

5.4 Simulation results

For our simulations, we consider a scenario with five DBSs serving $U = 300$ mobile users. The main simulation parameters are listed in Table I. In particular, we assume the quantity of each user request follows uniform distribution over the interval $[20, 600]$ Mbits, the occurrence of each user request follows Gaussian distribution with a standard deviation 1, and a mean ranges from 0 to 40 minutes. At every independent run of the Monte Carlo experiment studied in this section, we repeatedly deploy the users at some random locations that are uniformly distributed in every cluster, and sample a user access request realization from the assumed distribution. The value and policy functions at each DBSs are formulated by feed forward neural networks. The results of the proposed VD-RL algorithm are compared with the IAC [102] and the monotonic value function factorisation algorithm, called Q mix [103]. The IAC algorithm updates one policy function and one value function for all of the agents. It allows each DBS updates their policies relying only on their local actions and states, but toward the DBSs' maximal individual benefits. The comparison between the proposed algorithm and the IAC algorithm shows the gains one can get from the value decomposition scheme. The Q mix algorithm uses the same value and policy update procedure as our proposed VD-RL algorithm, but estimates the summation in (5.9) using a neural network. The results of the proposed meta-trained VD-RL are further compared to the results from the pre-trained VD-RL algorithm [151], the original VD-RL algorithm, as well as oracle results. Within the pre-trained VD-RL algorithm, the VD-RL algorithm starts from the optimal policy and value functions at former tasks. Here, the pre-trained VD-RL algorithm trains a VD-RL within exactly the same environments sampled at our meta training procedure. For the original VD-RL algorithm, the DBSs start with randomly initialized policy and value functions. The oracle results present the DBSs' performance with optimal strategies in the considered environment. All statistical results are averaged over a large number of independent runs.

Fig. 5.4 shows how successful service rates increase with time, in a network with two DBSs. From Fig. 5.4, we can see that the successful service rate achieved by each DBS increases when the DBS arrives at a new cluster $\xi_{n,k}$. Fig. 5.4 also shows that the two DBSs spend different amount of time on flying across clusters and serving clusters. In other words, the DBSs operate asynchronously in the considered area. In particular, when DBS 2 starts to serve its second cluster, DBS 1 is still serving its first cluster.

Table 5.1: Simulation Parameters [2]

Parameter	Value	Parameter	Value
P_u	20 dBm	V_s	30 m/s
N_0	-170 dBm/Hz	B	1 MHz
$\mu_{\text{LoS},}$	1.6	δ_{LoS}	8.41
$\mu_{\text{NLoS},}$	23	δ_{NLoS}	33.78

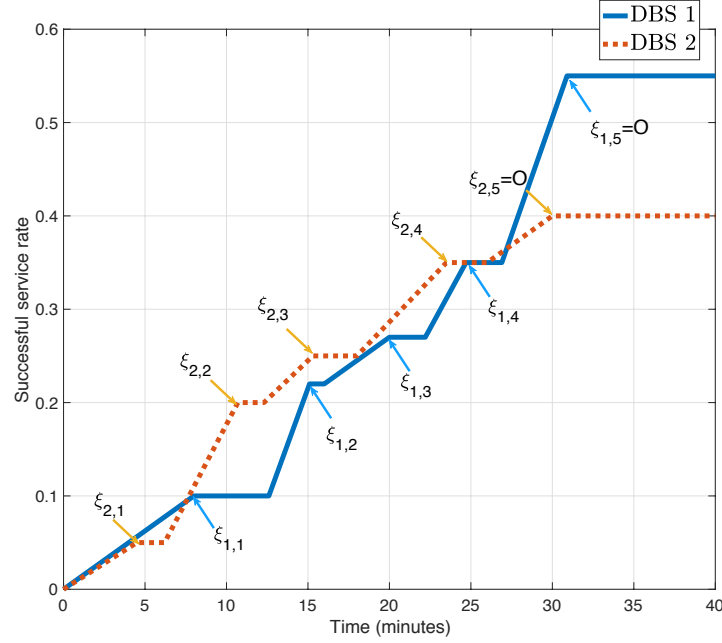


Figure 5.4: Successful service rate achieved at each DBS over time.

Fig. 5.5 shows how the decomposition of the value function $V(\mathcal{S}_k)$ is learned using the proposed VD-RL algorithm. Here, the value of function $V(\mathcal{S}_0)$ is the estimation of the discounted future reward at the initial state $\mathcal{S}_0$. In Fig. 5.5, the value of function $V(\mathcal{S}_0)$, as well as the individual values of functions $\tilde{V}_{\theta_{c,1}}(\mathcal{S}_0)$ and $\tilde{V}_{\theta_{c,2}}(\mathcal{S}_0)$ follow the same trend of as the team utility of all DBSs. This implies that the individual value functions also estimate the team utility. With such values, the DBSs can independently find the strategies that maximize the team utility without sharing information such as actions and states. In particular, at the convergence of the VD-RL algorithm, DBS 2 sacrifices its own utility for team optimality. From Fig. 5.5, we can also observe that, with the VD-RL algorithm, the individual value at DBS 1 keeps increasing, yet the successful service rate at DBS 2 increases at first and then decreases while reaching convergence. The sum of individual values at two DBSs always equals to the value of $V(\mathcal{S}_k)$, which follows the trend of the team utility. This is due to the fact that, in the proposed VD-RL algorithm, each DBS is rewarded with the utility achieved by all the DBSs, i.e. team utility. Thus, the DBSs are training their own policies toward a maximal team utility instead of their own utilities. In this case, DBS

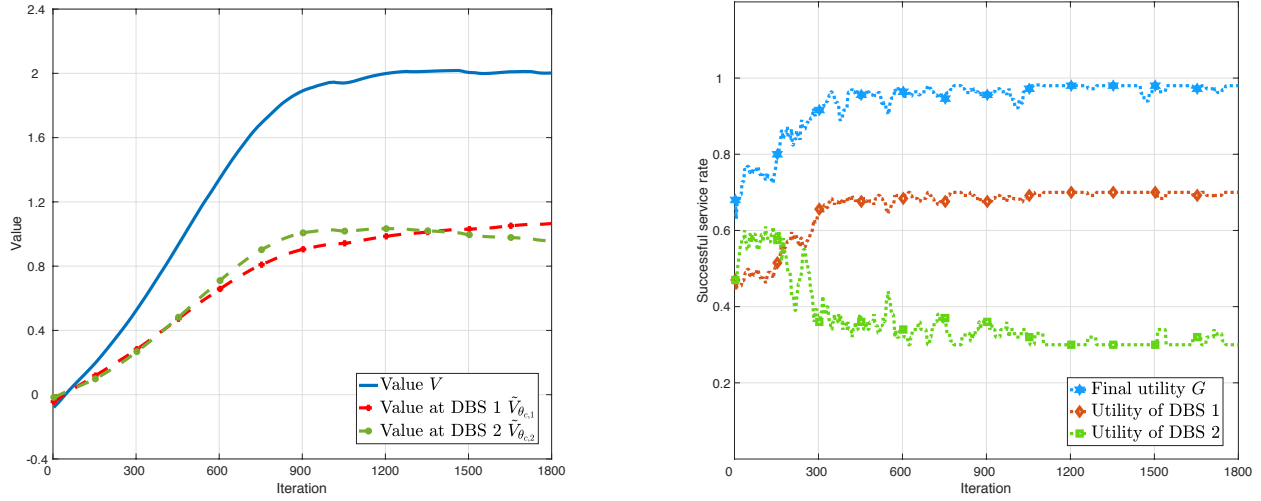


Figure 5.5: The value decomposition learned by the proposed VD-RL algorithm.

2 will be prone to choose policies that yield a lower individual utility but a higher team utility. In other words, the VD-RL algorithm learns to decompose the value function into two individual values, each of which can direct each DBS to independently find its strategy that maximizes the team utility, instead of its individual utility.

Fig. 5.6 shows the value decomposition within 5 VD-RL training iterations. In this figure, we can see that the value of $V(S_0)$ increases at the first step within each VD-RL training iteration. This is because, function $V(S_0)$ estimates the cumulative future team reward that the DBSs can get at the initial state. Thus, the DBSs update value function $V(S_0)$ using the rewards that they achieved from serving the ground users at the last training iteration, once they return to the origin. Fig. 5.6 also shows that the decomposed individual value $\tilde{V}_{\theta_{c,2}}(S_0)$ at DBS 2 is much larger than the individual value $\tilde{V}_{\theta_{c,1}}(S_0)$ at DBS 1. This is because DBS 2 selects the actions that are beneficial for the team utility. A large individual value $\tilde{V}_{\theta_{c,2}}(S_0)$ reinforces such actions to maintain a high team utility. However, DBS 1 does not select actions that improve the team utility. Hence, the individual value $\tilde{V}_{\theta_{c,1}}(S_0)$ attributed to DBS 1 is small, which keeps DBS 1's strategy unchanged. DBS 1 will, thus, act "greedily" by selecting all possible actions with its current strategy, until it finds the optimal one.

In Fig. 5.7, we show the convergence of the proposed VD-RL algorithm. In this figure, we can see that the proposed VD-RL algorithm requires approximately 1,300 iterations to reach convergence, which improves the convergence speed by up to 30.6% compared to the Q mix algorithm. This stems from the fact that the neural network used to estimate the summation in (5.9) remarkably increases the complexity of Q mix. Meanwhile, Fig. 5.7 also shows that the proposed VD-RL algorithm yields a 53.2% higher final utility than IAC. This is because the VD-RL algorithm can find a team optimal strategy to maximize the team reward. The IAC algorithm, however, finds a strategy that maximizes the DBSs' individual utilities. From Fig. 5.7, we also observe that the

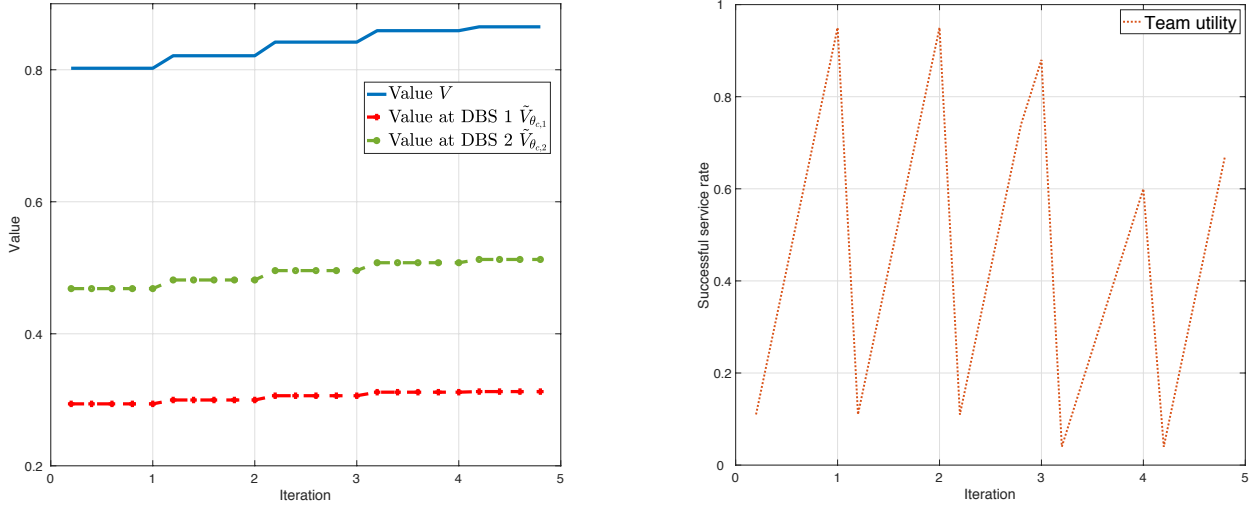


Figure 5.6: Value decomposition within 5 VD-RL iteration.

proposed VD-RL algorithm has a similar convergence speed to the IAC algorithm. This stems from the fact that, in the VD-RL and IAC algorithm, each DBS updates the policy and value functions using its own experience, and, thus, the proposed VD-RL algorithm and the IAC algorithm all have low time complexities that are comparable to a single agent RL algorithm.

Fig. 5.8 shows a snapshot of the trajectories resulting from the proposed meta-trained VD-RL, and the pre-trained VD-RL algorithm in an environment that was not experienced during the meta training procedure. Here, the trajectories in Figs. 5.8b, 5.8c, 5.8e, and 5.8f are selected, respectively, at the 100-th iteration of both considered algorithms. In this figure, we can see that, when faced with an unseen environment, the meta-trained VD-RL scheme can effectively find the optimal trajectories for the DBSs within a considerably small number of iterations. However, the trajectories resulting from the pre-trained VD-RL are still far from the optimal trajectories. Fig. 5.8 also shows that the proposed meta-training method can adapt the DBSs in unseen environments much faster than the pre-trained VD-RL algorithm. This is because the meta-training method can find a policy and value function parameter initialization that is close to optimal policy and value functions at all possible user requests in $p(\mathcal{Z})$. By using this meta-learning-based initialization, the meta-trained VD-RL can reach the optimal policy, value functions and find a team optimal strategy in an unseen environment within $p(\mathcal{Z})$, using a small number of iterations.

In Fig. 5.9, we show the convergence of the meta-trained VD-RL. From this figure, we can see that, with initial value functions $\tilde{V}_{\theta_{c,n}}^*$ and policy functions $\pi_{\theta_{a,n}}^*$ provided by the meta training procedure in Algorithm 7, the meta-trained VD-RL converges at approximately the 700-th iteration, which is 36.4%, and 53.8% faster than the pre-trained VD-RL algorithm and the original VD-RL algorithm. Fig. 5.9 also shows that the meta-trained VD-RL achieves a successful service rate that is equal to the one reached by the original VD-RL algorithm. The successful service rate achieved by the meta-training method is 9.2% and 53.2% higher than the one achieved by,

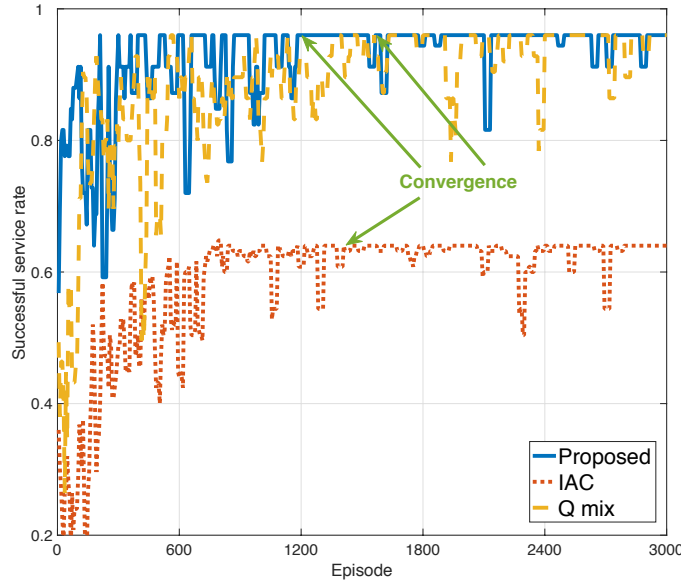


Figure 5.7: Convergence of all considered multi-agent RL algorithms.

respectively the pre-trained VD-RL algorithm and the IAC algorithm. This gain stems from the fact that the meta-training method trains a set of policy and value functions with proper estimation on various user access requests. By starting the VD-RL policy updating procedure from such initialization, the DBSs can find a team optimal strategy within a small number of policy update steps. The pre-trained VD-RL, however, converges to a much lower successful service rate, since it could make the DBSs to start their policy updating procedure from the initializations that are far from the optimal policy and value functions.

Fig. 5.10 shows the convergence of the proposed meta-training method as the number of DBSs varies. Fig. 5.10 shows that, as the number of DBSs increases, the number of iterations needed for convergence increases because of the associated growth in the size of the action and state spaces. From Fig. 5.10, we observe that, as the number of DBSs increases, the number of iterations needed for convergence will grow at a higher speed. This is because the action and state spaces within the problem increase exponentially with the number of DBSs. This is consistent with the complexity analysis in Section III. A. Moreover, the number of iterations that the Q mix and the proposed VD-RL algorithm need for convergence increases much slower compared to the traditional IAC algorithm. This is because the value decomposition scheme reduces the dimensionality of the action and state spaces in the considered problem. In particular, the value decomposition scheme enables each DBS to update its strategy using its own actions and states, thus simplifying the considered problem. However, the neural networks deployed for value decomposition within the Q mix algorithm introduces extra complexity at the DBSs, especially in simpler system as shown in Fig. 5.10. The convergence speed of the meta-trained VD-RL scheme decreases relatively slower compared to the original VD-RL algorithm, since the initializations of the policy and value function

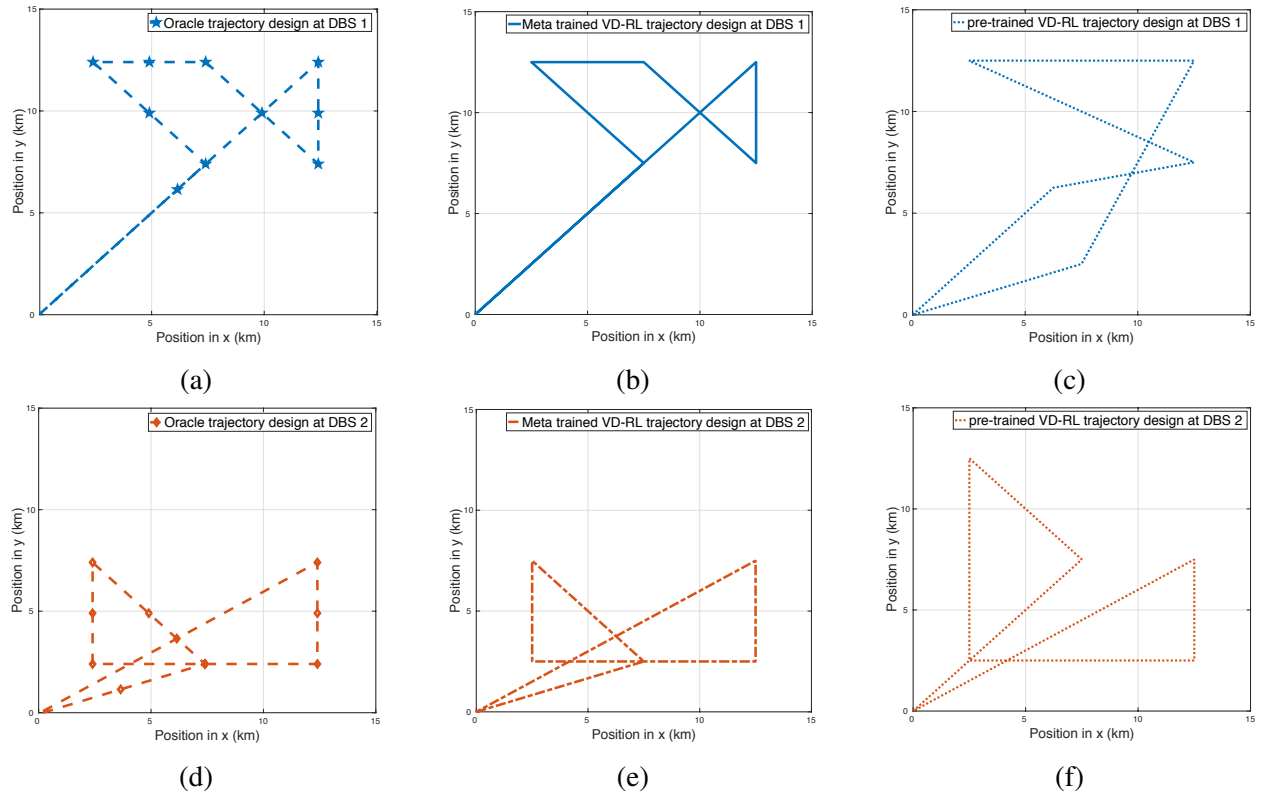


Figure 5.8: Snapshots of trajectories resulting from all meta-trained and pre-trained algorithms in a sampled unseen environment.

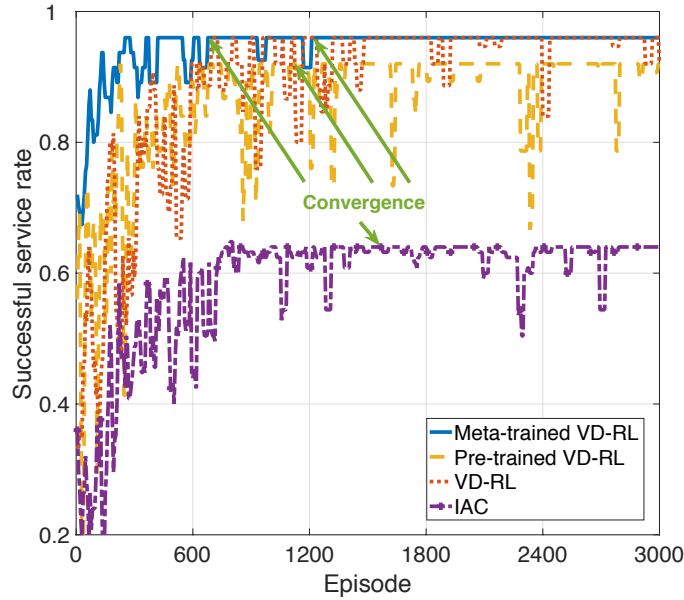


Figure 5.9: Convergence of the meta-trained VD-RL.

parameters in the meta-trained VD-RL algorithm have optimized performance across diverse user access request realizations. Moreover, the pre-trained VD-RL algorithm only slightly improves the VD-RL algorithm's convergence speed, as the initialization of pre-trained VD-RL algorithm could be far from the optimal policy and value functions in unseen environments.

Fig. 5.11 shows how the successful service rate achieved by the DBSs varies with the number of DBSs and the number of clusters. From Fig. 5.11, we can see that the successful service rate provided by the DBSs increases with the number of deployed DBSs, but it decreases with the number of active clusters. This is because a larger number of DBSs can lead to a better coverage to the ground users. Fig. 5.11 also shows that the successful service rate increases more slowly when the number of deployed DBSs increases, or when the number of clusters is smaller. This is because the number of covered user requests increases when there are more DBSs available. Fig. 5.11 shows that the successful service rate also decreases more slowly when the number of clusters in the simulated area is over 11. This is because the SLF time of the DBSs decreases when we have more clusters. From Fig. 5.11, we can also see that the meta-trained VD-RL algorithm yields a better coverage to the ground users, in particular, it yields a successful service rate that is 5.6% higher than the one resulting from the pre-trained VD-RL algorithm, 9.2% higher than the one resulting from the VD-RL algorithm, and 53.2% higher than the one resulting from the IAC algorithm. This is due to the fact that, by reducing the losses defined in (5.12) and (5.13) across various environments, the proposed meta-training method can find a set of policy and value functions that are close to the optimal policy and value functions at all possible user requests in $p(\mathcal{L})$. This guarantees a faster convergence to a team optimal strategy in an unseen environment. Note that, although the meta-training method only yields small improvements to the successful

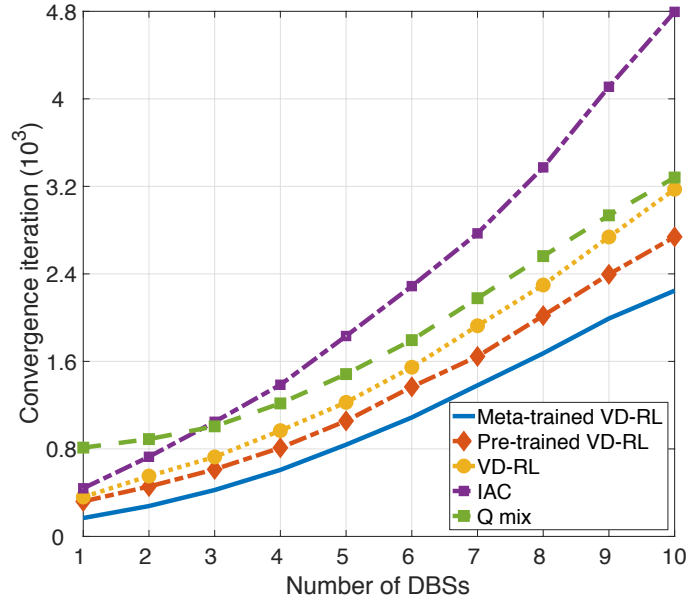


Figure 5.10: Convergence of the meta-trained VD-RL algorithm as the number of DBSs varies.

service rate, it can converge much faster than the VD-RL algorithm.

5.5 Conclusion

In this chapter, we have studied the problem of trajectory design for a group of DBSs in unpredictable, dynamic environments. In the considered system, the DBSs cooperatively fly around the considered environment to provide on demand uplink communication service to ground users. We have formulated the studied problem in an optimization setting and have proposed a VD-RL algorithm to solve this problem. The proposed VD-RL algorithm makes the DBSs independently update their individual strategies toward the maximal coverage in the system, by sharing only its utility and value to other DBSs. To improve the convergence speed of the VD-RL algorithm in unseen environments, we have also proposed a meta-training method to optimize the initializations in a VD-RL solution. Simulation results show that the proposed VD-RL algorithm with meta training mechanism outperforms the traditional MARL algorithms.

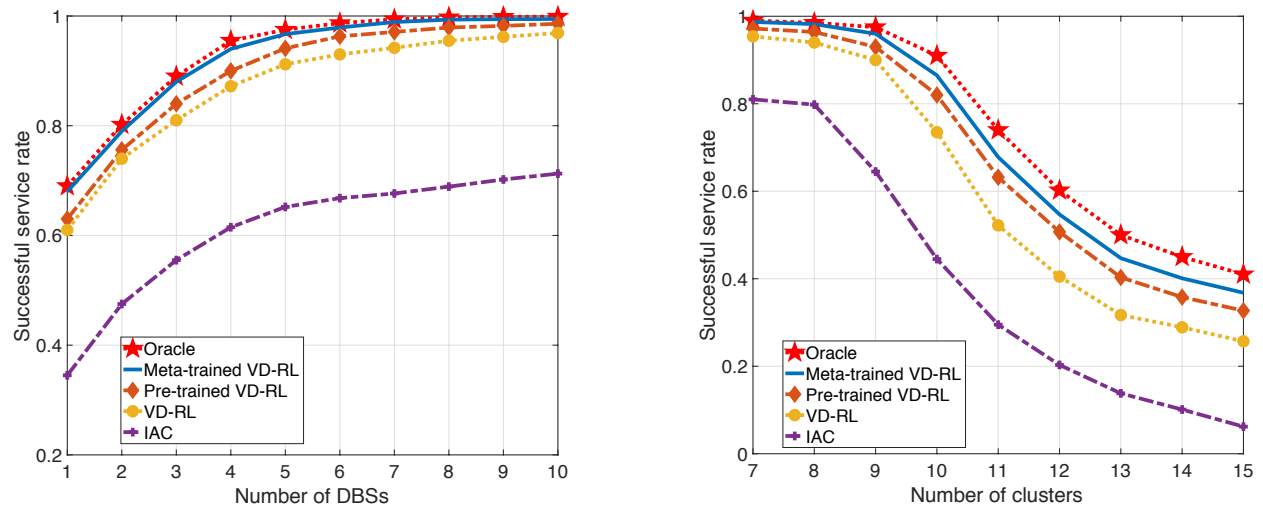


Figure 5.11: Successful service rate as the number of DBSs varies.

Algorithm 7 Meta training procedure for the trajectory design problem.

Require: A distribution of user requests $p(\mathcal{Z})$, user locations, time constraints.

Ensure: Initialize value functions $\tilde{V}_{\theta_{c,n}^{(1)}}$, and policy functions $\pi_{\theta_{a,n}^{(1)}}$, for $n \in \mathcal{N}$.

- 1: **for** Meta training epoch $i = 1 : I$ **do**
- 2: **for** Environment sampling epoch $j = 1 : J$ **do**
- 3: Sample an user request realization $\mathbf{z}_j \sim \mathcal{Z}$.
- 4: Carry out an experience by generating a sequence of actions that are randomly selected based on current policy functions, i.e. $\pi_{\theta_{a,n}^{(i)}}$, for $n \in \mathcal{N}$.
- 5: Calculate advantage A based on (5.8).
- 6: **for** Each DBS $n = 1 : N$ **do**
- 7: Update individual value functions

$$\theta_{c,n,j}^{(i)'} = \theta_{c,n}^{(i)} + \alpha_c \nabla_{\theta_{c,n}^{(i)}} \sum_{k=1}^K A^2 \left(\mathbf{a}_k^{(e_{n,j})}, \mathbf{s}_k^{(e_{n,j})} \right),$$

and policy functions

$$\begin{aligned} \theta_{a,n,j}^{(i)'} &= \theta_{a,n}^{(i)} + \alpha_a \sum_{k=1}^K \tilde{A}_n \left(\mathbf{a}_{n,k}^{(e_{n,j})}, \mathbf{s}_{n,k}^{(e_{n,j})} \right) \\ &\quad \nabla_{\theta_{a,n}^{(i)}} \log \pi_{\theta_{a,n}^{(i)}} \left(\mathbf{a}_{n,k}^{(e_{n,j})} \mid \mathbf{s}_{n,k}^{(e_{n,j})} \right). \end{aligned}$$

- 8: Carry out an experience by generating a sequence of actions that are randomly selected based on updated policy functions, i.e. $\pi_{\theta_{a,n,j}^{(i)'}}$.
 - 9: Calculate loss $\tilde{L}_{c,n} \left(\tilde{V}_{\theta_{c,n,j}^{(i)'}} , \mathbf{z}_j \right)$ and $\tilde{L}_{a,n} \left(\pi_{\theta_{a,n,j}^{(i)'}} , \mathbf{z}_j \right)$ with (5.12) and (5.13).
 - 10: **end for**
 - 11: **end for**
 - 12: **for** Each DBS $n = 1 : N$ **do**
 - 13: Update initial value parameters $\theta_{c,n}^{(i+1)} = \theta_{c,n}^{(i)} - \beta \nabla_{\theta_{c,n}^{(i)}} \sum_{j=1}^J \tilde{L}_{c,n} \left(\tilde{V}_{\theta_{c,n,j}^{(i)'}} , \mathbf{z}_j \right)$, and policy parameters $\theta_{a,n}^{(i+1)} = \theta_{a,n}^{(i)} - \beta \nabla_{\theta_{a,n}^{(i)}} \sum_{j=1}^J \tilde{L}_{a,n} \left(\pi_{\theta_{a,n,j}^{(i)'}} , \mathbf{z}_j \right)$.
 - 14: **end for**
 - 15: **end for**
 - 16: **return** Optimal initial value functions $\tilde{V}_{\theta_{c,n}^*}$, and initial policy functions $\pi_{\theta_{a,n}^*}$, for $n \in \mathcal{N}$.
-

Chapter 6

Meta-Reinforcement Learning for Trajectory Design in Wireless UAV Networks

The cooperation among DBSs within NTN was studied in Chapter 5 to provide maximal on-demand service to users with dynamic needs. In this chapter, the adjustment on the trajectory of a DBS under environmental dynamics is studied using online cross validation. Indeed, learning a DBS trajectory that allows it to provide timely on-demand service is particularly challenging within highly dynamic wireless environments. In this chapter, the DBS learns to service the unforeseeable needs from ground users with RL solutions, and adjusts its learning procedure for any possible changes on the wireless environment.

6.1 Background, Related Works, and Contributions

The existing literature such as in [8, 71, 96] has studied a number of problems related to trajectory design for drone-based systems. The work in [71] studied the problem of wireless-aware trajectory optimization for drone users using reinforcement learning. The authors in [8] propose a trajectory design that enhances physical layer security. The work in [96] studies how the 3D antenna radiation pattern and backhaul constraint affect a drone's 3D trajectory design. Despite their promising results, these existing works [8, 71, 96] do not consider the design of drone trajectory in a network where the service requests from ground users vary over time. Meanwhile, prior works such as [152] and [74] study the use of machine learning for drone trajectory optimization in dynamic environments. The authors in [152] use machine learning to deploy a drone to provide on-demand wireless service to cellular users. The authors in [74] propose a multi-agent Q-learning-based algorithm to design trajectories for multiple UAVs based on the prediction of users' mobility. However, the RL-based solutions in [152] and [74] are mostly over-fitted to certain environments,

and, thus, they cannot rapidly cope with new or highly-dynamic environments. In particular, the internal dynamics of the RL in [152] and [74], such as hyper-parameters, exploration strategies, and loss functions, are manually pre-determined to help the algorithms to design trajectories in specific environments. As such, even the slightest changes in the environment could introduce significant negative influence on the RL performance. In contrast, a meta-learning based trajectory design solution [153] that automatically tunes the RL internal dynamics would be more appropriate for dynamic wireless environments. Meta-learning, also known as “learning to learn”, provides a solution that adapt the DBS to new environments rapidly with only a few training examples.

The main contribution of this chapter is, thus, a novel meta-learning approach for optimizing the trajectory of a DBS while considering the uncertainty and dynamic of terrestrial users’ service request. In particular, we consider a trajectory design problem that enables a DBS to effectively provide on-demand uplink service to users and solve it with a meta-reinforcement learning based algorithm. While prior works such as [154–156] used meta-learning to study various wireless problems, those works have not considered the DBS trajectory design problem. In contrast, here, we introduce a learning solution tailored to the drone-aided system whose goal is to provide on time service to ground users. Unlike previous meta-learning solutions such as [155], and [156], which aim at performing prediction with less training examples, we propose a novel, lightweight meta-learning approach that adapts the DBS to unseen environments. The proposed meta-learning based solution tunes the hyper-parameters of RL algorithm with online cross validation. An approximation on the hyper-parameter update direction is also proposed to reduce the computational complexity. Simulation results show that, by tuning the hyper-parameters, the proposed meta-learning solution yields a 25% improvement in the convergence speed, and about 10% improvement in the DBS’ communication performance, compared to a vanilla policy gradient algorithm. With the proposed algorithm, the probability that the DBS serves over 50% of user requests increases about 27%. The computational complexity of the proposed algorithm is also shown to be similar to that of the vanilla policy gradient algorithm.

The rest of this chapter is organized as follows. The system model and problem formulation are described in Section 6.2. In Section 6.3, the proposed algorithm is discussed. In Section 6.4, numerical simulation results are presented. Finally, conclusions are drawn in Section 6.5.

6.2 System Model and Problem Formulation

Consider a geographical area within which a set $\mathcal{U}$ of U randomly deployed terrestrial users request uplink data services. A fixed-wing DBS is dispatched to satisfy the uplink access requests of the terrestrial users, as shown in Fig. 6.1. In such an area, a given user that is requesting data services is called an *active user*, otherwise, it is called an *inactive user*. We assume the users to be separated into several groups, each of which is called a *cluster*. The set of all possible clusters is denoted as $\mathcal{L}$. The DBS will travel among the groups in a steady SLF, and it uses a circle to hover over each group, in a steady circular flight, with constant speed V and constant altitude

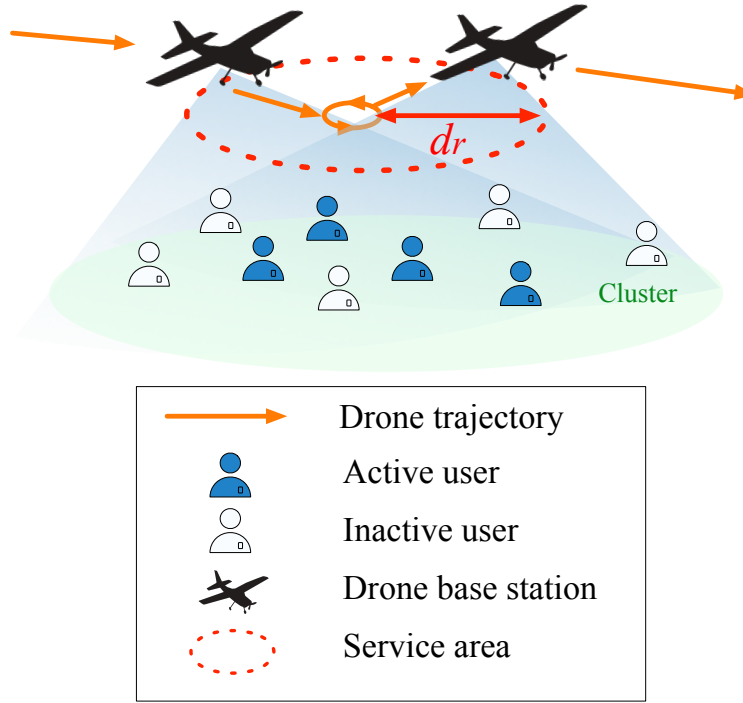


Figure 6.1: Network topology.

H [92]. Meanwhile, the DBS must return to its original location O within a time period T for battery charging. The trajectory that traces the DBS's movement within the studied time duration is captured by a vector $\xi = [l_1, \dots, l_K]^\top$, with $l_k \in \mathcal{L}$ being the k -th cluster that the DBS serves and K being the number of served clusters along the trajectory. The set of all possible trajectories of the DBS is given by $\mathcal{E}$. Let $\pi(l | l_k, \tau_k)$ be the probability that the DBS moves toward cluster $l \in \mathcal{L}$ after arriving at l_k at time epoch $T - \tau_k$, in which τ_k is the remaining time for the DBS to return to the origin, after serving cluster l_k .

6.2.1 Communication Performance Analysis

In our network, the users adopt an OFDMA technique and transmit data over a set of uplink RBs. The DBS will arbitrarily allocate one RB to each one of its associated users within a cluster. We assume that the DBS can keep serving its associated users within a d_r -meter radius over each cluster as shown in Fig. 6.1. The area within this range is called a *service area*. Also, user u is assumed to request a total of b_u data (in bits) at time epoch t_u . The DBS must satisfy the user's request within the studied duration T . Let $\mathbf{b} = [b_1, \dots, b_U]$ and $\mathbf{t} = [t_1, \dots, t_U]$ be, respectively, the vector of quantity and occurrence of the users' access request in the network. Here, the quantity b_u and active time t_u are assumed to be independent random variables that follow unknown distributions. Every possible $\mathbf{b}$ and $\mathbf{t}$ that could be observed in this network is called one *realization* of the users' access request. In this model, the communication performance variation caused by the DBS's movement within the service area is considered to be negligible, as the variation on the

DBS-user distance is considerably small with d_r being much smaller than H . The SNR at the link between the DBS and user u , will thus be

$$\gamma_u = \frac{Ph_{u,\phi_u}}{\sigma^2}, \quad (6.1)$$

where P represents the transmit power of user u , which is assumed to be equal for all users. h_{u,ϕ_u} is the path loss between user u and the DBS, with ϕ_u is the resource block allocated to user u . Here, the Nakagami channel model is applied to characterize a wide range of fading environments for the link between user u and the DBS, such that $h_{u,\phi_u} = \varepsilon d_u^{-\alpha}$ with ε following Gamma distribution with shape parameter m , d_u being the distance between user u and the DBS, and α being the path loss exponent. The data rate at the link between the DBS and user u will then be $c_u = B \log(1 + \gamma_u)$ with B being the RB bandwidth (equal for all RBs).

6.2.2 Utility Function Model

In the studied network, the transmission delay of user $u \in \mathcal{U}$, when being served, is $D_u = \frac{b_u}{c_u}$. In this case, the time that DBS consumes to hover over cluster l is given by

$$D_l^* = \max_{u \in \mathcal{U}_l} D_u - \frac{2d_r}{V}, \quad (6.2)$$

where $\mathcal{U}_l$ is the set of active users within cluster l . Note that, here, $\max_{u \in \mathcal{U}_l} D_u$ is the time the DBS consumes within the service area of cluster l . D_l^* and $\frac{2d_r}{V}$ are the time the DBS consumes for hovering, and SLF, respectively, within the service area. Meanwhile, the DBS has to return to the origin within the considered, fixed time duration T , that is $t_0(\xi) = \sum_{k=0}^K \frac{d_{l_k, l_{k+1}}}{V} + \sum_{k=1}^K D_{l_k}^* \leq T$, where t_0 is the time when the DBS returns to origin O along trajectory ξ , with $d_{l_k, l_{k+1}}$ being the distance between cluster l_k and l_{k+1} . Also, l_0 and l_{K+1} represent, respectively, the DBS's original and final locations. The available time that the DBS can use to return to origin after serving cluster l_k on trajectory ξ is $\tau_k = T - \sum_{\kappa=0}^{k-1} \frac{d_{l_\kappa, l_{\kappa+1}}}{V} - \sum_{\kappa=1}^k D_{l_\kappa}^*$.

However, with one DBS serving the whole area, some users may not have the opportunity to finish their data transmission during the studied period. In such a case, the utility function of the DBS along trajectory ξ is defined as the *service success rate*, which captures the fraction of users being served in the studied duration, and can be given by

$$\mu(\xi) = \frac{\sum_{k=1}^K \sum_{u \in \mathcal{U}} \mathbb{1}_{\{u \in \mathcal{U}_{l_k}, t_u \leq T - \tau_k\}}}{\sum_{u \in \mathcal{U}} \mathbb{1}_{\{0 \leq t_u \leq T\}}}, \quad (6.3)$$

where $\mathbb{1}_{\{x\}} = 1$ when x is true, otherwise, we have $\mathbb{1}_{\{x\}} = 0$. Here, $\sum_{u \in \mathcal{U}} \mathbb{1}_{\{0 \leq t_u \leq T\}}$ is the number of users request data service within the studied time duration, and $\sum_{u \in \mathcal{U}} \mathbb{1}_{\{u \in \mathcal{U}_{l_k}, t_u \leq T - \tau_k\}}$ is the number of served user with cluster l_k . The value of the utility function that can be achieved with

policy π is given by

$$\bar{\mu}(\pi) = \sum_{\xi \in \mathcal{E}} \mu(\xi) \prod_{k=1}^{K-1} \pi(l_{k+1} | l_k, \tau_k), \quad (6.4)$$

where π defines the probabilities of each of DBS' moving direction at different cluster and time epoch.

6.2.3 Problem Formulation

Our goal is to find an optimal DBS trajectory that can effectively serve all the users. Thus, we formulate an optimization problem whose objective is to find the policy that maximizes the expected utility of the network

$$\max_{\pi} \bar{\mu}(\pi), \quad (6.5)$$

$$\text{s. t. } t_0(\xi) \leq T, \xi \in \mathcal{E}, \quad (6.5a)$$

$$0 \leq \pi(l_{k+1} | l_k, \tau_k) \leq 1, 1 \leq k \leq K-1, \quad (6.5b)$$

where (6.5a) means that the DBS must return to the origin before time epoch T . From (6.5), we can see that, the network utility is determined by the DBS's policy, which is, in turn, decided based on the quantity and occurrence of users' access requests. In essence, each realization of user requests needs a novel optimal trajectory design for the DBS. However, the user requests in the studied network is dynamic and unpredictable. In order to provide best coverage to such users, the DBS must find an algorithm that can effectively guide it in various environments. Here, we notice that the use of optimization algorithms, such as branch and bound or nonlinear programming, is not suitable to solve (6.5), as the values of $\mathbf{b}$ and $\mathbf{t}$ follow unknown distributions. Traditional machine learning algorithms [136] are also not suitable to solve (6.5), as most of them are overfitted to their training tasks, and, thus they are not capable of guiding the DBS in unseen tasks. To effectively solve (6.5), we propose a meta-reinforcement learning (MRL) algorithm that prepares the DBS for novel tasks by tuning an RL solution for each task. In particular, the MRL solution adaptively tunes the RL hyper-parameters so as to guide the DBS in unseen wireless environments. This, in turn, can guarantee fast convergence to maximal service success rate. The proposed MRL algorithm is introduced in the next section.

6.3 Meta Gradient Policy Gradient Meta-Reinforcement Learning

We now introduce an MRL algorithm, called meta-gradient policy gradient (MGPG), that merges the concept of online cross-validation [148] with the PG framework. As a classical RL algorithm, the PG algorithm finds the optimal trajectory for the DBS by running gradient descent over the

policy space toward the maximal expected utility. However, this process is fitted only to certain environments, which can lead to a poor performance in dynamic networks. To address these challenges, we propose a meta-learning approach that tunes the hyper-parameters of the PG algorithm thus enabling the DBS adapt to the dynamic and unknown environments. Next, we first introduce the components of the proposed MGPG algorithm. Then, we explain how to use the MGPG algorithm to solve our problem.

6.3.1 MGPG Components

An MGPG algorithm consists of five components:

1. *Agent*: Our agent is the DBS whose goal is to design its trajectory.
2. *Actions*: The action of the agent is the cluster it targets at each step, the vector of actions taken by the DBS until step k is denoted $\mathbf{a}_k = [a_0, a_1, \dots, a_k]$ with $a_k \in \mathcal{L}$ with a_0 being the first cluster the DBS targets.
3. *States*: Each state considers both the agent's location, represented by the cluster it currently serves, and the DBS' energy level, represented by the remaining time for the DBS to return to the origin, and it is given by $\mathbf{s}_k = [a_{k-1}, \tau_k]$, where a_{k-1} indicates the DBS' location at step k . τ_k is the remaining time for the DBS to return to the origin at step k .
4. *Policy*: The policy of the DBS is defined as the probability of choosing a given action at a given state and is denoted by $\pi_{\theta}(a_k | \mathbf{s}_k)$. The policy is approximated by a deep neural network parametrized by θ . This neural network takes observation on the state of the agent and outputs the probabilities of taking each action at this state. The DBS's policies at all states is denoted in vector π_{θ} .
5. *Reward*: The reward of the DBS measures the benefit of each action. This reward is denoted as $r(a_k | \mathbf{s}_k) = \frac{\sum_{u \in \mathcal{U}} \mathbb{1}_{\{u \in \mathcal{U}_{\mathbf{a}_k}\}}}{\sum_{u \in \mathcal{U}} \mathbb{1}_{\{0 \leq t_u \leq T\}}}$, which is the service success rate the DBS achieves with action a_k at state $\mathbf{s}_k$, such that $\sum_{k=1}^K r(\mathbf{a}_k | \mathbf{s}_k) = \mu(\mathbf{a}_K)$. The average of the accumulated reward is the the expected utility or the objective function the DBS tries to optimize, that is,
$$\sum_{\mathbf{a}_K \in \mathcal{E}} \sum_{k=1}^K r(\mathbf{a}_k | \mathbf{s}_k) \prod_{k=1}^K \pi_{\theta}(a_k | \mathbf{s}_k) = \bar{\mu}(\pi_{\theta}).$$

6.3.2 Policy training procedure

In the proposed MGPG algorithm, the goal of the DBS is to find the optimal policy that maps its states to the actions leading to a maximum service success rate in various environments. The DBS will collect experiences at a realization of the user service request and then update its policy based on the utility value it obtains from these experiences. In particular, the DBS updates its policy

toward the maximized utility, i.e. in the direction to the gradient $\nabla_{\theta} \bar{\mu}(\pi_{\theta})$, which, based on the the policy gradient theorem, is approximated in the form of:

$$\mathbb{E}_{\pi} \left(\sum_{k=1}^K r(a_k | s_k) \left(\sum_{k=1}^K \nabla_{\theta} \log \pi_{\theta}(a_k | s_k) \right) \right). \quad (6.6)$$

The DBS, then, tunes its update direction after each policy update. Next, the procedure of the policy training and direction tuning is explained in detail.

The DBS is trained on its historical experience under a realization of the users' access requests that follow unknown distribution. The training process is given as follows:

1. Initialize a policy $\pi_{\theta^{(1)}}$ with random parameter values $\theta^{(1)}$ and $\eta^{(1)}$. Note that, here $\theta^{(1)}$ and $\eta^{(1)}$ are, respectively, the policy parameter and hyper-parameter at training episode 1.
2. Within each training episode i , the DBS carries out an experience by generating a sequence of actions that are randomly selected based on policy $\pi_{\theta^{(i)}}$. Each of such procedure generates an experience defined in a vector $e = [s_1, a_1, r(a_1 | s_1), \dots, s_K, a_K, r(a_K | s_K)]$ with K being the terminal step. Here, s_k and a_k are, respectively, the state and action at the k -th step of experience e .
3. The discounted future reward, denoted return G_k , at each step k of the DBS' experience is given by

$$G_k(\eta^{(i)}) = \sum_{\kappa=k}^K (\eta^{(i)})^{\kappa-k} r(a_{\kappa} | s_{\kappa}), \quad (6.7)$$

where hyper-parameter $\eta^{(i)}$ is the discount factor of the return function at training episode i .

4. The DBS then updates its policy with a reinterpreted objective defined as $J(\eta^{(i)}, \theta^{(i)}) = \sum_{k=1}^K G_k(\eta^{(i)}) \log \pi_{\theta^{(i)}}(a_k | s_k)$, one the sampled experience, in particular

$$\theta^{(i+1)} = \theta^{(i)} + \alpha \nabla_{\theta^{(i)}} J(\eta^{(i)}, \theta^{(i)}), \quad (6.8)$$

where α is the updating step size.

5. The DBS carries out an experience e' with the updated policy $\pi_{\theta^{(i+1)}}$.
6. The DBS measures current hyper-parameter $\eta^{(i)}$'s performance on directing its policy to the optimal utility with the updated measurement function $\tilde{J}(\tilde{\eta}, \theta^{(i+1)})$ with a fixed hyper-parameter $\tilde{\eta}$. The update function is given as

$$\eta^{(i+1)} = \eta^{(i)} - \beta \nabla_{\eta^{(i)}} \tilde{J}(\tilde{\eta}, \theta^{(i+1)}). \quad (6.9)$$

Steps 2 through 6 are repeated until training is complete and yields an optimal policy π_{θ^*} . Note that, instead of performing the expensive deviation calculations $\nabla_{\theta^{(i)}} J(\eta^{(i)}, \theta^{(i)})$, $\nabla_{\eta^{(i)}} \tilde{J}(\tilde{\eta}, \theta^{(i+1)})$, the DBS can approximate the update and tuning direction based on Proposition 1.

Proposition 3. At each iteration i of the policy update procedure, the DBS calculates the update direction using

$$\begin{aligned} & \nabla_{\theta^{(i)}} J \left(\eta^{(i)}, \theta^{(i)} \right) \\ &= \sum_{k=1}^K \sum_{\kappa=k}^K \left(\eta^{(i)} \right)^{\kappa-k} r(a_{\kappa} | s_{\kappa}) \frac{\nabla_{\theta^{(i)}} \pi_{\theta^{(i)}}(a_{\kappa} | s_{\kappa})}{\pi_{\theta^{(i)}}(a_{\kappa} | s_{\kappa})}. \end{aligned} \quad (6.10)$$

After every policy update, the DBS estimate the hyper-parameter tuning direction $\nabla_{\eta^{(i)}} \tilde{J} \left(\tilde{\eta}, \theta^{(i+1)} \right)$, with $\tilde{\eta}$ being 1, in the form of

$$\nabla_{\eta^{(i)}} \tilde{J} \left(\tilde{\eta}, \theta^{(i+1)} \right) = \alpha \sum_{k'=1}^K \sum_{k=1}^K A_{k'} B_k x_{k'} y_k. \quad (6.11)$$

where $A_{k'} = \sum_{\kappa=k'}^K r(a_{\kappa} | s_{\kappa})$, $x_{k'} = \frac{\nabla_{\theta^{(i+1)}} \pi_{\theta^{(i+1)}}(a_{k'} | s_{k'})}{\pi_{\theta^{(i+1)}}(a_{k'} | s_{k'})}$, $B_k = \sum_{\kappa=1}^K (\kappa - k) \left(\eta^{(i)} \right)^{\kappa-k-1} r(a_{\kappa} | s_{\kappa})$, and $y_k = \frac{\nabla_{\theta^{(i)}} \pi_{\theta^{(i)}}(a_k | s_k)}{\pi_{\theta^{(i)}}(a_k | s_k)}$.

Proof. See Appendix C.1. □

As such, using Proposition 1, the DBS can estimate the update and tuning direction simply based on the collected rewards and policy deviation $\frac{\nabla_{\theta} \pi_{\theta}(a_k | s_k)}{\pi_{\theta}(a_k | s_k)}$. The mini-batch training procedure of the MGPG based solution is shown in Algorithm 1. With this mini-batch training procedure, the DBS updates its policy over a whole experience to reduce the variance caused by the action sampling without the need to store a big dataset. At the beginning of the algorithm, the DBS randomly chooses its policy parameter $\theta^{(1)}$ and hyper-parameter $\eta^{(1)}$. The DBS, then, carries out an experience by selecting a trajectory with policy $\pi_{\theta^{(1)}}$. The DBS, then, flies along the selecting trajectory to serve the users. The service success rate resulted from the selected trajectory is recorded by the DBS. After returning to the origin, the DBS updates its policy parameter based on (6.8), with the service success rate it obtains from the selected trajectory. This update is then evaluated with (6.11) on an experience collected with the updated policy. The hyper-parameter is tuned with function (6.9) based on such evaluation. This policy training procedure is repeated by the DBS flying around the network until a convergence is reached. In essence, the DBS updates its policy episodically with a formerly generated experience at its origin, and tunes its hyper-parameter online right after each policy update. The complexity of the the MGPG based solution is $\mathcal{O}((v+1)nC)$, where v is the iteration at which the convergence is reached, n is the number of elements within policy parameter θ and C is the time complexity of calculating the gradient of each element (i.e. each element in θ). This complexity is considerably low, as it is close to the one of the vanilla policy gradient algorithm, $\mathcal{O}(vnC)$. Other advanced PG solutions such as advantage actor-critic (A2C), deep deterministic policy gradient (DDPG), however, have a huge time complexity, since each neural network they adopt have a time complexity of $\mathcal{O}(vn^5)$. In the studied drone-aided network, the lightweight MGPG algorithm is more desirable and practical, as the computation capacity of a drone is limited.

Algorithm 8 Proposed MGPG algorithm for trajectory design.**Require:** The user locations, time constraints.**Ensure:** Initialize parameter $\theta^{(1)}$ and $\eta^{(1)}$.

- 1: **for** Policy training epoch $i = 1 : I$ **do**
- 2: Carries out an experience e .
- 3: Update policy parameter based on (6.10).
- 4: Carries out an experience e' with updated policy.
- 5: Update hyper-parameter based on (6.11).
- 6: **end for**

Table 6.1: Simulation Parameters [3]

Parameter	Value	Parameter	Value
P_u	20 dBm	α	2
σ^2	-104 dB	B	1 MHz
V	30 m/s	m	3

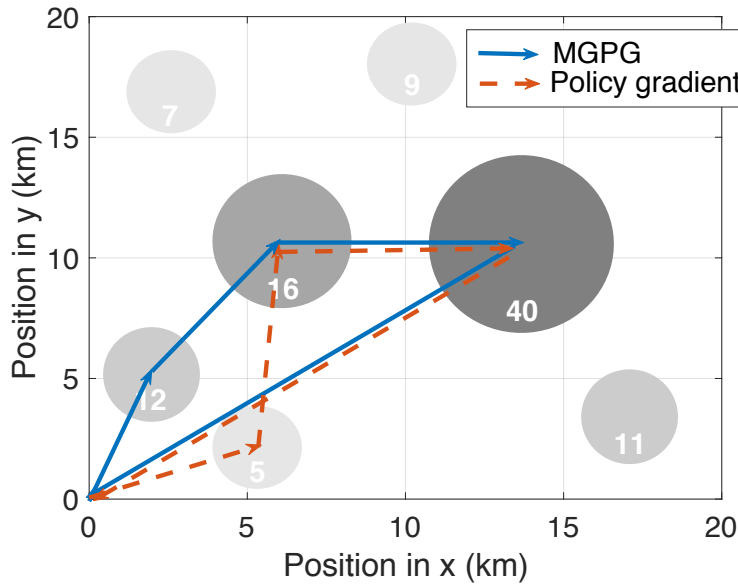


Figure 6.2: Snapshot of trajectories resulting from all considered algorithms in a sample unseen environment.

6.4 Simulation Results and Analysis

For our simulations, we consider a scenario with one DBS serving $U = 100$ mobile users. Other parameters used in the simulations are listed in Table 8. The proposed MGPG algorithm results are compared to the vanilla policy gradient algorithm [137], called policy gradient algorithm hereinafter. All statistical results are averaged over a large number of independent runs.

Fig. 6.2 shows a snapshot of the trajectories resulting from the proposed MGPG and the policy

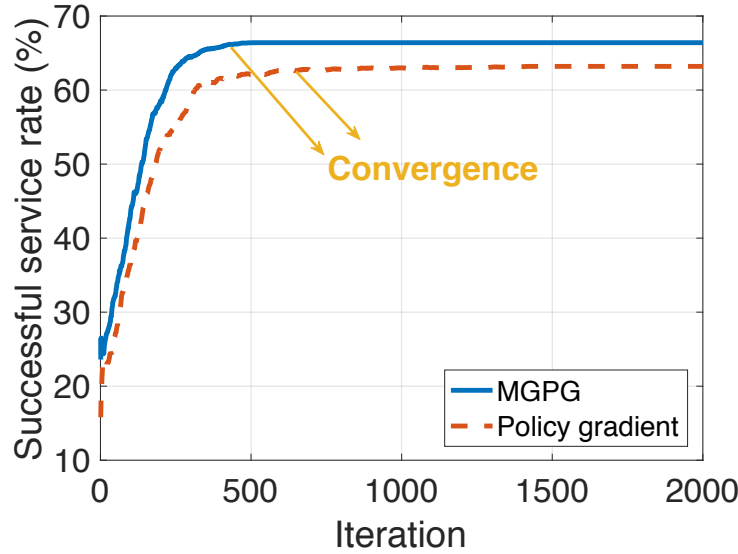


Figure 6.3: Convergence of all considered algorithms.

gradient algorithm in an illustrative unseen realization of user requests. The grey circles are the user clusters labeled by the number of active users. In this figure, we can see that, facing an unseen task with its manually pre-defined hyper-parameters, the policy gradient algorithm leads the DBS to only a suboptimal solution. The MGPG algorithm, however, effectively finds the optimal trajectory for the DBS by tuning the hyper-parameters in this unseen environment.

Fig. 6.3 shows the convergence of the proposed MGPG algorithm. From Fig. 6.3, we observe that the MGPG algorithm requires approximately 450 iterations to reach convergence, which is 25% less than the number of iterations required for convergence of the policy gradient algorithm. This is because the update direction of MGPG keeps being tuned in the policy update process, and, thus, it has a smaller angle towards the optimal result. than the sub-gradient algorithm. Fig. 6.3 also shows that the proposed MGPG algorithm achieves a service success rate of 68% at its convergence, which is about 10% higher the one reached by the policy gradient baseline. This stems from the fact that the proposed MGPG algorithm finds the hyper-parameter results in the best performance of the RL algorithm, as the MGPG algorithm keeps tuning the hyper-parameters in policy update process.

Fig. 6.4 shows the CDF of the service success rate at the convergence of all independent runs. For example, from Fig. 6.4, we can see that the proposed algorithm converges to a service success rate lower than 50% within only about 16% of the independent runs. In Fig. 6.4, we can see that the proposed algorithm achieves about 27% gain in the frequency of convergence at service success rate over 50%, compared to the policy gradient algorithm. This means that the proposed algorithm is more likely to converge to a higher service success rate, that is, it reduces the variance of algorithm performance. The main reason is that the proposed algorithm tunes the hyper-parameter toward a return G_t resulting in the best performance of the DBS.

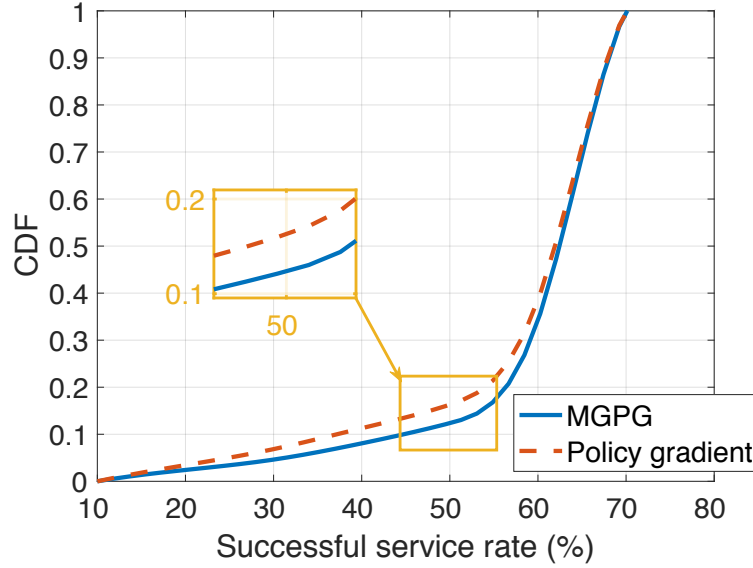


Figure 6.4: CDF in terms of service success rate.

6.5 Conclusion

In this chapter, we have studied a challenging trajectory design problem for UAVs in dynamic and unknown wireless network environments. We formulated an optimization problem that seeks to maximize the DBS service success rate. To solve this problem, we have developed a novel meta-reinforcement learning based algorithm that enables the DBS to adapt to unseen environments quickly. Simulation results have shown that, by tuning the hyper-parameters in RL, the proposed MGPG solution can yield a 25% improvement in the convergence speed, and an about 10% improvement in the final service success rate, compared to the vanilla policy gradient algorithm. Future work can consider various extensions such as the cases with multiple DBSs and other dynamic factors.

Chapter 7

Psychological Game Theory for Secured Internet of Battlefield Things Systems

Besides the challenges from the communication perspective, the security optimization of the ISDTC system must also be considered when implementing ISDTC system for communication application. In particular, the security of the the battlefield ISDTC system, i.e. the IoBT system, must be carefully studied, as the IoBT system is much more vulnerable to cyber and physical attacker than other ISDTC systems. In particular, with in an IoBT system, the connectivity between the wearables carried by the soldiers and other IoBT devices, such as multipurpose sensors, autonomous vehicles, and drones, plays a significant role in the mission-critical battlefield operations. Yet the connectivity between these devices is highly vulnerable to cyber attacks, given the the adversarial nature of the battlefield coupled with the limitations of the IoBT devices' security mechanisms. Moreover, as discussed in Chapter 1, in the adversarial battlefields, the psychology of the soldiers and attackers could significantly influence their behavior, and, subsequently influence the security of the IoBT network. Thus, in this chapter, we investigate the optimal anti-jamming solutions within human intervened IoBT system. Within this system, the defender will selectively choose IoBT device to connect for seamless, low latency data service. A game theory based framework is developed to analyze and predict the psychological attacking operations within the system. Then, a Bayesian learning based solution is proposed to learn the optimal connecting solutions for the defender.

7.1 Background, Related Works, and Contributions

The existing literature has studied a number of problems related to the security of the IoBT [58–60, 157]. In [58], the communications and information management challenges of the IoBT are investigated. The work in [59] integrates IoT and network centric warfare for the enhancement of the IoBT integrity. The authors in [60] use a feedback Stackelberg solution to dynamically

optimize the connectivity of an adversarial IoBT network. The work in [157] develops a mean-field game approach to analyze the spread of misinformation in an adversarial IoBT. Despite the promising results, these existing works [58–60, 157] mostly rely on static constructs and do not consider the influence of the human players' psychology and potential bounded rationality when making decisions or choosing strategies within an IoBT setting. Indeed, the behavioral aspect of human decision making processes, leading agents to deviate from the fully rational objective behavior in an IoBT, has a direct impact on the security of the IoBT network. Hence, this aspect must be accounted for and thoroughly studied within the context of studying and assessing the security of the IoBT.

In this chapter, we develop a novel framework to dynamically optimize the connectivity between a soldier and the IoBT network, considering the soldier's and attacker's psychological behaviors in the battlefield. We consider a battlefield in which a soldier must accomplish a time-critical mission that requires traversing the battlefield while maintaining connectivity with the IoBT network. Meanwhile, the attacker in the battlefield is interested in compromising the soldier's IoBT connectivity, by selectively jamming the IoBT network at each time instant in the battlefield. The soldier, acting as a defender, will selectively connect to certain IoBT devices at each time instant along its mission path, so as to evade the attack. We formulate this IoBT security problem as a dynamic game, in which the soldier attempts to predict and evade the attacker's attack at each time instant in the battlefield to minimize its cumulative expected retransmission delay, while the attacker aims at optimally targeting the IoBT devices to maximize the soldier's retransmission delay while accounting for its limited cumulative jamming power. Both the soldier's time limitation and the attacker's power limitation are considered in the formulated game. In this regard, we prove the uniqueness of the Nash equilibrium (NE) of this game, under a set of defined conditions, and we study the resulting NE strategies, which allows analysis of the optimal decision making processes of the soldier and attacker based on their built set of beliefs over the strategy on their opponent's strategies. We perform fundamental analysis on the soldier's and attacker's psychology in the battlefield using the framework of psychological game theory [104]. In the formulated psychological game, the psychology of the players (i.e. the soldier and the attacker) is modeled as their intention to *frustrate* each other. The frustration of the players is quantified as the gap, if positive, between their expected payoff and actual payoff. A psychological equilibrium (PE) is used to solve the psychological IoBT game. In this regard, we prove the uniqueness of the PE for our proposed psychological game, under the same set of conditions at which the NE is unique. In addition, our analytical results show that, in an attempt to frustrate the soldier, at the PE, the attacker is more prone to attack the IoBT device with the best channel conditions. The results also show that, with the error-free beliefs on the attacker's psychological strategies and beliefs, the soldier can obtain a notable improvement on its expected material payoff at a PE. Meanwhile, using Bayesian updating, the soldier and the attacker can achieve ϵ -like beliefs, such that an ϵ -like self-confirming psychological equilibrium of the formulated psychological game can be reached. Simulation results also validate that, the non-error-free beliefs achieved by the proposed algorithm can help obtain better connectivity at the soldier.

The rest of this chapter is organized as follows. The system model and problem formulation are

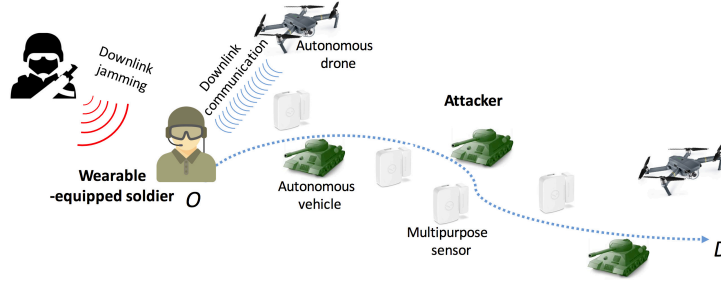


Figure 7.1: Soldier battlefield security graph.

described in Section 7.2. The psychological analysis of the soldier and attacker is represented in Section 7.3. The Bayesian updating-based solution of the psychological IoBT game is proposed in Section 7.4. In Section 7.5, simulation results are presented and analyzed. Finally, conclusions are drawn in Section 7.6.

7.2 System Model and Problem Formulation

Consider a battlefield in which a soldier seeks to move from an origin O to a destination D along a predefined path, using a minimum amount of time as shown in Fig. 1. At the same time, this soldier tries to communicate with a total of X IoBT devices in a set $\mathcal{X}$ that is uniformly deployed along this path, to get access to situational awareness within the battlefield and to receive instructions from the battlefield commander. The soldier can only associate with one IoBT device at each location. The soldier should communicate with $J < X$ IoBT devices along the path, so as to maintain its total downlink transmission delay lower than Δ , while getting access to the required information on time. Meanwhile, in this battlefield, an attacker seeks to disrupt the connectivity between the soldier and the IoBT devices by jamming the communication links. Given the limitation on its total power E , the attacker can only compromise (i.e. jam) the IoBT network at most J' times along the path. At each step in this battlefield, the soldier and attacker will sequentially choose strategies to realize their objective, based on their perfect observation on what happened in the battlefield. Here, the soldier's objective is minimizing its communication delay, the attacker's objective is maximizing the soldier's communication delay, while minimizing its total power consumption.

7.2.1 Soldier's communication delay

We assume that the soldier (attacker) chooses to connect with (jam) the IoBT network at each step, sequentially, until the soldier arrives at D .

The soldier communicates with each IoBT device $x \in \mathcal{X}$ over a downlink channel c_x . The SINR γ_x of the downlink channel between the soldier and IoBT device x is given by:

$$\gamma_x = \begin{cases} \frac{P_S l_x^s}{P_A l_x^a + \sigma^2}, & \text{if channel } c_x \text{ is jammed,} \\ \frac{P_S l_x^s}{\sigma^2}, & \text{otherwise,} \end{cases} \quad (7.1)$$

where P_S and P_A are, respectively, the transmit powers of soldier and the attacker. $l_x^s = g_s d_s^{-\lambda}$ is the path loss between the soldier and IoBT device x , with g_s being the Rayleigh fading parameter, d_s being the distance between the soldier and IoBT device x , and λ the path loss exponent. $l_x^a = g_a d_a^{-\lambda}$ is the path loss between the attacker and the IoBT device x , with g_a being the Rayleigh fading parameter, and d_a being the distance between the attacker and the IoBT device x . σ^2 is the power of the Gaussian noise. At each step x , based on the probability distribution of the Rayleigh fading parameter g_s and g_a , the probability that the soldier's received SINR, γ_x , is higher than a threshold, $\hat{\gamma}$, in one time slot is given by:

$$q_x = \begin{cases} \int_V^\infty f(g_s) dg_s, & \text{if channel } c_x \text{ is not jammed,} \\ \int_0^\infty \int_0^{\frac{g_s - V}{W}} f(g_a) f(g_s) dg_a dg_s, & \text{otherwise,} \end{cases} \quad (7.2)$$

where $V = \frac{\hat{\gamma} \sigma^2}{d_s^{-\lambda} P_S}$, $W = \frac{d_a^{-\lambda} P_A \hat{\gamma}}{d_s^{-\lambda} P_S}$. Here, $f(g_s)$ and $f(g_a)$ are the probability density functions of the Rayleigh fading parameters g_s and g_a , respectively. In the studied battlefield, the soldier attempts to maintain a probability of achieving an SINR exceeding $\hat{\gamma}$, q_x , that is higher than a threshold $\hat{q}$. Hence, the soldier will request k retransmissions of the downlink data from IoBT device x . However, the soldier will perform $k < \hat{k}$ retransmissions, in the case that the channel is occasionally experiencing a small scale fading. Thus, k is given by:

$$k = \begin{cases} \left\lceil \frac{\log(1-\hat{q})}{\log(1-q_x)} \right\rceil, & \text{if } k < \hat{k}, \\ \hat{k}, & \text{if } k \geq \hat{k}, \end{cases} \quad (7.3)$$

under the goal of maintaining $(1 - q_x)^k < 1 - \hat{q}$. Thus, the retransmission delay τ of the soldier at each step is given by $k t_x$. Here, $t_x = \frac{S}{I_x \log(1 + \gamma_x)}$ is the average unit transmission delay, which is the average duration of a successful packet transmission at the physical medium of one resource block, at step x [158]. S is the size of one resource block, I_x is the bandwidth of channel c_x .

7.2.2 Strategies of the players

In the studied battlefield, the objective of the soldier is to effectively maintain a low transmission delay. Thus, the soldier will attempt to communicate with the IoBT devices that will not be attacked. $\mathcal{A} = \{a_1, a_2\}$ represents the soldier's action space at each step $x \in \mathcal{X}$. Here, at each step x , $a_1 \in \mathcal{A}$ indicates that the soldier builds a communication link with IoBT device x , whereas $a_2 \in \mathcal{A}$ indicates that the soldier does not communicate with IoBT device x .

On the other hand, the objective of the attacker is to increase the retransmission delay of the soldier. As such, under a constraint on its total power consumption, the attacker will find the best time instant to launch an attack on the IoBT network, so as to decrease the SINR of the communication

channel between the soldier and IoBT network. The attacker's set of the possible actions at each step x can be represented by $\mathcal{B} = \{b_1, b_2\}$. Here, action $b_1 \in \mathcal{B}$ indicates that the attacker chooses to compromise the IoBT network, while action $b_2 \in \mathcal{B}$ indicates that the attacker does not jam the IoBT network. Note that the jamming power, P_K , that will be used by the attacker is assumed to be constant.

In addition, we use h^x to represent the sequence of actions that have been taken by each of the players before reaching step x . We refer to h^x as the history at step x . In this respect, the set of all possible histories h^x at step x is denoted by $\mathcal{H}^x$. In addition, we let $\hat{h}^x$ denote the sequence of actions that have been taken by each player up to step x , including the action pair taken at step x . After observing history $h^x \in \mathcal{H}^x$ at step x , the soldier and attacker will find the optimal strategies at the current step to realize their individual objectives. The set of soldier's feasible actions at history h^x is, then, represented by $\mathcal{A}_{h^x}$, while the set of the attacker's feasible actions at history h^x is represented by $\mathcal{B}_{h^x}$. The actions that are chosen by the soldier and the attacker at history h^x are represented, respectively, by $a(h^x)$ and $b(h^x)$. In addition, the set of possible terminal histories $\hat{h}^X$, at which point the soldier reaches D , is represented by $\mathcal{L}$, where $\mathcal{L} = \{\mathcal{H}^X, (\mathcal{A}_{h^x}, \mathcal{B}_{h^x})\}$.

In an adversarial IoBT environment, the soldier will randomize its action selection at each history such as to make it more complex for the attacker to guess the IoBT device to which the soldier aims to connect. The soldier will, hence, choose a probability distribution $\alpha_{h^x} = [\alpha_1, \alpha_2]$ over its feasible action set $\mathcal{A}_{h^x}$ at history h^x . In this regard, α_i denotes the probability of choosing action $a_i \in \mathcal{A}_{h^x}$ at history h^x , where $i \in \{1, 2\}$. This probability distribution α_{h^x} denotes the soldier's *mixed strategy* at history h^x . A possible strategy for the soldier in the battlefield can, then, be represented by a set $\alpha = \{\alpha_{h^x} | h^x \in \mathcal{H}^x, x \in \mathcal{X}\}$. The set of all feasible strategies of the soldier is denoted by $\mathcal{C}$.

A similar randomization logic is used by the attacker. The attacker seeks to choose a probability distribution $\beta_{h^x} = [\beta_1, \beta_2]$ over its feasible action set $\mathcal{B}_{h^x}$ at each history h^x , so as to maximize the soldier's transmission delay while keeping its consumed jamming power at a minimum. In this respect, β_i corresponds to the probability of choosing action $b_i \in \mathcal{B}_{h^x}$ at history h^x , where $i \in \{1, 2\}$. This probability distribution β_{h^x} is the attacker's *mixed strategy* at history h^x . A possible strategy for the attacker can, hence, be denoted by a set $\beta = \{\beta_{h^x} | h^x \in \mathcal{H}^x, x \in \mathcal{X}\}$. The set of all possible strategies of the attacker is denoted by $\mathcal{D}$.

7.2.3 Material payoff

We define the soldier's *material payoff* as the normalized gap between the sum of the soldier's actual communication delay and the soldier's maximum tolerable communication delay. Meanwhile, we define the attacker's *material payoff* as the weighted sum of the soldier's time delay and the attacker's power consumption.

Note that, in (7.2), q_x depends on both the soldier and attacker's actions $a(h^x)$ and $b(h^x)$ in the

form:

$$q_x(a(h^x) = a_1, b(h^x)) = (1 - \mathbb{1}_{b(h^x)=b_1}) \int_V^\infty f(g_x) dg_x \\ + \mathbb{1}_{b(h^x)=b_1} \int_0^\infty \int_0^{\frac{g_x - V}{W}} f(g_a) f(g_x) dg_a dg_x, \quad (7.4)$$

where $\mathbb{1}_{b(h^x)=b_1}$ is an indicator function that only equals to 1 when the current action of the attacker is b_1 . Hence, the soldier's time delay, when attempting to communicate with $x \in \mathcal{X}$ at history h^x , is given by:

$$\tau(a(h^x) = a_1, b(h^x)) \\ = \begin{cases} \left\lceil \frac{\log(1-\hat{q})}{\log(1-q_x(a(h^x), b(h^x)))} \right\rceil t, & \text{if } k < \hat{k}, \\ \hat{k}t, & \text{otherwise.} \end{cases} \quad (7.5)$$

In case the soldier does not communicate with the IoBT device at history h^x , the soldier will naturally not incur any delay which leads to $\tau(a(h^x) = 2, b(h^x)) = 0$. As such, the soldier's retransmission delay is a function of the soldier's and attacker's actions. At the terminal history $\hat{h}^X$, the soldier's accumulated communication delay will be:

$$\tau(\hat{h}^X) = \sum_{v=1}^X \tau(a(h^v), b(h^v)), \quad (7.6)$$

where $a(h^v)$ and $b(h^v)$ represent, respectively, the soldier's and attacker's action at step v in $\hat{h}^X$.

Note that, under each terminal history $\hat{h}^X \in \mathcal{Z}$, $\sum_{v=1}^X \mathbb{1}_{a(h^v)=a_1} = J$ and $\sum_{v=1}^X P_K \mathbb{1}_{b(h^v)=b_1} \leq E$.

Here, we note that, even though not communicating with any device will lead to a minimum delay for the soldier, this is not a feasible strategy for the soldier, since by definition, the soldier has to communicate with $0 < J \leq X$ devices in the battlefield so as to acquire situational awareness. Based on its primary objective, the soldier will determine an optimal strategy α that minimize its expected total time delay. This, hence, requires maximizing the normalized gap between the cumulative retransmission delay and the maximum tolerable delay, which is defined as :

$$\pi(\alpha, \beta) = \frac{\Delta - \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\alpha, \beta}(\hat{h}^X) \tau(\hat{h}^X)}{\Delta}, \quad (7.7)$$

where $Q_{\alpha, \beta}(\hat{h}^X)$ is the probability of occurrence of terminal history $\hat{h}^X \in \mathcal{Z}$, and is induced by the soldier's and the attacker's mixed-strategies, α and β as follow:

$$Q_{\alpha, \beta}(\hat{h}^X) = \prod_{x=1}^X (\alpha_{h^x}(1) \mathbb{1}_{a(h^x)=a_1} + \alpha_{h^x}(2) \mathbb{1}_{a(h^x)=a_2}) \\ \times (\beta_{h^x}(1) \mathbb{1}_{b(h^x)=b_1} + \beta_{h^x}(2) \mathbb{1}_{b(h^x)=b_2}), \quad (7.8)$$

where history h^x is part of $\hat{h}^X$. In other words, h^x represents the sequence of actions in $\hat{h}^X$ taken before x . Hence, $\pi(\alpha, \beta)$ represents the soldier's expected utility (or, equivalently, expected material payoff) achieved under the strategy pair (α, β) .

Meanwhile, the material payoff of the attacker at the terminal history $\hat{h}^X$ is defined as

$$\pi'_0(\hat{h}^X) = \frac{\sum_{v=1}^X \tau(a(h^v), b(h^v))}{\theta_1 \Delta} + \theta_2 \frac{E - \sum_{v=1}^X P_K \mathbb{1}_{b(h^v)=b_1}}{E}, \quad (7.9)$$

where $\theta_1 \geq 0$ and $\theta_2 \geq 0$ represent, respectively, the weight of time delay and power consumption, with $\theta_1 + \theta_2 = 1$. Thus, in this battlefield, the attacker will select the optimal strategy β that maximizes the soldier's time delay¹, while minimizing its power consumption, which can be captured by maximizing the following expected utility (i.e. expected material payoff):

$$\pi'(\alpha, \beta) = \sum_{\hat{h}^X \in \mathcal{X}} Q_{\alpha, \beta}(\hat{h}^X) \pi'_0(\hat{h}^X). \quad (7.10)$$

In this battlefield, the attacker can track the soldier's location via GPS, and it can gather intelligence (i.e., knowledge) on the soldier's associated objective. However, it does not know the IoBT devices to which the soldier will connect. Meanwhile, the soldier knows that the attacker is present, but does not know which IoBT devices it will target. Then, to determine their optimal actions at each history, the soldier and attacker aim at forming an estimation of their opponent's actions (e.g. the attacker estimates the soldier's actions, and the soldier estimates the attacker's actions). This estimation is defined as the soldier and attacker's *first-order beliefs* on each other. Let $\delta_{h^x}^1 = [\delta_{h^x}^1(1), \delta_{h^x}^1(2)]$ be the soldier's vector of beliefs on the probability distribution of the attacker's actions b_1 and b_2 at history h^x , and let $\rho_{h^x}^1 = [\rho_{h^x}^1(1), \rho_{h^x}^1(2)]$ be the attacker's belief vector on the probability distribution of the soldier's actions a_1 and a_2 at history h^x , respectively. As such, we let $\delta_{\mathcal{H}^x}^1$ and $\rho_{\mathcal{H}^x}^1$, denote the set of first-order beliefs of, respectively, the soldier and attacker for each possible history at step x . Hereinafter, we use $\delta^1 = \{\delta_{\mathcal{H}^1}^1, \dots, \delta_{\mathcal{H}^X}^1\}$ to denote a set of soldier's first-order beliefs on the attacker, and $\rho^1 = \{\rho_{\mathcal{H}^1}^1, \dots, \rho_{\mathcal{H}^X}^1\}$ to denote a set of attacker's first-order beliefs on the soldier, at each possible history.

Based on belief δ^1 , the soldier's perceived (i.e. belief-based) expected material payoff will be given by:

$$\bar{\pi}(\alpha, \delta^1) = \frac{\Delta - \sum_{\hat{h}^X \in \mathcal{X}} Q_{\alpha, \delta^1}(\hat{h}^X) \tau(\hat{h}^X)}{\Delta}, \quad (7.11)$$

where $Q_{\alpha, \delta^1}(\hat{h}^X)$ is the belief-based probability of occurrence of the terminal history $\hat{h}^X \in \mathcal{X}$ induced by α and δ^1 :

$$\begin{aligned} Q_{\alpha, \delta^1}(\hat{h}^X) &= \prod_{x=1}^X (\alpha_{h^x}(1) \mathbb{1}_{a(h^x)=a_1} + \alpha_{h^x}(2) \mathbb{1}_{a(h^x)=a_2}) \\ &\quad \times (\delta_{h^x}^1(1) \mathbb{1}_{b(h^x)=b_1} + \delta_{h^x}^1(2) \mathbb{1}_{b(h^x)=b_2}). \end{aligned} \quad (7.12)$$

¹Here, we assume that Δ can be learnt by the attacker using its knowledge of the IoBT devices' quantity and channel condition, or through, for example, a prior reconnaissance phase about the soldier and its objectives.

Similarly, given its first-order belief ρ^1 , the attacker's perceived (i.e. belief-based) expected material payoff under strategy β will be:

$$\bar{\pi}'(\beta, \rho^1) = \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\rho^1, \beta}(\hat{h}^X) \pi'_0(\hat{h}^X), \quad (7.13)$$

where $Q_{\rho^1, \beta}(\hat{h}^X)$ is the belief-based probability of occurrence of the terminal history $\hat{h}^X \in \mathcal{Z}$ induced by ρ^1 and β :

$$\begin{aligned} Q_{\rho^1, \beta}(\hat{h}^X) &= \prod_{x=1}^X (\rho_{h^x}^1(1) \mathbb{1}_{a(h^x)=a_1} + \rho_{h^x}^1(2) \mathbb{1}_{a(h^x)=a_2}) \\ &\quad \times (\beta_{h^x}(1) \mathbb{1}_{b(h^x)=b_1} + \beta_{h^x}(2) \mathbb{1}_{b(h^x)=b_2}). \end{aligned} \quad (7.14)$$

7.2.4 Game formulation

In the studied IoBT scenario, the primary objective of the soldier is to find a strategy to effectively evade the jamming attack of the attacker, while the objective of the attacker is to find an attack strategy that effectively jams the soldier's communication with the IoBT devices. As such, we formulate a dynamic game $[\mathcal{P}, \mathcal{H}, \mathcal{Z}, \pi, \pi', \bar{\pi}, \bar{\pi}']$ to capture the dependence between the objectives and the actions of the soldier and the attacker. Here, $\mathcal{P}$ is the set of players which includes the soldier and attacker. $\mathcal{H}$ is the set of histories representing the sequence of actions that have been taken by each of the players before reaching a certain stage, and $\mathcal{Z}$ represents the set of terminal histories, at which point the soldier reaches its destination, D , and the game ends. π and π' are the expected utilities of the soldier and the attacker, respectively, defined in (7.7), and (7.10), while $\bar{\pi}$ and $\bar{\pi}'$ are their belief-dependent (i.e. perceived) expected utilities, defined in (7.11) and (7.13). In this formulated game, each of the soldier and the attacker aim at maximizing their (belief-based) expected utilities. When the beliefs of each player accurately predict the strategy of the opponent, and when each player chooses a strategy that maximizes its expected utility based on those beliefs, these strategies give rise to a *NE* for the proposed game, which is formally defined as follows:

Definition 4. A *NE* for the formulated dynamic game is defined as $(\alpha^*, \beta^*, \rho^{1*}, \delta^{1*})$, in which α^* , and β^* are rational, such that:

$$\alpha^* \in \arg \max_{\alpha \in \mathcal{C}} \bar{\pi}(\alpha, \delta^{1*}), \quad (7.15)$$

$$\beta^* \in \arg \max_{\beta \in \mathcal{D}} \bar{\pi}'(\beta, \rho^{1*}), \quad (7.16)$$

while beliefs ρ^{1*} and δ^{1*} are error-free such that for all $a_n \in \mathcal{A}_{h^x}$ at each history h^x in the game:

$$\rho_{h^x}^{1*}(n) = \alpha_{h^x}^*(n), \quad (7.17)$$

and for all $b_m \in \mathcal{B}_{h^x}$ at each history h^x in the game:

$$\delta_{h^x}^{1*}(m) = \beta_{h^x}^*(m). \quad (7.18)$$

Thus, at an NE of the proposed game, both the soldier and attacker correctly estimate their opponents' strategies (represented by an error-free set of beliefs over the opponent's strategy) and make rational determinations on their strategies based on their error-free beliefs, at every history of the game. As shown in Definition 1, the rational strategies of the players (i.e. the soldier and the attacker) are the strategies that maximize the players' error-free belief-based perceived expected payoff, $\bar{\pi}$ and $\bar{\pi}'$. At each history $h^x \in \mathcal{H}^x$, the players' error-free first-order beliefs (i.e. δ^{1*} and ρ^{1*}) on each of their opponents' feasible action (i.e. $b_m \in \mathcal{B}_{h^x}$, $a_n \in \mathcal{A}_{h^x}$) equals the probability that their opponents choose this action with their rational strategies (i.e. α^* and β^*). The players, including the soldier and the attacker, are considered to hold accurate (error-free) beliefs in the computation of their respective NE strategies. Hence, these error-free beliefs require that, at equilibrium, beliefs should accurately predict the opponent's strategy. However, in practical networks, the players' beliefs may not be fully accurate, when no effective prediction method is used. Hence, when solving (7.15) and (7.16), the resulting soldier and attacker strategies are rational (i.e. optimal), but are based on their respective beliefs. If these beliefs are not accurate (i.e. if (7.17) and (7.18) are not met), even through each of the players is still acting rationally, their strategies may deviate from the NE strategies.

Moreover, in practice, as emotional human players, the soldier and the attacker may also deviate from their NE strategies [159]. In this case, despite being theoretically valid, the error-free beliefs, δ^{1*} and ρ^{1*} defined in Definition 1 may not be consistent with the players' actual emotional strategies. As such, the rational strategies, α^* and β^* , that maximize the players' δ^{1*} -based and ρ^{1*} -based perceived expected payoffs, may not maximize the players real expected payoffs when the opponent deviates from its fully rational strategies, due to behavioral factors [104]. In addition, the soldier's and attacker's subjective emotions may also modify their objective functions to incorporate additional subjective goals. These psychological facets of the player's behavior in an IoBT network, are studied next using the framework of a dynamic psychological game [104].

7.3 Dynamic Psychological Game

The formulated dynamic game in Section 2.2 captures the primary objectives of the soldier and attacker and the interdependence between these objectives. In this respect, in this game, each player, by using a set of beliefs about the opponent's strategy, aims at computing its optimal strategy to maximize its respective expected utility. Hence, the beliefs are considered to be solely a means using which a player can estimate its opponent's strategy in order to choose its optimal strategy, but are not considered a part of the utility function of each player. However, given the psychological (i.e. human) nature of the players in our game, their expectations, beliefs, and emotions have a direct effect on the way they perceive the outcome of the game. Indeed, by not achieving their expected (or belief-based) payoff, the soldier or attacker will experience frustration or anger, which has a direct impact on the way they assess and perceive the outcome of the game. In addition, due to the adversarial nature of the relationship between the soldier and attacker, in addition to achieving their own objective by maximizing/minimizing the communication delay, each may also

strive to intentionally hurt the opponent, by aiming at frustrating the opponent or, more generally, causing a psychological (i.e. emotional) damage to this opponent. Hence, incorporating this psychological aspect in the formulation of the utility functions of each player enables a more general and representative game analysis that can realistically capture the psychological decision making processes and behavior of each of the soldier and attacker.

To this end, we next incorporate notions from psychological games [104] in our game formulation to capture and analyze this psychological aspect of the decision making processes of the attacker and soldier. As such, in our introduced psychological game, the players expectations and beliefs will now be directly incorporated in their utility functions. In addition, given their objective to frustrate and anger the opponent, each player aims at anticipating the payoff that the opponent expects. To this end, in addition to building beliefs over the opponent strategies, each player also aims at building a belief system over the opponent's beliefs. This would, hence, enable anticipating the expectations of the opponent and, as a result, maximize its frustration.

7.3.1 Psychology in the battlefield

In the aforementioned IoBT scenario, when one player (i.e. the soldier or the attacker) chooses its strategy such that its opponent receives a material payoff lower than expected, this player successfully frustrates its opponent. For example, if the soldier believes that the attacker did not launch an attack on the IoBT network at step x , it will communicate with this IoBT device x and expect to achieve a material payoff $\bar{\pi}$. If, in reality, the attacker attacked x , then the material payoff of the soldier will decrease to $\tilde{\pi}$. The gap between $\bar{\pi}$ and $\tilde{\pi}$ quantifies the soldier's frustration. Note that the soldier and attacker only feel frustrated when they get a lower material payoff, compared to their expected material payoff. Hence, in our proposed psychological game formulation, the soldier and attacker will intentionally attempt to frustrate each other while also seeking to achieve their own, individual objectives. Ultimately, the soldier and attacker's intention to frustrate each other, combined with their individual objectives (i.e. to minimize or maximize the soldier's communication delay), will determine the soldier and attacker's strategies in the battlefield.

To consider their opponents' frustration in their own payoffs, the soldier and attacker should estimate their opponent's expected payoffs. This estimation requires the soldier and the attacker to build beliefs about their opponent's first-order beliefs, i.e. to build *second-order beliefs*. The soldier's second-order belief on the attacker's first-order belief at history h^x , $\rho_{h^x}^1$, is denoted by a vector $\delta_{h^x}^2 = [\delta_{h^x}^2(1), \delta_{h^x}^2(2)]$. The attacker's second-order belief on the soldier's belief $\delta_{h^x}^1$ is denoted by $\rho_{h^x}^2 = [\rho_{h^x}^2(1), \rho_{h^x}^2(2)]$. Hereinafter, we use $\delta^2 = \{\delta_{\mathcal{H}^1}^2, \dots, \delta_{\mathcal{H}^X}^2\}$ to denote the set of soldier's second-order beliefs on the attacker, and $\rho^2 = \{\rho_{\mathcal{H}^1}^2, \dots, \rho_{\mathcal{H}^X}^2\}$ to denote the set of attacker's second-order beliefs on the soldier, for each possible history.

7.3.2 Soldier and attacker's frustration

We define the soldier's and attacker's frustration as the gap between their expected material payoffs, respectively defined in (7.11) and (7.13), and their actual material payoffs. This frustration, indeed, stems from the fact that the soldier (attacker) may choose an action $a_n \in \mathcal{A}_{h^x}$ ($b_m \in \mathcal{B}_{h^x}$) that may be different from what the attacker (soldier) has anticipated based on its belief $\rho_{h^x}^1$ ($\delta_{h^x}^1$). Thus, in the considered IoBT network, under terminal history $\hat{h}^X$, the soldier's frustration with strategy α and belief δ^1 will be given by (given that the soldier aims at maximizing $\bar{\pi}$ defined in (7.11)):

$$F(\alpha, \delta^1, \hat{h}^X) = [\bar{\pi}(\alpha, \delta^1) - \tilde{\pi}(\hat{h}^X)]^+, \quad (7.19)$$

where $[x]^+ = \max\{0, x\}$. $\tilde{\pi}(\hat{h}^X) = \frac{\Delta - \tau(\hat{h}^X)}{\Delta}$ is the soldier's actual payoff under terminal history $\hat{h}^X$. Note that the attacker has no knowledge of the soldier's first-order belief, δ^1 , and strategy, α . Hence, based on its sets of first-order and second-order beliefs, ρ^1 and ρ^2 , on the soldier's strategy α and first-order belief δ^1 , the attacker can form a belief-based perception of the soldier's frustration, denoted by $F_a(\rho^1, \rho^2, \hat{h}^X)$, when a terminal history $\hat{h}^X$ occurs, is expressed as follows:

$$F_a(\rho^1, \rho^2, \hat{h}^X) = [\bar{\pi}(\rho^1, \rho^2) - \tilde{\pi}(\hat{h}^X)]^+, \quad (7.20)$$

where $\bar{\pi}(\rho^1, \rho^2) = \frac{\Delta - \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\rho^1, \rho^2}(\hat{h}^X) \tau(\hat{h}^X)}{\Delta}$. In addition, $Q_{\rho^1, \rho^2}(\hat{h}^X)$ is the attacker's perceived belief-based probability of occurrence of the terminal histories $\hat{h}^X \in \mathcal{Z}$ induced by its first-order and second-order beliefs, ρ^1 and ρ^2 :

$$\begin{aligned} Q_{\rho^1, \rho^2}(\hat{h}^X) &= \prod_{x=1}^X (\rho_{h^x}^1(1) \mathbb{1}_{a(h^x)=a_1} + \rho_{h^x}^1(2) \mathbb{1}_{a(h^x)=a_2}) \\ &\quad \times (\rho_{h^x}^2(1) \mathbb{1}_{b(h^x)=b_1} + \rho_{h^x}^2(2) \mathbb{1}_{b(h^x)=b_2}). \end{aligned} \quad (7.21)$$

Thus, combining the attacker's primary objective of maximizing the soldier's communication delay at a minimum needed total jamming power with its intention to frustrate the soldier results in the following belief-based expected psychological payoff (i.e. belief-based expected psychological utility):

$$\begin{aligned} u'(\beta, \rho^1, \rho^2) &= \bar{\pi}'(\beta, \rho^1) + \omega_a \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\rho^1, \beta}(\hat{h}^X) F_a(\rho^1, \rho^2, \hat{h}^X), \end{aligned} \quad (7.22)$$

where $\omega_a \in [0, 1]$ is a parameters that represents the attacker's motivation and willingness to frustrate the soldier.

Similarly, under history $\hat{h}^X$, the frustration of the attacker with strategy β , under the first-order belief ρ^1 , will be:

$$F'(\beta, \rho^1, \hat{h}^X) = [\bar{\pi}'(\beta, \rho^1) - \pi'_0(\hat{h}^X)]^+, \quad (7.23)$$

where $\pi'_0(\hat{h}^X)$ is the attacker's actual payoff at terminal history, as defined in (9). Based on its first-order and second-order beliefs, δ^1 and δ^2 , the soldier can form a belief-based perception of qualify the attacker's frustration, denoted by $F_s(\delta^1, \delta^2, \hat{h}^X)$, when terminal history $\hat{h}^X$ is achieved, as follows:

$$F_s(\delta^1, \delta^2, \hat{h}^X) = \left[\bar{\pi}'(\delta^1, \delta^2) - \pi'_0(\hat{h}^X) \right]^+, \quad (7.24)$$

where $\pi'(\delta^1, \delta^2) = \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\delta^2, \delta^1}(\hat{h}^X) \pi'_0(\hat{h}^X)$. Here, $Q_{\delta^2, \delta^1}(\hat{h}^X)$ is the perceived belief-based probability of occurrence of terminal history $\hat{h}^X \in \mathcal{Z}$ based on the soldier's first-order and second-order beliefs, δ^1 and δ^2 , and is defined as:

$$Q_{\delta^2, \delta^1}(\hat{h}^X) = \prod_{x=1}^X (\delta_{h^x}^2(1) \mathbb{1}_{a(h^x)=a_1} + \delta_{h^x}^2(2) \mathbb{1}_{a(h^x)=a_2}) \\ \times (\delta_{h^x}^1(1) \mathbb{1}_{b(h^x)=b_1} + \delta_{h^x}^1(2) \mathbb{1}_{b(h^x)=b_2}). \quad (7.25)$$

Then, the soldier's goal to minimize its expected communication delay combined with its intention to frustrate the attacker can be captured by the following belief-based expected psychological payoff (i.e. belief-based expected psychological utility):

$$u(\alpha, \delta^1, \delta^2) \\ = \bar{\pi}(\alpha, \delta^1) + \omega_s \sum_{\hat{h}^X \in \mathcal{Z}} Q_{\alpha, \delta^1}(\hat{h}^X) F_s(\delta^1, \delta^2, \hat{h}^X), \quad (7.26)$$

where $\omega_s \in [0, 1]$ is a parameter that represents the soldier's motivation to frustrate the attacker.

7.3.3 Dynamic Psychological game

To capture the decision making processes of the soldier and attacker, we introduce a *dynamic psychological game* $[\mathcal{P}, \mathcal{H}, \mathcal{Z}, u, u']$, where, similarly to the dynamic game defined in Section 7.2.4, $\mathcal{P}$ is the set of players including the soldier and attacker. $\mathcal{H}$ is the set of histories, and $\mathcal{Z}$ is the set of terminal histories. In addition, u and u' represent the soldier's and the attacker's psychological expected utility defined in (7.22) and (7.26), respectively. In this psychological game, the soldier and the attacker aim at maximizing their belief-based psychological expected utilities. In this regard, when the first-order and second-order beliefs of each player correctly predict the strategy and the first-order belief of the opponent, and when each player chooses a strategy that maximizes its belief-based psychological expected utility based on those correct beliefs, these strategies give rise to a *PE* [104]. In this respect, the PE of our proposed psychological game is formally defined as follows:

Definition 5. The *psychological equilibrium* of the formulated psychological game is defined as $(\alpha^*, \beta^*, \delta^{1*}, \delta^{2*}, \rho^{1*}, \rho^{2*})$, in which α^* , and β^* are rational, such that:

$$\alpha^* \in \arg \max_{\alpha \in \mathcal{C}} u(\alpha, \delta^{1*}, \delta^{2*}), \quad (7.27)$$

$$\beta^* \in \arg \max_{\beta \in \mathcal{D}} u'(\beta, \rho^{1*}, \rho^{2*}), \quad (7.28)$$

while the first-order and second-order beliefs, δ^{1*} , δ^{2*} , ρ^{1*} , and ρ^{2*} , are error-free such that for all $a_n \in \mathcal{A}_{h^x}$ at each history h^x :

$$\delta_{h^x}^{2*}(n) = \rho_{h^x}^{1*}(n) = \alpha_{h^x}^*(n), \quad (7.29)$$

and for all $b_m \in \mathcal{B}_{h^x}$ at each history h^x :

$$\rho_{h^x}^{2*}(m) = \delta_{h^x}^{1*}(m) = \beta_{h^x}^*(m). \quad (7.30)$$

The principal difference between the PE and the NE (which is an underlying difference between the proposed standard dynamic game and the proposed psychological game) is that the utility function of each player is not only dependent on the strategy or action chosen by the opponent, but also on the opponent's beliefs. In this regard, the payoff of each player in the psychological game does not only depend on what the opponent does, but also on what the opponent thinks. This enlarges the domain of analysis of the game to incorporate psychological aspects of the players' decision making processes, which are not typically present in a traditional dynamic game formulation. Hence, even though the definition of the PE still requires that the first-order and second-order beliefs of each player are error-free, since these beliefs are incorporated in the payoffs of each player, they will have a direct effect on their rationally chosen (i.e. PE) strategies. In essence, at a PE, the players' intention to frustrate one another is captured, as the soldier and the attacker make rational determination on their strategies to maximize both their belief-based expected material payoff and their opponents' frustration, based on their error-free first-order and second-order beliefs. Based on [104], there always exists at least one such PE in the formulated psychological game. In particular, under the assumptions that i) evaded attacks at history h^x yield higher expected material payoffs for the soldier and lower expected material payoffs for the attacker, at the current and future histories, and ii) a successful (unjammed) communication at history h^x yields a higher expected material payoff for the soldier and a lower expected material payoff for the attacker, at the current and future histories, we can derive Theorem 1 and Theorem 2:

Theorem 1. The NE and the PE of, respectively, the conventional dynamic game and the psychological game are unique.

Proof. See Appendix D.1 . □

The incorporation of the opponent's beliefs in the objective function of each player, and the possible willingness of each player to not only meet its own objective but to frustrate the opponent, introduce significant modifications to the equilibrium strategies of each player as shown in Theorem 2.

Theorem 2. In the psychological game, at the PE, the attacker is more likely to target the IoBT device having the best channel conditions as compared to the NE of the traditional dynamic game.

Proof. See Appendix D.2 . □

Hence, Theorem 2 shows the effect that the incorporation of beliefs in each player's objective function (as in the proposed psychological game) can have on their equilibrium strategies. In particular, even when the beliefs are error-free, the fact that the beliefs have a direct effect on how the outcome of the game is assessed by each player has a direct impact on the chosen equilibrium strategies. Indeed, Theorem 2 shows that, for the attacker, incorporating a belief over the soldier's strategy and a belief over the soldier's belief in its objective function allows the attacker to modify its equilibrium strategy in order to maximize the soldier's frustration.

In summary, the formulated psychological game enables analysis of the soldier's and attacker's psychological intention to frustrate each other and allows studying the effect of such a psychological decision making aspect on the chosen strategies of each player. As can be seen from Definition 2 and conditions (7.29) and (7.30), holding correct beliefs is necessary to reach a PE of the game. This highlights the importance of the beliefs and the effect that they have on the chosen soldier and attacker strategies. In practical applications, in which each player may not have the ability to analytically characterize a set of error-free beliefs, a learning algorithm could be applied to numerically synthesize these beliefs [160]. To this end, a Bayesian updating based algorithm is proposed next, which enables a numerical computation of first-order and second-order beliefs of each player, which as a result, allows computation of equilibrium strategies of the psychological game.

7.4 Learning to Be Rational in the Psychological Game

In the studied IoBT scenario, the soldier and attacker form first-order and second-order beliefs, which are used to compute their optimal (i.e. rational) strategy. As such, as shown in Definition 2, the PE strategies require the players to form correct (i.e. error-free) beliefs. Forming such beliefs analytically is a typically complex task especially when the size of the battlefield grows, which leads to having a significantly large number of possible histories in the game. Hence, rather than relying on complex analytical derivations, the soldier and attacker can rely on numerical techniques and observations to form such beliefs, and as a result, choose their strategies. A powerful tool which can be used to form the players' beliefs in our proposed psychological game is *Bayesian updating* [161], which enables the use of observations to form consistent beliefs. Hence, next, we develop a Bayesian updating based approach to solve the proposed psychological game, by first predicting the players' future strategies and beliefs.

By using Bayesian updating, the attacker and the soldier find, at each history h^x in the game, find the posterior probabilities as follows:

$$\Pr(a_n | h^x) = \frac{\Pr(h^x | a_n) \Pr(a_n)}{\Pr(h^x)}, \quad (7.31)$$

$$\Pr(b_m | h^x) = \frac{\Pr(h^x | b_m) \Pr(b_m)}{\Pr(h^x)}, \quad (7.32)$$

where $\Pr(a_n | h^x)$ represents the probability of the soldier choosing action a_n at history h^x , while $\Pr(b_m | h^x)$ is the probability of the attacker choosing action b_m at history h^x . $\Pr(h^x | a_n)$ repre-

sents the probability that the current history is h^x when the soldier chooses action a_n at step x . $\Pr(h^x | b_m)$ represents the probability that the current history is h^x when the soldier chooses action b_m at step x . $\Pr(a_n)$ and $\Pr(b_m)$, respectively, represent the probabilities that action a_n or b_m is chosen at step x . $\Pr(h^x)$ represents the probability that history h^x is reached at step x .

Under Bayesian updating, the soldier and the attacker build their first-order beliefs on their opponent's strategy, under each history h^x , according to (7.31) and (7.32). Meanwhile, based on Definition 2, the soldier and attacker's second-order beliefs are consistent with their own strategies. Based on their beliefs, the soldier and attacker determine their optimal strategies that maximize their belief-dependent utilities in (7.22) and (7.26). The specific process of our Bayesian updating based solution is represented in Algorithm 1. Note that, all of the aforementioned probabilities will be updated through the repetition of the game². At the beginning of algorithm, the soldier and attacker assume $\Pr(a_n) = \Pr(b_m) = \frac{1}{2}$ at each step x , as there is no reason for them to assume that their opponents has any preference on the choice of their actions. Similarly, $\Pr(h^x | a_n) = \frac{1}{V_{a_n, h^x}}$, $\Pr(h^x | b_m) = \frac{1}{V_{b_m, h^x}}$ are assumed at the beginning of the algorithm, with V_{a_n, h^x} , V_{b_m, h^x} representing the total number of times that when a_n or b_m are respectively chosen, they are chosen from history h^x . We also consider $\Pr(h^x) = 1$ when history h^x is reached at the first iteration of the game.

By extending the results in [162] to our formulated psychological game, when the soldier and the attacker maximize their payoffs based on their beliefs learnt from Bayesian updating, their beliefs will always converge to a value that ϵ -likes their opponents' strategies and beliefs. Here, a first-order belief at history h^x , $\delta_{h^x}^1 = [\delta_{h^x}^1(1), \delta_{h^x}^1(2)]$, is said to ϵ -like a player's strategy (i.e. $\beta_{h^x} = [\beta_1, \beta_2]$) when there exists an $\epsilon > 0$, such that (i) $\sum_{n=1}^2 \delta_{h^x}^1(n)$ and $\sum_{n=1}^2 \beta_n$ are greater than $1 - \epsilon$; (ii) $(1 - \epsilon) \beta_n \leq \delta_{h^x}^1(n) \leq (1 + \epsilon) \beta_n$, for $n \in \{1, 2\}$. During the Bayesian updating process, the soldier and the attacker predict their opponent's strategies based on the sequence of actions that the soldier and attacker had taken during the updating process. In this regard, each player, including the soldier and the attacker, optimizes its utilities based on the beliefs learned from the actual strategies played in the game. Hence, learning a belief system over the strategy the opponent takes at a certain history requires this history to be reached during the learning process. Hence, if the sequence of actions taken by the soldier and attacker do not lead to a certain history to be reached, the player's cannot use previous observations to build a belief system over that particular history. As such, based on the Bayesian updating and optimal strategy selection, the soldier and the attacker are guaranteed to reach an ϵ -like psychological self-confirming equilibrium (PSCE) defined as follow [163].

Definition 6. The ϵ -like PSCE of the formulated dynamic psychological game is defined as a vector $(\alpha^*, \beta^*, \delta^{1*}, \delta^{2*}, \rho^{1*}, \rho^{2*})$, where α^* , and β^* are rational; while ρ^{1*} , δ^{1*} , ρ^{2*} , and δ^{2*} respectively ϵ -like α^* , β^* , δ^{1*} , ρ^{1*} , for each history h^x such that $\Pr(\alpha^*, \beta^* | h^x) > 0$.

²In practical IoBT scenarios, there will be more than one soldier working on a same mission. The soldier and the attacker's beliefs, which are consistent with the posterior probabilities defined in (31) and (32), are updated through different soldiers' accomplishment of the same mission.

Algorithm 9 Bayesian updating solution for the dynamic psychological game

Require: The set of IoBT devices $\mathcal{X}$ in the battlefield, the number of required communication links between the soldier and IoBT devices J , and the power limitation of the attacker E .

Ensure: Initialize belief of the soldier and attacker.

```

1: for IoBT device 1 to  $X$  do
2:   Update  $\Pr(h^x)$ .
3:   Calculate  $\Pr(a_n|h^x)$ ,  $\Pr(b_m|h^x)$ , with  $\delta_{h^x}^{2*}(n) = \rho_{h^x}^{1*}(n) = \Pr(a_n|h^x)$ ,  $\rho_{h^x}^{2*}(m) = \delta_{h^x}^{1*}(m) = \Pr(b_m|h^x)$ .
4:   for the attacker do
5:     for all action  $b_m \in \mathcal{B}_{h^x}$  do
6:       Estimate the soldier expected material payoff based on  $\rho_{h^x}^{2*}$ .
7:     end for
8:     Find  $b^*(h^x) = \arg \max_{b_m \in \mathcal{B}_{h^x}} u'$ .
9:     Update  $\Pr(b_m)$ ,  $\Pr(h^x|b_m)$ , for all  $a_n \in \mathcal{A}_{h^x}$ .
10:  end for
11:  for the soldier do
12:    for all action  $a_n \in \mathcal{A}_{h^x}$  do
13:      Estimate the attacker's expected material payoff based on  $\delta_{h^x}^{2*}$ .
14:    end for
15:    Find  $a^*(h^x) = \arg \max_{a_n \in \mathcal{A}_{h^x}} u$ .
16:    Update  $\Pr(a_n)$ ,  $\Pr(h^x|a_n)$ , for all  $b_m \in \mathcal{B}_{h^x}$ .
17:  end for
18: end for

```

As such, in an ϵ -like PSCE, the soldier and the attacker's beliefs, which are updated based on previous actions that have been taken in the battlefield, ϵ -like the error-free beliefs, defined in the PE of this game. Hence, using the Bayesian updating process, the soldier and the attacker can dynamically predict their opponent's strategy. As such, rational strategies that maximize each player's payoff can be reached upon the convergence of the Bayesian updating algorithm.

7.5 Simulation Results and Analysis

For our simulations, we consider a battlefield in which multiple IoBT devices are randomly distributed along the soldier's path. The channel gain between the soldier and each IoBT device follows a Rayleigh distribution with unit variance. The parameters used in the simulations are listed in Table I. The *Bayesian updating based PSCE* results, denoted by BU herinafter, are compared to the NE and PE results.

Fig. 7.2 shows the way in which the soldier and attacker make decisions in an IoBT network with 3 devices, using the Bayesian updating algorithm. Fig. 7.2 shows that, at each history in the game, the soldier (attacker) makes decisions on whether to communicate with (attack) the current device

Table 7.1: Simulation Parameters [1]

Parameter	Value	Parameter	Value
P_S	20 dBm	P_A	20 dBm
$J = J'$	1	Δ	80 ms
$\theta_1 = \theta_2$	0.5	$\omega_s = \omega_a$	0.5
I_x	20 MHz	σ^2	-95 dBm

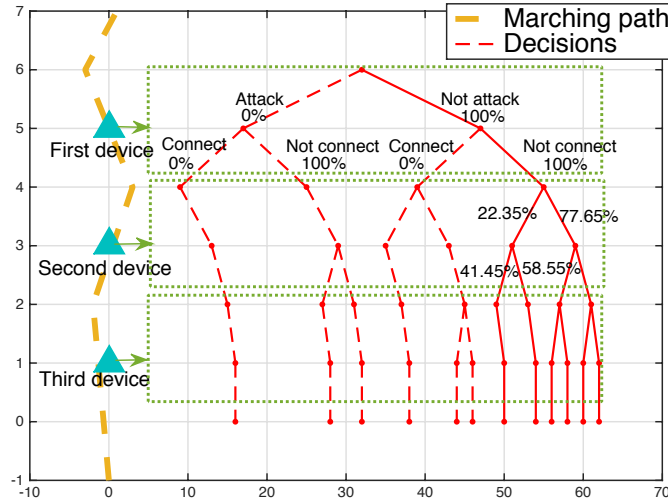


Figure 7.2: The soldier and attacker's strategy in the battlefield.

or not, based on their prediction on their opponent's strategies and beliefs. The soldier's belief on the attacker is updated based on the Bayesian updating algorithm. For example, at a certain history h^x in the game, an IoBT device is attacked $22.35\% \times M$ times in M iterations of the Bayesian updating algorithm, where history h^x is always reached at step x . As such, when history h^x is reached in the $M + 1$ -th iteration of the Bayesian updating algorithm, the soldier will believe that the attacker will attack the current IoBT device with a probability 22.35% . Note that, the soldier updates its beliefs based on the actual strategies played in the game. As the number of soldiers taking the same mission is limited, the opportunity with which the soldier updates its beliefs is limited. Hence, the soldier's beliefs in the battlefield may not be error-free since not enough previous observations are available for the generated beliefs to ϵ -like the correct strategies. In the following simulation results, the effect of the players' non-error-free beliefs will also be studied, in addition to a complete numerical analysis investigating the NE, PE, and BU of the proposed games as well as studying the effects that the various game parameters have on these equilibria. For the following simulations, we consider 5 IoBT devices.

Fig. 7.3 shows the effect that the weights of frustration of each player (i.e. ω_s and ω_a) – which reflects the importance each player assigns on frustrating the opponent – have on the chosen PE strategies. In this regard, Fig. 7.3 shows that, as the weight ω_a increases, the probability that

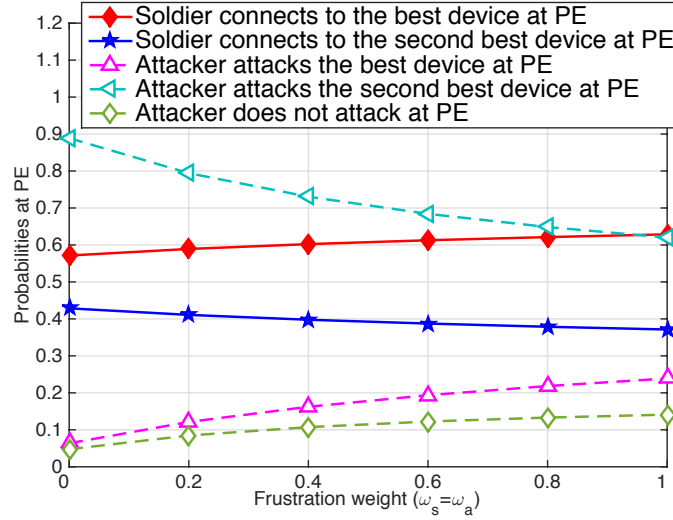


Figure 7.3: The soldier and the attacker's PE strategies as the frustration weights vary.

the attacker attacks the best IoBT device increases. Here, the ranking of best, second best device correspond to the channel quality of that device. Also, the probability that the attacker attacks the IoBT devices decreases with the increase of ω_a , signifying that the attacker's likelihood of launching any jamming attack decreases. This is due to the fact that, attempting to frustrate the soldier, the attacker's strategy can be designed to increase the likelihood of compromising the best IoBT device along the path. Meanwhile, since the attacker also aims at minimizing the total jamming power consumed, its equilibrium strategy will tend towards attacking the IoBT network with a lower probability. In addition, Fig. 7.3 also shows that, as the weight ω_s increases, the probability that the soldier communicates with the IoBT device with best channel quality increases, as likelihood of occurrence of attacks decreases. Note that, when $\omega_s = 0$ and $\omega_a = 0$, the attacker's frustration is not considered in the soldier's payoff, while the soldier's frustration is not considered in the attacker's payoff, which corresponds to the NE of the conventional dynamic game introduced in Section 7.2.4.

Fig. 7.4 shows the convergence of the soldier's and attacker's strategies using the proposed Bayesian updating algorithm. Fig. 7.4 (a) and Fig. 7.4 (b) show that, as the number of iteration increases, the soldier and attacker's strategies ϵ -like their PE counterparts, upon convergence. In the results shown in Fig. 7.4 (a) and Fig. 7.4 (b), Bayesian updating algorithm approximately requires 210 iterations to reach convergence. Fig. 7.5 shows the convergence of the BU expected material payoffs of the soldier and attacker, as their BU beliefs as well as their BU strategies in Fig. 7.4 converge. Fig. 7.5 also shows that, when the player's belief are not error-free, their strategies will not be rational, such that the player's expected material payoff could be less than their expected material payoffs at the PE.

Fig. 7.6 shows the effect of the variation in the frustration weights on the achieved expected material payoffs, of the soldier and the attacker, at PE and BU. In this regard, Fig. 7.6 shows

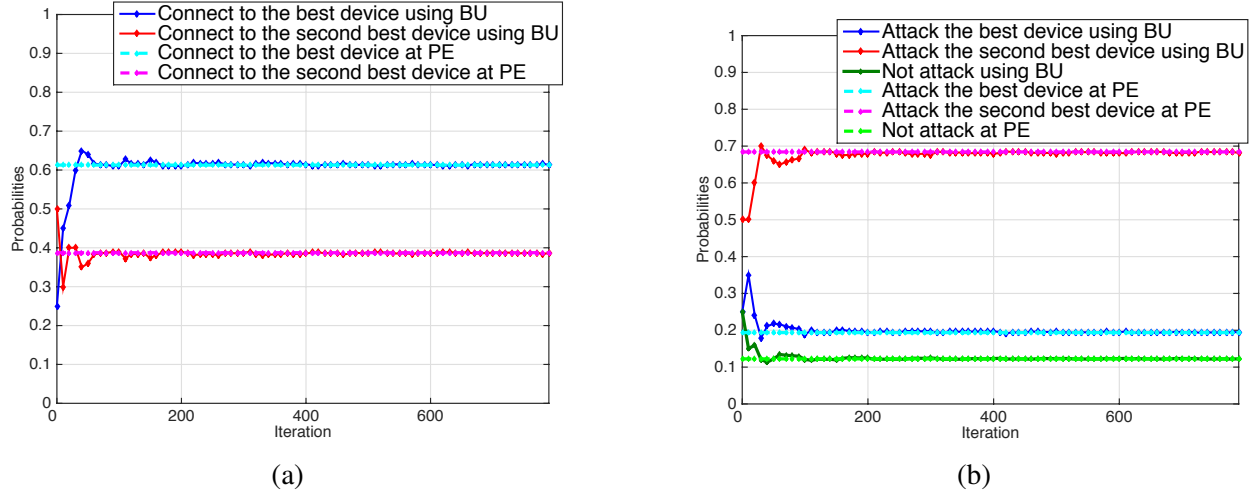


Figure 7.4: Convergence of the BU strategies.

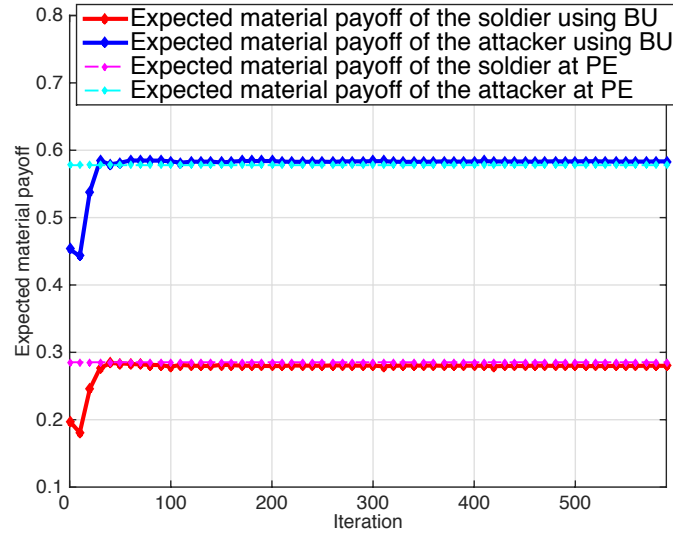


Figure 7.5: Convergence of the BU expected material payoffs.

that, as the weight ω_a increases, the attacker's expected material payoff decreases, as the attacker becomes less apt to launching any attack. Fig. 7.6 also shows that, as the weight ω_s increases, the soldier's expected material payoff increased by up to 15.11% as compared to the expected material payoff the soldier achieves in the conventional dynamic game (which corresponds to the PE expected material payoff of the soldier at $\omega_s = 0$). This stems from the results shown in Fig. 7.4 that at higher frustration weights, the attacker becomes less prone to launching attacks, which leads to an increase in the soldier's expected material payoff when the soldier's beliefs are error-free. Fig. 7.6 also shows that potentially inaccurate beliefs, which occur after 10 iterations in the Bayesian updating algorithm, yield up to 9.12% loss on the soldier's and the attacker's expected

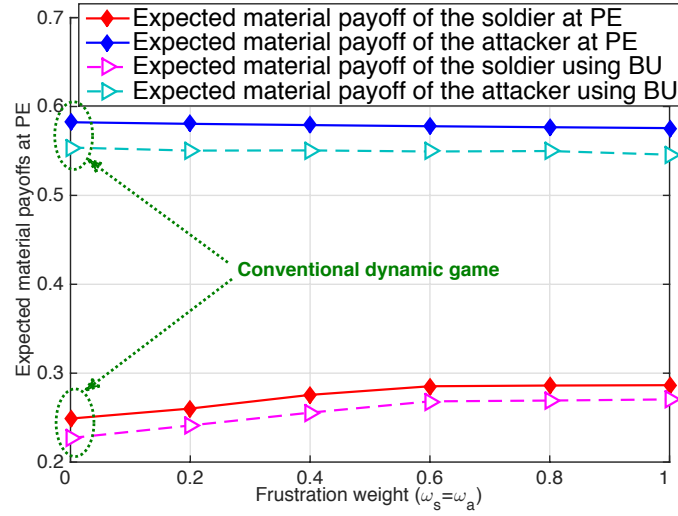


Figure 7.6: The soldier's and the attacker's expected material payoffs as the frustration weights vary (BU results from 10 iterations in Algorithm 1).

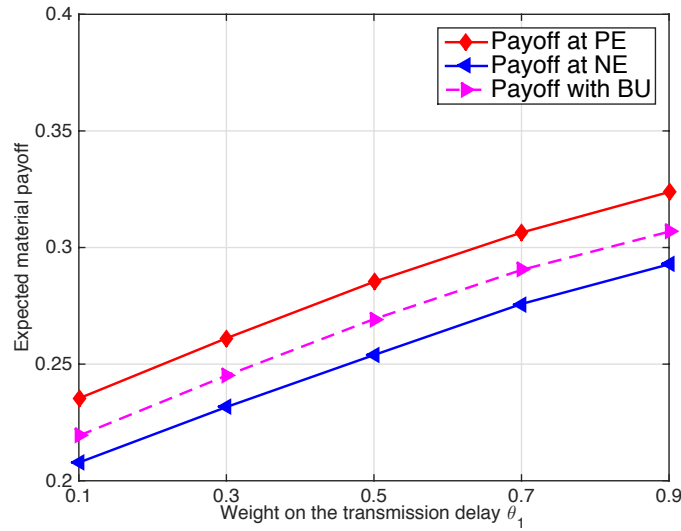


Figure 7.7: The soldier's expected material payoff as the attacker's assigned weight, θ_1 , on maximizing the soldier's communication delay increases.

material payoffs.

Fig. 7.7 highlights the variation in the soldier's and attacker's expected material payoffs at the equilibria as the weight, θ_1 , that the attacker assigns as part of its utility function to maximizing the soldier's communication delay varies. Fig. 7.7 shows that, as the weight θ_1 increases, the soldier's expected material payoff at the NE, PE and BU increases. Here, the BU solution is generated by running 10 iterations of the proposed Bayesian updating algorithm. In this respect, Fig. 7.7 shows that, as the weight θ_1 increases from 0.1 to 0.9, the soldier's expected material payoffs at

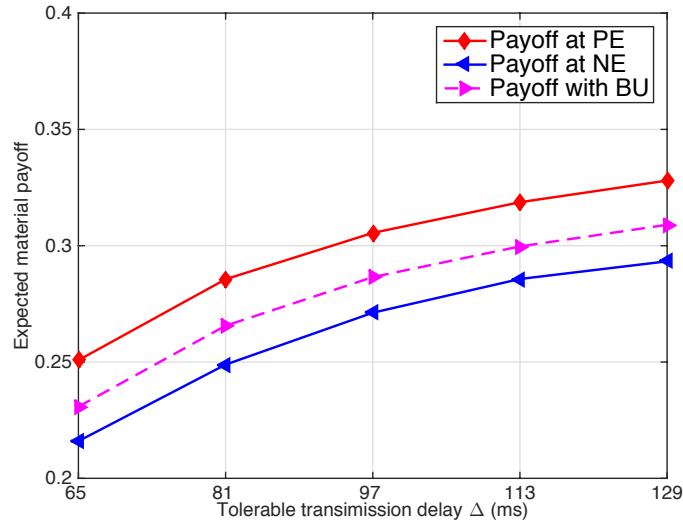


Figure 7.8: The soldier's expected material payoff as its tolerable transmission delay varies.

NE and PE increase up to 37.56% and 39.05%, respectively. This stems from the fact that the attacker's increased intention to maximize the soldier's communication delay causes an increased likelihood (at the equilibrium) that the attacker compromises the best IoBT device along the path. This increased likelihood of attacking the IoBT device with the best channel leads to a decrease in the likelihood of attacking the remaining IoBT devices. With the error-free beliefs on the attacker, the soldier becomes more apt to connecting to the second best device along the path. As such, the soldier's expected material payoff still increases with the increased weight θ_1 . Meanwhile, at 10 iterations of the Bayesian updating algorithm, the resulting beliefs of each of the players would not have totally converged. Hence, at this point, the beliefs of each of the players are not completely error-free. This causes the chosen rational strategies not to completely align with the equilibrium strategies leading to a decrease of up to 9.23% in the resulting expected material payoff of the soldier, as compared to its PE expected material payoff. Hence, similarly to Fig. 7.6, this result also highlights the effect that non-error-free beliefs have on the chosen strategies as well as on the outcome of the game.

Fig. 7.8 shows how the soldier and attacker's expected material payoffs at an equilibrium vary as the soldier's tolerable communication delay Δ increases. In this respect, Fig. 7.8 shows that the soldier's expected material payoff at the NE, PE and BU increases with an increase in Δ . Fig. 7.8 also shows that, as Δ increases, the soldier's expected material payoff at NE and PE increases up to 35.79% and 38.11%, respectively, while the soldier's expected material payoff at the BU increases up to 36.88%. This is due to the fact that, as Δ increases from 65 ms to 129 ms, the difference between the normalized communication delay (i.e., $\frac{\tau}{\Delta}$) of the best channel and other channels decreases. Hence, the soldier becomes less likely, at the equilibrium, to connect to the IoBT device that has the best channel condition. As the probability that the attacker compromises the suboptimal IoBT devices decrease, the likelihood that the soldier evades the attacker's jamming attacker increases. Fig. 7.8 also shows that, when the soldier and attacker choose optimal strategies based

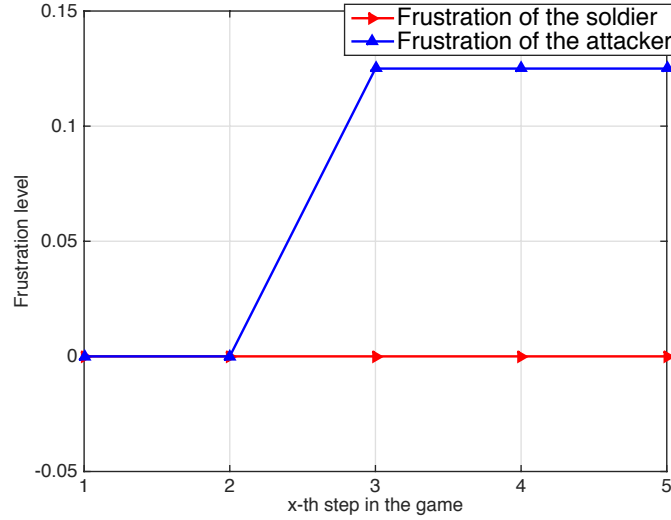


Figure 7.9: Frustration of the soldier and the attacker at the different steps in the battlefield.

on inaccurate beliefs (e.g., after only 10 iterations of Bayesian updating), the soldier's expected material payoff will decrease by up to 8.70%, which is aligned with the results of Fig. 7.6 and Fig. 7.7.

Fig. 7.9 shows the soldier's and attacker's frustration as the soldier progresses from one step to the other in the battlefield. At the used simulation parameters, at the third step of this mission, the attacker attacks the IoBT network, while the soldier does not connect yet to the network. After observing previous sequence of actions in the battlefield, the soldier will connect to the IoBT network at the fourth step. Fig. 7.9 also shows that, as the soldier successfully connects to the IoBT network, without being jammed, the frustration level of the soldier stays at 0. On the other hand, the frustration level of the attacker increases at step 3 since it has attacked a device to which the soldier has not communicated. Fig. 7.9 shows that the frustration level of the attacker increases at step 3, and remains constant until the end of this mission.

Fig. 7.10 shows the variation of the soldier's and attacker's frustration levels at different iterations of the Bayesian updating algorithm. In this regard, Fig. 7.10 shows that, as time elapses, the players' frustration under each of their chosen strategies decreases, as the players' beliefs converge. This, as a result, highlights the effect of having inaccurate beliefs, not only on the expected material payoffs of each player, but also on their resulting frustration levels.

7.6 Conclusion

In this chapter, we have considered an anti-jamming problem in an IoBT network in which an adversary attempts to interdict the connection between a soldier and IoBT devices using jamming.

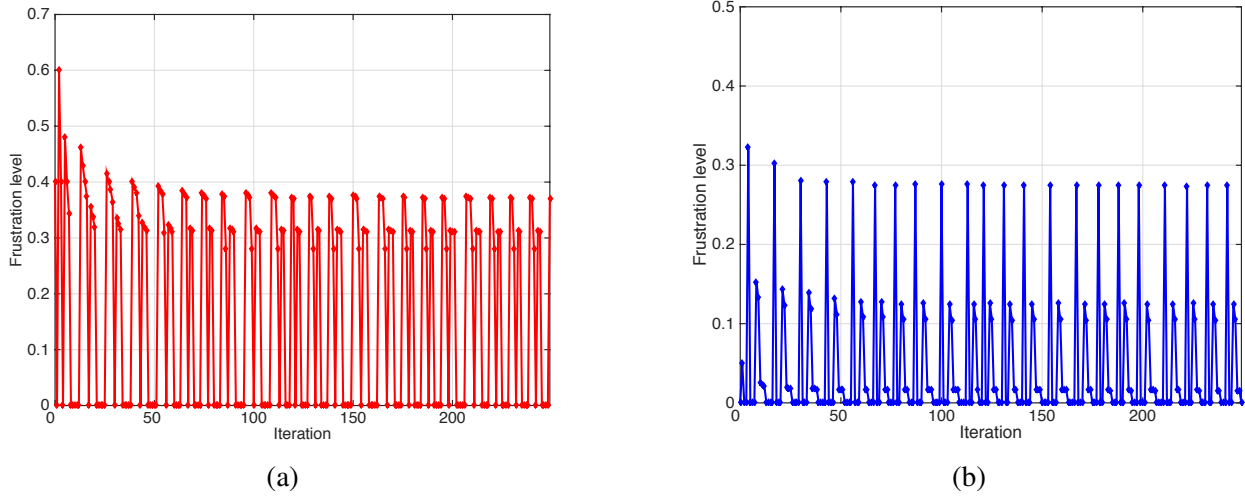


Figure 7.10: Frustration of the soldier and the attacker resulting from BU, as time elapses.

We have formulated this problem as a dynamic psychological game. Due to the reliance of the players' actions on their beliefs, we have used the Bayesian updating to solve this game. The psychological game enables the soldier to determine its actions based on its estimation on the attacker's behavior and belief. Simulation results have shown that, by explicitly intending to frustrate its opponent, the soldier's and attacker's strategies will deviate from their strategies at a conventional, non-psychological NE. Simulation results have also shown that, using the proposed Bayesian updating algorithm, the soldier and the attacker update their beliefs toward their opponent and can reach ϵ -like PSCE strategies for our proposed psychological game.

Chapter 8

Dynamic Psychological Game for Adversarial Internet of Battlefield Things Systems

To maintain low latency data service in battlefield applications, the work in Chapter has investigate the optimal anti-jamming solution within human intervened IoBT system using selective connections. While, in this chapter, we focus on designing soldier trajectories to evade jamming attacks within the battlefield. We propose a psychological game theory based framework to analyze attacking behaviors and optimal trajectories within in the human intervened IoBT systems.

8.1 Background, Related Works, and Contributions

The rapid growth of the IoT technologies has led to significant changes in how autonomous systems are managed [7, 9, 164–167]. IoT technologies provide new ways for managing and operating a battlefield by interconnecting combat equipment, IoT devices, and other battlefield resources [58]. This integration of the IoT with military networks is referred to as the *IoBT* [164]. In a mission-critical IoBT, the connectivity between the IoBT devices, such as wearables carried by soldiers, autonomous vehicles, and drones, will play a significant role in the battlefield operations [59]. However, the adversarial nature of the battlefield coupled with the limitation of the IoBT devices' security mechanisms will render the IoBT network highly vulnerable to cyber attacks [60].

The existing literature has studied a number of problems related to the security of the IoBT [58–60, 99]. The work in [58] investigates the communications and information management challenges of the IoBT. In [59], the authors propose a novel architectural framework which integrates IoT and network centric warfare to enhance the integrity of IoBT networks. In [60], a feedback Stackelberg solution is used to dynamically optimize the connectivity of an adversarial IoBT network. Despite the promising results, these existing works [58–60] do not consider the IoBT devices' mobility,

which introduces a higher degree of dynamics in the IoBT network. The authors in [99] propose a zero-sum network interdiction game based security solution for drone delivery systems, with the mobility of the drone being considered. However, the work in [99] does not consider any communication network. Moreover, in an adversarial battlefield environment, the psychology of the soldiers could significantly influence their behavior, and, subsequently influence the security of the IoBT network, which is not accounted for in [58–60, 99].

More recently, there has been significant interest in studying how human psychology influences game-theoretic decision making [104, 168–170]. In [104], the authors formulate a game-theoretic model that captures dynamic psychological effects and develop new psychological game solution concepts. The authors in [168] study the behavioral consequences of psychology in presence of blaming behaviors. In [169], the effect of the human psychology in mean-field-type games is presented. In [170], the authors investigate the effect of a player's psychology on the solution of one-shot bimatrix games. However, most existing works on psychological game theory and its applications [104, 168–170] have focused on economic or social situations. In fact, none of these works analyzed the potential adoption of psychological game approaches in security scenarios. In an IoBT, psychological game theory provides a suitable mathematical tool to study the psychology of the soldier and attacker, and how this psychology can impact the system's security.

The main contribution of this chapter is a novel security solution for an IoBT network that considers the soldier's mobility along with the psychological factors pertaining to both the soldier and the attacker. In particular, we consider a network interdiction game between a soldier and a malicious attacker. In this game, the soldier acts as an evader who seeks to choose an optimal path along with the best connection with IoBT network devices, to evade attacks and minimize the expected travel time along its chosen travel path. On the other hand, the attacker, referred to as an interdictor, seeks to jam the connection between the soldier and IoBT devices by choosing most suitable IoBT device to compromise. Unlike previous studies [99], which formulate the network interdiction problem with a static game, we propose a dynamic game with imperfect information to capture the dynamic nature of the battlefield. Moreover, to capture the psychology of the soldier and attacker, we incorporate mathematical tools from psychological game theory [104]. In this regard, the soldier and attacker's intention to frustrate each other is formulated based on their estimation on each other's behavior and belief. Based on this estimation, the soldier and attacker can determine their strategies without having perfect information on the IoBT network. To our best knowledge, *this is the first work that exploits the framework of psychological game theory to solve network interdiction problems in an IoBT*. Simulation results show that, when the soldier's psychological intention is accounted for, the proposed approach can improve the soldier's payoff by up to 15.6% compared to a conventional dynamic Bayesian game.

The rest of this chapter is organized as follows. The system model and problem formulation is described in Section 8.2. The psychological forward induction based solution is proposed in Section 8.3. In Section 8.4, numerical simulation results are presented. Finally, conclusions are drawn in Section 8.5.

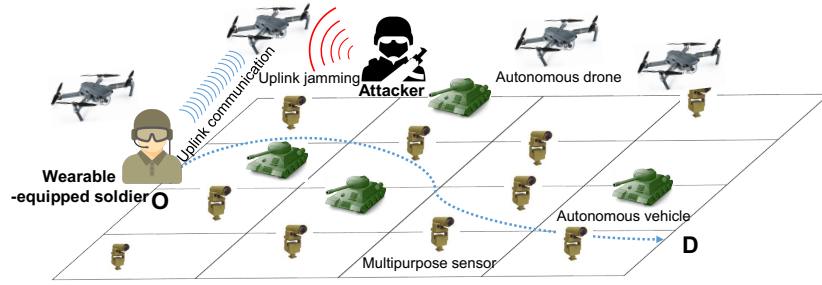


Figure 8.1: Soldier battlefield security graph.

8.2 System Model and Problem Formulation

Consider a $v \times v$ square battlefield, in which a soldier seeks to move from an origin O to a destination D using a minimum amount of time as shown in Fig.1. At the same time, this soldier must maintain a desired bit error rate (BER) performance for its communication links with a total of X uniformly deployed IoBT devices. The soldier can only associate with one IoBT device at each location. In this network, an attacker seeks to disrupt the connectivity between the soldier and an IoBT device by jamming the communication link. As a result, instead of choosing the shortest nonstop physical path, the soldier must consider alternative paths that can potentially improve its BER performance. For example, the soldier will move to certain locations to connect to IoBT devices that have not been attacked yet.

The battlefield is represented by a graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$, with $\mathcal{N}$ being the set of $N = (v + 1)^2$ vertices in the battlefield, and $\mathcal{E}$ being the set of directed edges connecting each pair of vertices. The set of IoBT devices that the soldier can communicate with at vertex i is $\mathcal{X}^i = \{x_1^i, x_2^i, \dots, x_{X_i}^i\}$. The set of IoBT devices that the soldier can communicate with at vertex i and all its neighboring vertices is $\mathcal{X}^{\prime i} = \{x_1^{\prime i}, x_2^{\prime i}, \dots, x_{X_i'}^{\prime i}\}$. As shown in Fig. 2, at vertex 9, the IoBT devices that the soldier can communicate with are $\{x_1^9, x_2^9, \dots, x_5^9\}$. The soldier communicates with each IoBT device $x_l^i \in \mathcal{X}^i$ over an uplink channel l . The SINR γ_l of the uplink channel between the soldier and IoBT device x_l^i is given by:

$$\gamma_l = \begin{cases} \frac{P_S h_l^s}{P_K h_l^a + \sigma^2}, & \text{if channel } l \text{ is attacked,} \\ \frac{P_S h_l^s}{\sigma^2}, & \text{otherwise,} \end{cases} \quad (8.1)$$

where P_S and P_K are, respectively, the transmit powers of the soldier and attacker. $h_l^s = g_l d_l^{-\lambda}$ is the path loss between the soldier and IoBT device x_l^i , with g_l being the Rayleigh fading parameter, d_l being the distance between the soldier and x_l^i , and λ the path loss exponent. $h_l^a = g_l d_l'^{-\lambda}$ is the path loss between the attacker and IoBT device x_l^i , with d_l' being the distance between the attacker and x_l^i . σ^2 is the power of the Gaussian noise. The soldier's BER is given by $r(\gamma_l)$ which is determined by the SINR γ_l , equalization, and modulation [171]. The soldier can only move to one of its neighboring vertices at one step. The set of N_i neighboring vertices of vertex i is represented by $\mathcal{N}_i$. For example, at vertex 9 in Fig. 2, the set of neighboring vertices for the

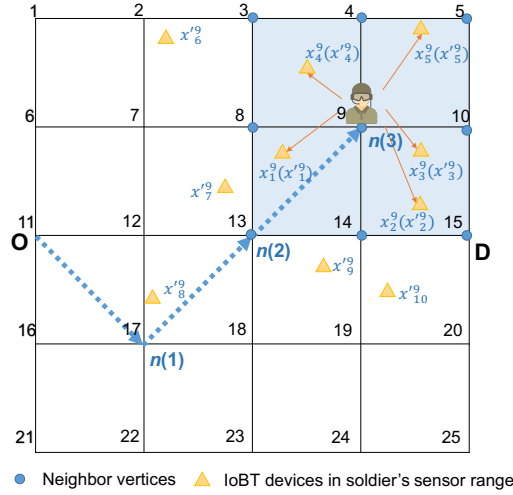


Figure 8.2: Illustration of a soldier's movement in a 4×4 battlefield.

soldier is $\{3, 4, 5, 8, 10, 13, 14, 15\}$. τ_{ij} is the time that the soldier needs to move from vertex i to vertex j over edge e_{ij} . The soldier must be located at one vertex $i \in \mathcal{N}$ and communicate with one IoBT device $x_l^i \in \mathcal{X}_i$ at each step until it arrives at D .

The objective of the soldier is to effectively maintain a low BER while minimizing the travel time. To minimize the BER $r(\gamma_l)$, the soldier will selectively move to one of its neighbor vertices and communicate with one of the IoBT devices located around these vertices. Let $n(t)$ be the soldier's current location. $\mathcal{A} = \{a_1, a_2, \dots, a_{N_{n(t)}}\}$ and $\mathcal{B} = \{b_1, b_2, \dots, b_{X'_{n(t)}}\}$ represent the soldier's feasible actions at each step t . Here, $a_n \in \mathcal{A}$ represents the soldier's movement to vertex $n \in \mathcal{N}_{n(t)}$. Since the soldier can only move to one of its neighbor vertices at each step, the set of feasible actions is determined by the soldier's current location $n(t)$. The set of feasible actions for the soldier at step t will then be represented by $\mathcal{A}(n(t))$. Action $b_l \in \mathcal{B}$ indicates that the soldier communicates with IoBT device x_l^i . In this model, the sets of IoBT devices that the soldier can communicate with at different vertices are different. Thus, the feasible actions for the soldier at step t are captured by $\mathcal{B}(n(t))$. For example, in Fig. 2, the set of feasible actions for the soldier at step 3 are $\mathcal{A}(n(3)) = \{a_3, a_4, a_5, a_8, a_{10}, a_{13}, a_{14}, a_{15}\}$ and $\mathcal{B}(n(3)) = \{b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9, b_{10}\}$.

On the other hand, the objective of the attacker is to increase the BER of the soldier. The attacker will select an IoBT device to compromise, so as to decrease the SINR of the communication channel between this IoBT device and the soldier. The set of the attacker's possible actions can be represented by $\mathcal{C} = \{c_1, c_2, \dots, c_{X'_{n(t)}}\}$. Here, action $c_l \in \mathcal{C}$ indicates that the attacker chooses to attack the IoBT device x_l^i . In the scenario of Fig. 2, the attacker's feasible action set at step 3 is $\mathcal{C}(n(3)) = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}\}$. Given the limited number of IoBT devices located near the soldier's location $n(t)$, the soldier will move to neighboring vertices and will attempt to communicate with IoBT devices that have not been attacked yet. In this case, the soldier's action

$\mathbf{a}_n \in \mathcal{A}(n(t))$ and $b_m \in \mathcal{B}(n(t))$ are all coupled with the attacker's action $c_l \in \mathcal{C}(n(t))$. The soldier's travel time and BER performance are influenced by the attacker's action $c_l \in \mathcal{C}(n(t))$. As a result, the SINR in (1) will be function of the soldier's actions $\mathbf{a}_n$, b_m , and the attacker's action c_l , i.e. $\gamma(\mathbf{a}_n, b_m, c_l)$. The soldier's BER will also be a function of the soldier and attacker's actions, i.e. $r(\mathbf{a}_n, b_m, c_l)$.

In this battlefield, the attacker compromises the IoBT devices to jam the soldier's connections with these devices before the soldier connects to them and, subsequently, the soldier chooses its actions to evade the attack. Both soldier and attacker have an imperfect observation of the IoBT network. In particular, the attacker has a knowledge of the location of the soldier and its associated objective. It does not know the IoBT device that the soldier will be connected to, and the vertex the soldier will move to. The soldier knows that the attacker is present, but does not know which IoBT device the attacker compromised. Thus, the soldier and attacker will determine their optimal actions based on their estimation of their opponent's actions (e.g. the attacker estimates the soldier's actions, and the soldier estimates the attacker's action). This estimation is captured by the soldier and attacker's *beliefs*, respectively. In real battlefield scenarios, the soldier and attacker choose actions based on their beliefs. However, if their opponents did not choose actions according to these beliefs, the soldier and attacker will get a lower payoff than what they expected, and feel *frustrated*. For example, if the soldier believes that the attacker attacked IoBT device x_1^i , it will communicate with IoBT device x_2^i with a BER of $\bar{r}$. If, in reality, the attacker attacked x_2^i , then the BER of the soldier will increase to $\hat{r}$. The gap between $\bar{r}$ and $\hat{r}$ captures the soldier's frustration. Given that the attacker's objective is to maximize the soldier's BER, and the soldier's objective is to minimize its BER, the soldier and attacker will intentionally attempt to frustrate each other while also seek to achieve their own, individual objectives.

Given the clear dependence between the objectives and the actions of the soldier and the attacker, we formulate the IoBT security problem using a dynamic psychological network interdiction game, which will be introduced next.

8.3 Psychological Game Formulation and Solution

8.3.1 Psychological Game Formulation

Network interdiction games are game-theoretic models for scenarios in which evaders attempt to traverse a network without being detected by interdictors [172]. The evader in the studied IoBT scenario is the soldier whereas the interdictor is the attacker. The objective of the attacker is to find a strategy that effectively jams the soldier's communication with IoBT devices. The objective of the soldier is to find a strategy that minimizes its BER as well as the travel time needed to reach the destination.

To capture the belief-dependent motivation of the soldier and attacker, we formulate the IoBT network interdiction game as a psychological game $[\mathcal{P}, u, u', \mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \mathcal{L}]$, where $\mathcal{P}$ is the

set of players which includes the soldier and attacker. u and u' are the utilities of the soldier and the attacker given their mutual actions. $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$ are the sets of the two players' historical actions at each step in the game. Here, we represent the soldier's action history at time t by $\mathbf{h}_1^t$ and $\mathbf{h}_2^t$, with $\mathbf{h}_1^t$ being the set of soldier's actions $\mathbf{a}_n \in \mathcal{A}$ and $\mathbf{h}_2^t$ being the set of the soldier's actions $b_m \in \mathcal{B}$ until step t . Note, at each step t , action $\mathbf{a}_n \in \mathcal{A}$ leads the soldier to move from vertex $n(t)$ to the vertex $n(t+1)$. $\mathbf{h}_1^t$ can, then, be represented by the sequence of traversed vertices $\mathbf{h}_1^t = (n(1), n(2), \dots, n(t))$. An example of $\mathbf{h}_1^3$ is shown in Fig. 2, where the soldier's history $\mathbf{h}_1^3$ at step 3 is (17, 13, 9). Meanwhile, we define the attacker's action history until step t as $\mathbf{h}_3^t = \{c(1), c(2), \dots, c(t)\}$. $\mathcal{Z}$ represents the set of terminal histories, when the soldier reaches the destination D .

In this game, we define the behavioral strategy of the player as the probability distribution over its feasible actions at each history. We define the behavioral strategy vector over the action set $\mathcal{A}$ of the soldier as $\alpha_t = [\alpha_t(1), \alpha_t(2), \dots, \alpha_t(N_{n(t)})]$, and the behavioral strategy vector over the action set $\mathcal{B}$ of the soldier as $\beta_t = [\beta_t(1), \beta_t(2), \dots, \beta_t(X'_{n(t)})]$. The behavioral strategy vector over the attacker's feasible action set $\mathcal{C}$ is $\varepsilon_t = [\varepsilon_t(1), \varepsilon_t(2), \dots, \varepsilon_t(X'_{n(t)})]$. We assume that the soldier and attacker's actions at each step t are independent across different histories from $\mathcal{H}_1$, $\mathcal{H}_2$, and $\mathcal{H}_3$. Then, we can define the probability of the soldier choosing $\mathbf{a}_n$ and b_m conditional on history $\hat{\mathbf{h}}_1^t$ and $\hat{\mathbf{h}}_2^t$, as follows:

$$\begin{aligned} \Pr(\mathbf{a}_n | \hat{\mathbf{h}}_1^t) &= \prod_{\mathbf{h}_1 \in \mathcal{H}_1 / \mathcal{Z}, \mathbf{h}_1 \not\prec \hat{\mathbf{h}}_1^t} \alpha_t(\mathbf{a}_n | \mathbf{h}_1), \\ \Pr(b_m | \hat{\mathbf{h}}_2^t) &= \prod_{\mathbf{h}_2 \in \mathcal{H}_2 / \mathcal{Z}, \mathbf{h}_2 \not\prec \hat{\mathbf{h}}_2^t} \beta_t(b_m | \mathbf{h}_2). \end{aligned} \quad (8.2)$$

where $\mathbf{h}_1 \not\prec \hat{\mathbf{h}}_1^t$ means that $\mathbf{h}_1$ is not a predecessor of $\hat{\mathbf{h}}_1^t$. The probability over the attacker's actions c_l conditional on $\hat{\mathbf{h}}_3^t$ can be given by:

$$\Pr(c_l | \hat{\mathbf{h}}_3^t) = \prod_{\mathbf{h}_3 \in \mathcal{H}_3 / \mathcal{Z}, \mathbf{h}_3 \not\prec \hat{\mathbf{h}}_3^t} \varepsilon_t(c_l | \mathbf{h}_3). \quad (8.3)$$

Let $\delta_t^1 = [\delta_t^1(1), \delta_t^1(2), \dots, \delta_t^1(X'_{n(t)})]$ be the soldier's vector of beliefs on the probability distribution of the attacker's action, and let $\rho_t^1 = [\rho_t^1(1), \rho_t^1(2), \dots, \rho_t^1(N_{n(t)})]$ and $\rho_t'^1 = [\rho_t'^1(1), \rho_t'^1(2), \dots, \rho_t'^1(X'_{n(t)})]$ be the attacker's belief vectors on the probability distribution of the soldier's actions $\mathbf{a}_n \in \mathcal{A}$ and $b_l \in \mathcal{B}$ at step t , respectively. The soldier's beliefs on the attacker's beliefs ρ_t^1 and $\rho_t'^1$ are given by $\delta_t^2 = [\delta_t^2(1), \delta_t^2(2), \dots, \delta_t^2(N_{n(t)})]$ and $\delta_t'^2 = [\delta_t'^2(1), \delta_t'^2(2), \dots, \delta_t'^2(X'_{n(t)})]$, respectively. The attacker's belief on the soldier's belief δ_t^1 is given by $\rho_t^2 = [\rho_t^2(1), \rho_t^2(2), \dots, \rho_t^2(X'_{n(t)})]$. Hereinafter, we use δ_t^1 , δ_t^2 , and $\delta_t'^2$ to denote the soldier's first and second order beliefs on the attacker, and ρ_t^1 , $\rho_t'^1$, ρ_t^2 to denote the attacker's first and second order beliefs on the soldier at step t . We define the soldier's *material payoff* as

the weighted sum of its BER performance and travel time. Meanwhile, we define the attacker's *material payoff* as the BER performance of the soldier. For example, at each step t , the soldier will determine an action that minimizes its BER and travel time, which are captured by the following function:

$$\bar{\pi}(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1) = \theta_1 \frac{\Delta - \tau(\mathbf{h}_1^t) - \tau(\mathbf{a}_n)}{\Delta} + \theta_2 \left[1 - \sum_{y=1}^{X'_{n(t)}} \delta_t^1(y) r(\mathbf{a}_n, b_m, c_y) \right], \quad (8.4)$$

where $\theta_1 \geq 0$ and $\theta_2 \geq 0$ are the respective weights for the travel time and BER, with $\theta_1 + \theta_2 = 1$. $\mathbf{a}_n$ and b_m represent the soldier's action at step t . Δ is the maximum allowed time for the soldier to arrive at D . $\tau(\mathbf{h}_1^t) = \sum_{v=1}^t \tau_{n(v-1), n(v)}$ is the travel time of the soldier under history $\mathbf{h}_1^t$. $\tau(\mathbf{a}_n)$ is the travel time of the soldier under action $\mathbf{a}_n$. δ_t^1 is the soldier's first-order belief.

Similarly, at step t , the attacker's expected payoff when taking action c_l and given beliefs $\rho_t^1, \rho_t'^1$ will be:

$$\bar{\pi}'(c_l, \rho_t^1, \rho_t'^1) = \sum_{x=1}^{N_{n(t)}} \sum_{y=1}^{X'_{n(t)}} \rho_t^1(x) \rho_t'^1(y) r(\mathbf{a}_x, b_y, c_l). \quad (8.5)$$

Note that, the material payoffs of the soldier and attacker do not consider the soldier and attacker's psychology, which will be defined as their intent to frustrate, as explained next.

We define the frustration of the players as the gap between their material payoffs in (4) and (5) and their best material payoffs under their adversary's actual action. For example, at step t , given the attacker's action c_l , the soldiers' frustration F with action $(\mathbf{a}_n, b_m)$ and belief δ_t^1 will be:

$$F(\mathbf{a}_n, b_m, c_l, \mathbf{h}_1^t, \delta_t^1) = [\bar{\pi}(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1) - \hat{\pi}(c_l)]^+, \quad (8.6)$$

where $\hat{\pi}(c_l) = \max_{\mathbf{a}_m \in \mathcal{A}(n(t)), b_m \in \mathcal{B}(n(t))} \{\bar{\pi}(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1)\}$ represents the soldier's best material payoff under the attacker's action c_l . $[x]^+ = \max\{0, x\}$. Note that the attacker has no knowledge of the soldier's belief δ_t^1 and strategies. Based on its beliefs $\rho_t^1, \rho_t'^1$, and ρ_t^2 on the soldier's strategies and first-order belief δ_t^1 , the attacker can formulate the soldier's frustration in the form:

$$F_k(c_l, \mathbf{h}_1^t, \rho_t^1, \rho_t'^1, \rho_t^2) = \left[\theta_1 \sum_{x=1}^{N_{n(t)}} \rho_t^1(x) \frac{\Delta - \tau(\mathbf{h}_1^t) - \tau(\mathbf{a}_x)}{\Delta} + \theta_2 [1 - \tilde{r}] - \hat{\pi}(c_l) \right]^+, \quad (8.7)$$

where $\tilde{r} = \sum_{x=1}^{N_{n(t)}} \sum_{y=1}^{X'_{n(t)}} \sum_{z=1}^{X'_{n(t)}} \rho_t^1(x) \rho_t'^1(y) \rho_t^2(z) r(\mathbf{a}_x, b_y, c_z)$.

If the soldier selects actions $\mathbf{a}_n$ and b_m at step t , the frustration of the attacker with action c_l , under beliefs ρ_t^1 and $\rho_t'^1$, will be:

$$F'(c_l, \mathbf{a}_n, b_m, \rho_t^1, \rho_t'^1) = [\bar{\pi}'(c_l, \rho_t^1, \rho_t'^1) - \hat{\pi}'(\mathbf{a}_n, b_m)]^+, \quad (8.8)$$

where $\hat{\pi}'(\mathbf{a}_n, b_m) = \max_{c_l \in \mathcal{C}(n(t))} r(\mathbf{a}_n, b_m, c_l)$ represents the attacker's best material payoff under the

soldier's actions $\mathbf{a}_n, b_m$. Based on its beliefs $\delta_t^2, \delta_t'^2$ on the attacker's first order beliefs, the soldier can formulate the attacker's frustration in the form:

$$F'_s(\mathbf{a}_n, b_m, \delta_t^1, \delta_t^2, \delta_t'^2) = \left[\sum_{x=1}^{N_{n(t)}} \sum_{y=1}^{X'_{n(t)}} \sum_{z=1}^{X'_{n(t)}} \delta_t^2(x) \delta_t'^2(y) \delta_t^1(z) r(\mathbf{a}_x, b_y, c_z) - \hat{\pi}'(\mathbf{a}_n, b_m) \right]^+. \quad (8.9)$$

Here, the soldier's intention to frustrate the attacker is captured in a belief-dependent utility of the form:

$$u(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1, \delta_t^2, \delta_t'^2) = \bar{\pi}(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1) + \omega_s F'_s(\mathbf{a}_n, b_m, \delta_t^1, \delta_t^2, \delta_t'^2), \quad (8.10)$$

where ω_s represent the soldier's sensitivity parameter on its motivation to frustrate the attacker. $\mathbf{a}_n$ and b_m represent the soldier's action at step t , and $\mathbf{h}_1^t$ represents the set of the soldier's historical action at step t . δ_t^1, δ_t^2 , and $\delta_t'^2$ represent the soldier's first and second order beliefs at step t , respectively. Similarly, the belief-dependent utility of the attacker can be given by:

$$u'(c_l, \mathbf{h}_1^t, \rho_t^1, \rho_t'^1, \rho_t^2) = \bar{\pi}'(c_l, \rho_t^1, \rho_t'^1) + \omega_k F_k(c_l, \mathbf{h}_1^t, \rho_t^1, \rho_t'^1, \rho_t^2), \quad (8.11)$$

where ω_k represents the attacker's sensitivity parameter on its motivation to frustrate the soldier. Given that the players' actions are coupled to their first and second order beliefs on one another, we use the psychological forward induction approach to solve this network interdiction problem, as discussed next.

8.3.2 Psychological Forward Induction Solution

In this section, we develop a psychological forward induction approach to solve the proposed psychological network interdiction game [104]. Using the psychological forward induction solution, the soldier and the attacker determine their optimal strategies considering their belief-independent utilities, at each step t . Here, we first introduce the solution concept of a *psychological sequential equilibrium (PSE)*. Next, the specific process of the psychological forward induction solution is introduced.

We define the assessment of the soldier and attacker as the profile of their beliefs and behavioral strategies. Let $\delta_t^{1*}, \rho_t^{1*}, \rho_t'^{1*}, \delta_t^{2*}, \rho_t^{2*}, \rho_t'^{2*}$ be the beliefs of the soldier and attacker, and $\alpha_t^*, \beta_t^*, \varepsilon_t^*$ be the strategies of the soldier and attacker at the PSE. The PSE assessment in the proposed game is defined as [104]:

Definition 7. An assessment is a PSE if for all $c \in \mathcal{C}$:

$$\begin{aligned} \rho_t^{2*}(l) &= \delta_t^{1*}(l) = \Pr(c_l | \hat{\mathbf{h}}_3^t), \\ \delta_t^{2*}(n) &= \rho_t^{1*}(n) = \Pr(\mathbf{a}_n | \hat{\mathbf{h}}_1^t), \\ \delta_t'^{2*}(m) &= \rho_t'^{1*}(m) = \Pr(b_m | \hat{\mathbf{h}}_2^t), \end{aligned} \quad (8.12)$$

and, for all $\mathbf{a} \in \mathcal{A}, b \in \mathcal{B}, c \in \mathcal{C}$, and for every element $\alpha_t^*(n)$ in the support of α_t^* , we have:

$$\mathbf{a}_n \in \arg \max_{\mathbf{a}_n \in \mathcal{A}(n(t))} u(\mathbf{a}_n, b_m, \mathbf{h}_1^t, \delta_t^1, \delta_t^2, \delta_t'^2), \quad (8.13)$$

for every element $\beta_t^*(m)$ in the support of β_t^* , we have:

$$b_m \in \arg \max_{b_m \in \mathcal{B}(n(t))} u(a_n, b_m, h_1^t, \delta_t^1, \delta_t^2, \delta_t'^2), \quad (8.14)$$

for every element $\varepsilon_t^*(l)$ in the support of ε_t^* , we have:

$$c_l \in \arg \max_{c_l \in \mathcal{C}(n(t))} u'(c_l, h_1^t, \rho_t^1, \rho_t'^1, \rho_t^2). \quad (8.15)$$

Thus, the PSE assessment requires that both the soldier and attacker hold correct beliefs about their mutual strategies and beliefs, and make rational determinations on their strategies at every step of the game. The proposed psychological forward induction procedure is shown in Algorithm 1.

In the formulated network interdiction problem, knowing that the soldier will rationally choose its actions to maximize its utility in (10), the attacker can build its first-order belief on the soldier's strategies. Here, the rational strategy of the soldier is to communicate with the IoBT devices having the best BER performance and lowest travel time, under the condition that the devices have not been attacked. Given that the rational strategy of the attacker is to attack the IoBT devices which optimize the soldier's BER performance, the attacker can build its first-order belief on the soldier's behavior. Under the assumption that both the soldier and the attacker make rational determinations on their strategies, their first-order beliefs δ_t^{1*} , ρ_t^{1*} , and $\rho_t'^{1*}$ must be equal to each another's behavioral strategies, respectively. Also, if the soldier and attacker are rational, they must choose their actions based on their beliefs. Under this assumption, the soldier and attacker can estimate their opponents' first-order beliefs based on their opponents' strategies and, then, build their second-order beliefs. Using the illustrated induction procedure, the soldier and attacker can build their beliefs that satisfy (12).

Given its belief ρ_t^{1*} and $\rho_t'^{1*}$, the attacker will find all the actions $c^* \in \mathcal{C}$ that :

$$u'(c^*, h_1^t, \rho_t^{1*}, \rho_t'^{1*}, \rho_t^{2*}) \geq u'(c_l, h_1^t, \rho_t^{1*}, \rho_t'^{1*}, \rho_t^{2*}), \quad (8.16)$$

for all $c_l \in \mathcal{C}$. To find c^* , at each step t , the attacker will first value its material payoffs with each of its feasible strategies, under its first-order belief on the soldier. Subsequently, the attacker will value the soldier's frustration under each one of its feasible strategies in two steps. First, based on (7), at each step t , the attacker will solve the following linear program with the objective to find the soldier's best material payoff under each one of the attacker's feasible strategies :

$$\max_{a_n \in \mathcal{A}(n(t)), b_m \in \mathcal{B}(n(t))} \{\bar{\pi}(a_n, b_m, h_1^t, \delta_t^{1*})\}, \quad (8.17)$$

Second, the attacker captures the soldier's expected payoff based on its first and second-order belief with (4). Then, using (7) the attacker can estimate the soldier's frustration $F_k(c_l, h_1^t, \rho_t^1, \rho_t'^1, \rho_t^2)$ under each one of its feasible strategies $c_l \in \mathcal{C}$. Finally, the attacker can find its optimal actions c^* that maximizes its utility. In a similar way, the soldier will find all of the actions a^* and b^* , such that:

$$u(a^*, b^*, h_1^t, \delta_t^{1*}, \delta_t^{2*}, \delta_t'^{2*}) \geq u(a_n, b_m, h_1^t, \delta_t^{1*}, \delta_t^{2*}, \delta_t'^{2*}), \quad (8.18)$$

for all $a_n \in \mathcal{A}$ and $b_m \in \mathcal{B}$. Then we can formulate the soldier and attacker's optimal strategies α_t^* , β_t^* , ε_t^* , based on (13), (14), and (15). After this process, (12) is satisfied with the soldier and attacker's induction, (13), (14), and (15) are satisfied for the soldier and attacker's optimization

Algorithm 10 Psychological forward induction solution in the dynamic psychological IoBT network interdiction game

Require: The graph $\mathcal{G}$ of the battlefield, set of IoBT devices $\mathcal{R}$ in the battlefield.

Ensure: Initialize belief of the soldier and attacker, such that $\delta_t^{1*}, \rho_t^{1*}, \rho_t'^{1*}, \delta_t^{2*}, \rho_t^{2*}, \rho_t'^{2*}$ satisfies equation (12).

```

1: for all step  $t$  do
2:   for the attacker do
3:     for all action  $c_l \in \mathcal{C}(n(t))$  do
4:       Calculate the soldier's best material payoff based on (17).
5:       Estimate the attacker expected material payoff with  $\rho_t^2$  and  $\rho_t'^2$ .
6:     end for
7:     Find action  $c^* = \arg \max_{c_l \in \mathcal{C}(n(t))} u'(c_l, h_1^t, \rho_t^1, \rho_t'^1, \rho_t^2)$ , build  $\varepsilon_t^*$ .
8:   end for
9:   for the soldier do
10:    for all action  $a_n \in \mathcal{A}(n(t))$  do
11:      for all action  $b_m \in \mathcal{B}(n(t))$  do
12:        Calculate the attacker best material payoff.
13:        Estimate the attacker expected material payoff with  $\delta_t^2$ .
14:      end for
15:      Find  $b^*(t) = \arg \max_{b_m \in \mathcal{B}(t)} u(a^*, b_m, h_1^t, \delta_t^{1*}, \delta_t^{2*}, \delta_t'^{2*})$ , build  $\beta_t^*$ .
16:    end for
17:    Find  $a^*(t) = \arg \max_{a_n \in \mathcal{A}(t)} u(a_n, b^*, h_1^t, \delta_t^{1*}, \delta_t^{2*}, \delta_t'^{2*})$ , build  $\alpha_t^*$ .
18:  end for
19: end for

```

Table 8.1: System parameters [1]

Parameter	Value	Parameter	Value
P_A	20 dBm	P_K	20 dBm
$\theta_1 = \theta_2,$	0.5	σ^2	-95 dBm

procedure. Thus, the resulting assessment $(\delta_t^{1*}, \rho_t^{1*}, \rho_t'^{1*}, \delta_t^{2*}, \rho_t^{2*}, \rho_t'^{2*}, \alpha_t^*, \beta_t^*, \varepsilon_t^*)$ will be a PSE, when Algorithm 1 converges.

8.4 Simulation Results and Analysis

We evaluate the performance of the proposed psychological forward induction solution using simulations. We consider a single $v \times v$ battlefield with uniformly distributed IoBT devices. The channel gain between the soldier and IoBT device follows a Rayleigh distribution with unit variance. The BER $r(\gamma_l)$ of the soldier is calculated using the LTE system toolbox in MATLAB [173]. The parameters are listed in Table I. The results are compared to a baseline dynamic Bayesian game, which is dynamic imperfect information game with only the material payoffs (with no psychology) of the soldier and attacker being considered.

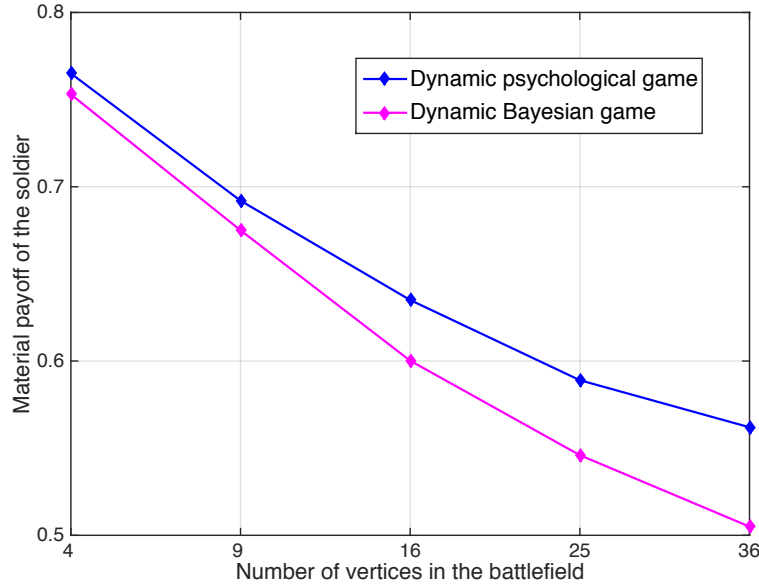


Figure 8.3: The soldier's material payoff as the number of vertices in the battlefield varies. ($X = 20, \omega_s = 0.5$).

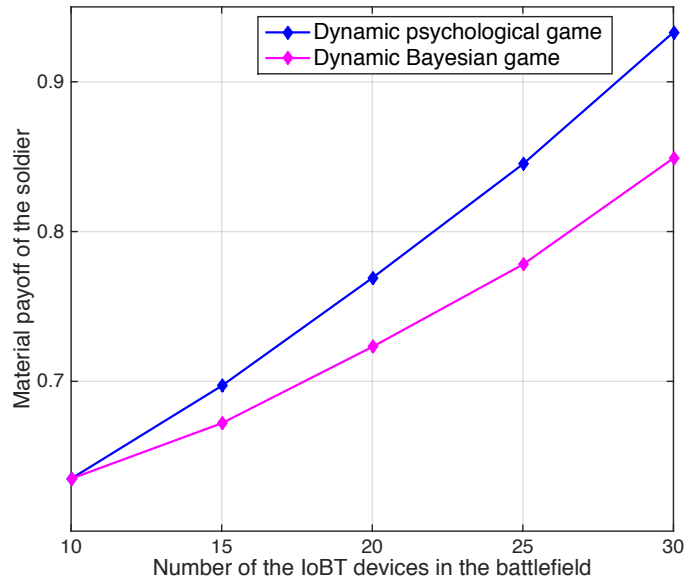


Figure 8.4: The soldier's material payoff as the number of IoBT devices varies. ($v = 3, \omega_s = 0.5$).

Fig. 3 shows how the soldier's material payoff varies with the number of vertices in the battlefield. Note that the maximum step in the game increases with the increase of the number of vertices in the battlefield. In this figure, we can see that, as v increases, the soldier's material payoff decreases with all algorithms. This is because the maximum step of the game increases, and the distances from the IoBT devices to the soldier increase. Fig. 3 also shows that, when the dynamic psycholog-

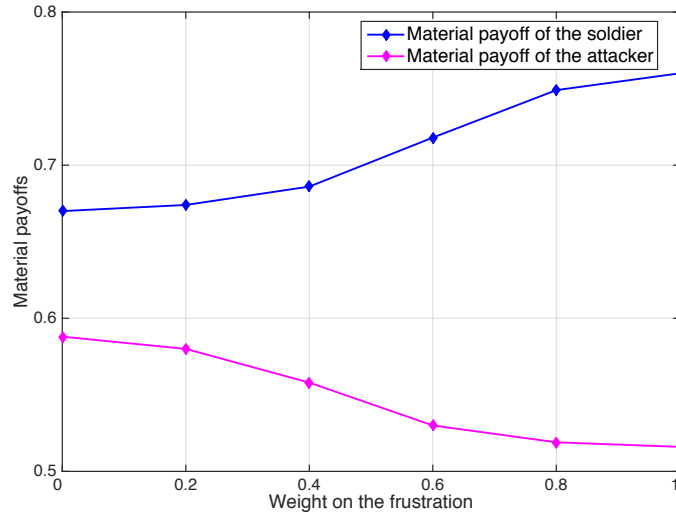


Figure 8.5: The soldier's material payoff as the weight ω of the attacker's frustration in the soldier's utility varies. ($v = 3$, $X = 20$).

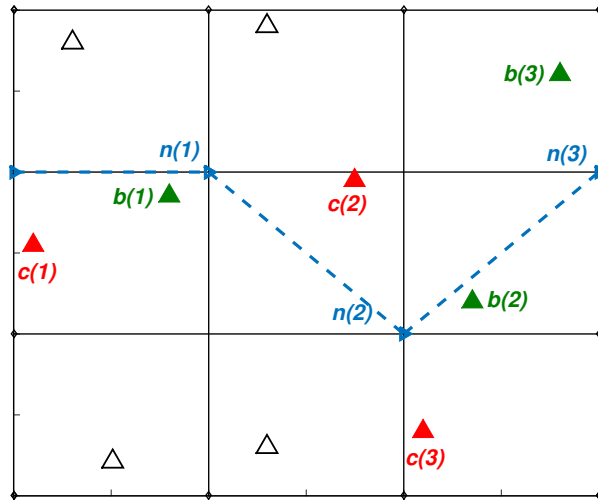


Figure 8.6: The soldier and attacker's strategy. ($v = 3$, $X = 10$).

ical game is used, the soldier's material payoff decreases at a considerably slower rate than when the dynamic Bayesian game is used. This is because, under the objective to evade the attack along the path, the soldier's travel time will increase at a faster rate than the size of the battlefield. Fig. 3 shows that the dynamic psychological game can yield up to 12.5% of improvements in the soldier's material payoff compared with dynamic Bayesian game for a network with 20 IoBT devices. This

is due to the fact that, using the psychological forward induction, the soldier can correctly estimate the attacker's strategy. Then, the soldier's intention to frustrate the attacker leads the soldier choose actions that minimize the attacker's payoff, with the soldier's payoff being increased as a result.

In Fig. 4, we show how the soldier's material payoff varies with the number of IoBT devices. As the number of IoBT devices increases, both solutions yield an increased material payoff because the players have more IoBT devices to choose from and the distances from the IoBT devices to the soldier decrease. Fig. 4 shows that the proposed approach can yield up to 15.6% of improvements in the soldier's material payoff compared with a dynamic Bayesian game. These gains mainly stem from the soldier's better determination on its actions, as a result of its beliefs that are built with the psychological forward induction, and its intention to frustrate the attacker.

Fig. 5 shows how the soldier and attacker's material payoffs vary as the weight of attacker's frustration in the soldier's utility increases. Fig. 5 shows that, as the weight increases, the attacker's material payoff decreases, as the soldier becomes more apt to choose actions that minimize the attacker's payoff. From Fig. 5, we can also see that, as the weight increases, the soldier's material payoff increases. This is due to that the soldier's increased intention to frustrate the attacker leads the soldier to choose actions that minimize the attacker's payoff, resulting with an increase of the attacker's payoff. Note that, when $\omega_s = 0$, the attacker's frustration is not considered on the soldier's payoff, which corresponds to the dynamic Bayesian game.

Fig. 6 shows the soldier and attacker's strategy in a battlefield with 10 uniformly distributed IoBT devices. In this figure, $c(1)$, $c(2)$, and $c(3)$ represent the devices that the attacker compromised and $b(1)$, $b(2)$, and $b(3)$ represent the devices the soldier connected at step 1, 2, and 3, respectively. From Fig. 6, we can see that, with psychological forward induction, the attacker will always attack the devices that can provide the best connections to the soldier, i.e. devices located closest to the soldier. Subsequently, with psychological forward induction, the soldier effectively evades the attacker and frustrates the attacker by connecting to the second best devices located around it.

8.5 Conclusion

In this chapter, we have considered a network interdiction problem in an IoBT network in which an adversary attempts to interdict the connection between a soldier and IoBT devices using jamming. We have formulated this problem as a dynamic psychological game. Due to the reliance of the players' actions on their beliefs, we have used the psychological forward induction to solve this game. The psychological game enables the soldier to determine its actions based on its estimation on the attacker's behavior and belief. Simulation results have shown that, by explicitly intending to frustrate its opponent, the soldier can get a better payoff than the case in which a conventional Bayesian game is used.

Chapter 9

Conclusion and Future Works

9.1 Summary of Contributions

Thus far, the main body of our work in this dissertation can be summarized as follows. In Chapter 1, we have presented the main benefits, applications, and challenges of wireless ISDTC systems. Then, we have investigated a number of the technical challenges within the ISDTC systems. We firstly study the implementation of the integrated communication system. In Chapter 2 and 3, we investigate the improvement on performance of NTN that are integrated with terrestrial communication infrastructures. In the integrated communication system, the LEO satellite constellation provides backhaul connectivity, supplementally, to the DBSs whose demand cannot be satisfied by terrestrial MBSs. In particular, to enable efficient cooperation among the NTN devices (i.e. DBSs and satellites), we have proposed joint access and backhaul resource allocation scheme on both the downlink and uplink services. We have modeled this joint user association and resource allocation problem in a competitive market setting, so as to capture the performance of each communication links, and their dependence on one another. We have proposed a distributed algorithms which are proved to reach a Walrasian equilibrium of the formulated competitive market, at which the maximal sum rate of the system, and data rate at each communication link are reached.

Then, we have investigate the effective deployment of NTNs for on demand service delivery to users with unforeseeable needs. In particular, in Chapter 4, we have studied the problem of contact plan design on cube satellite constellations that are trying to provide low latency service coverage to ground users with diverse and dynamic access needs. We proposed a distribution-robust meta reinforcement learning framework that routs data among the resource constraint cube satellites with prior estimation on the user requests at different service occasions. The proposed solution helps the cube satellites reach optimal service coverage with reduced computation complexity, and communication overhead. We then study the implementation of aerial communication platforms carried out by the UAVs and satellites.

In Chapter 5 and 6, we studied the problem of trajectory design for UAV carried aerial radio ac-

cess points, which aim at providing on-demand coverage to dynamic, unpredictable ground access requests. To this end, we proposed a reinforcement learning based framework that designs trajectories on DBSs for effective wireless service coverage and proposed a distributed MARL algorithm that coordinates the DBSs with low time complexity, and minimum overhead. Then, we proposed meta-training based mechanisms for reduced computation cost on learning the optimal trajectories in unseen environments.

After investigating the challenges on the ISDTC system from the communication perspective, we then looked into the security related problems. In particular, we have optimized the security of an IoBT system. We have optimized a soldier's connectivity with the IoBT devices under the jamming attacks from an adversarial attacker, while considering the soldier and attacker's psychological behavior in the battlefield. We have captured the soldier and attacker's psychological decision making processes using a dynamic psychological game framework, and proposed a novel Bayesian updating based solution which is proved to converge to an ϵ -like the error-free belief. With an error-free belief, the soldier can reach a primary estimation on the attacker's behaviors and, thus, make decisions that lead to optimal connectivity.

While we have addressed several key challenges in ISDTC-based wireless communication systems, some of our results need to be further extended and a number of new directions can be explored. Therefore, in this chapter, we present an overview of the future directions of this dissertation. For future work, we will consider three research directions and applications of UAVs and satellites in wireless networks.

9.2 Meta-Federated Learning for Beam Training within Satellite Drone Networks

In Chapter 6, we have investigated the spectrum management problem within an ISDN system operating on millimeter wave (mmWave) frequencies. However, to cope with the severe path loss of the mmWave links for reliable connectivity, the ISDN base stations (including all the terrestrial, aerial, and space base stations) must exploit highly directional communication. Within such mmWave communication systems, even a slight beam mis-alignment between two communicating devices can generate a significant performance drop. In fact, the communication devices must perform a time-consuming beam training procedure in order to determine the best directions of transmission/reception, which incurs significant overhead (and waste of network resources). In addition, the beam training procedure must be frequently repeated which could lead to a massive overhead in the ISDN system, with its highly mobile devices including the DBSs, satellites, and ground users (i.e. CAVs). To sum up, a fast and efficient beam training scheme is of paramount importance to maintain seamless, reliable connectivities in the mmWave ISDN system.

Thus, in this future work, our goal is to investigate effective beam training for an ISDN system. In this ISDN system, the UAVs provide downlink connectivity, supplementally, to ground users

whose demand cannot be satisfied by terrestrial SBSs. Meanwhile, a satellite system and a set of terrestrial MBSs are used to provide backhaul connectivity for the DBSs and SBSs. In this system, the ground users and UAVs move along certain unknown trajectories, the satellites move on their pre-defined orbits. Such mobility causes difficulties for beam training in the system, and a light weight ML based beam training solution must be introduced to provide reliable connectivities in the ISDN system. In particular, we propose to use a meta-federated learning algorithm to predict channel directions, i.e. the angle of arrivals (AoAs), for effective beam training at the ISDN system. Such solution attributes the AoA prediction task to each base station to reduce the complexity of the the studied problem. The associated meta-training mechanism enhances the base stations beam training performance in unseen environments.

9.3 Trajectory design with Variable Drone Functionality

In Chapter 5, we have investigated the trajectory design at multiple UAVs for on-demand coverage to certain ground users. However, in practical wireless scenarios, drone carried aerial radio access points can be highly vulnerable to environmental changes, because of: 1) the high sensitivity of the UAVs' performance to wireless parameters such as propagation environment, or user mobility, 2) the need to maintain consistent backhaul connectivity and desirable radio service within already disadvantaged areas, 3) efficient management of coordination and cooperation among multiple UAVs for optimal wireless performance, and 4) increased exposure to cyber and physical attacks, severe weather conditions, and so on. This means that, the launched DBSs may not be able to finish their data delivery tasks when affected by the poor propagation on their associated radio access or backhaul links. At the meantime, some of launched DBSs may not be able to fly along the trajectories designed for maximal wireless service coverage. Indeed, the DBSs may have to return to their charging stations when facing severe weather conditions, and can be incapable of flying with improper configuration.

Thus, to enable a continuous and efficient deployment of drone carried base stations, the control solutions on the DBSs must learn to adjust the trajectories with the variable functionality of applicable drones. Thus, one must firstly predict the relationship between the number of DBSs the DBSs' achievable service coverage. By relying on such prediction, a machine learning mechanism that designs trajectories on DBSs for maximal service coverage, with the variable drone functionality, should be studied. Then, meta training mechanisms should be investigated to help the DBSs to adjust their trajectories online when some of the DBSs fail on their service delivery missions.

9.4 Secure Satellite Drone Systems from Data Leakage

To make its pervasive data service reliable, the integrated satellite drone system must secure its data transmission from several cyber threats such as jamming, and eavesdropping attacks. How-

ever, data transmission over NTN channels are more vulnerable, compared to the terrestrial communication channels, to cyber attacks since: 1) information transmitted over the LoS dominated NTN communication links are more likely to be intercepted by undesired receivers (i.e eavesdroppers) and, thus, cause information leakage. 2) the high altitude NTN transceivers can be easily exposed to malicious jamming attacks through their LoS dominated communication links. 3) the NTN-ground connections suffer from high communication latency, which causes the loss of time synchronization, and, thus, the effectiveness of security mechanisms in the satellite communication systems. Yet it is rather challenging to design security schemes for satellites and UAVs with their limited computation ability and energy resource.

Thus, a key goal here is to develop novel security solutions considering the complex, multi-layer structure, and limited resource capacity of the integrated satellite drone systems, such solutions include: 1) Privacy-reserving meta federated learning based 3D beamforming to reduce the possibilities of information leakage with minimal training cost. In particular, it is of interest to develop algorithms that can train wireless strategies distributedly at each interdependent device in the system, so as to reduce overhead, and complexity of the considered privacy-reserving connectivity optimization problem. 2) Artificial noise intelligently carried out by UAVs and satellites to reduce the capacity of eavesdropping channels. In particular, we will propose task-robust meta learning solutions to quickly adapt UAVs in various unknown environments, which includes expected environments from certain distribution and the ones out of the expected distribution. The proposed solution will adjust UAV learning process within the task of reaching a noise launching scheme effectively interfering every possible eavesdroppers from unknown locations.

Appendices

Appendix A

Appendices for Chapter 2

A.1 Proof of Proposition 3

Proof. Based on (2.1), (2.11), and (2.14), the data rate of link I_n^S at slot t can be given as:

$$c_{n,S,t} = \beta_{nt} B_S \log \left(1 + \frac{P_S L_1 Q_{MM} Q_{RM} d_{n,S}^{-0.1}(t) 10^{-0.1\alpha' - 0.1\chi}}{\Omega_{n,S,t} + \Omega_c + \sigma^2} \right), \quad (\text{A.1})$$

where $d_{n,S}(t) = \sqrt{(x_0 + v\tau - x_{2,n})^2 + (y_0 - y_{2,n})^2 + (f_0 - f_{2,n})^2}$ is the distance between the satellite and BS n at slot t , with $\frac{\partial d_{n,S}(t)}{\partial t} = \frac{(x_0 - x_{2,n})v\tau + v^2\tau^2 t}{d_{n,S}(t)}$. Also, we define function $g(t)$ in a form of:

$$g(t) = d_{n,S}^{-0.1}(t) 10^{-0.1\alpha' - 0.1\chi}, \quad (\text{A.2})$$

such that:

$$\frac{\partial g(t)}{\partial t} = -\frac{\alpha(x_0 v\tau + v^2\tau^2 t)}{10d_{n,S}^2(t)} g(t). \quad (\text{A.3})$$

Denote $P'_S = \frac{P_S L_1 Q_{MM} Q_{RM}}{\Omega_{n,S,t} + \Omega_c + \sigma^2}$, such that $c_{n,S,t}(t) = \beta_{nt}^* B_S \log_2(1 + P'_S g(t))$. In such a case:

$$\begin{aligned} \frac{\partial c_{n,S,t}(t)}{\partial t} &= \frac{\beta_{nt}^* B_S}{\ln 2} \frac{P'_S g'(t)}{1 + P'_S g(t)} \\ &= -\frac{\alpha \beta_{nt}^* B_S P'_S}{10 \ln 2} \frac{((x_0 - x_{2,n})v\tau + v^2\tau^2 t) g(t)}{(1 + P'_S g(t)) d_{n,S}^2(t)}. \end{aligned} \quad (\text{A.4})$$

As each studied time slot has a rather short time duration τ , we can approximate the variation of $c_{n,S,t}$ in one time slot as $\frac{\partial c_{n,S,t}(t)}{\partial t} \tau$. This completes the proof. $\square$

A.2 Proof of Lemma 3

Proof. We first, relax the integer constraints (4.10p)-(4.10r) into $\rho_{nut}, \theta_{nut}, \delta_{nmt}, \varphi_{nmt}, \beta_{nt}, \varepsilon_{nt} \in [0, 1]$, and, thus, the maximization problem (4.10) becomes a convex problem. Based on [174], when the continuous relaxation of (4.10) is feasible with at least one integer solution, and the constraints of (4.10) can be written in the form of $A\psi \leq b$, with A being an integer matrix whose entries are all integers and $\psi = [\rho, \delta, \beta, \theta, \phi, \varepsilon]^T$, the problem (4.10) will have at least one optimal integer solution. In particular, when the value of the data rates studied in the constraints are rounded into rational numbers, constraints (4.10a)-(4.10s) can be transformed in the form of $A\psi \leq b$. This completes the proof. $\square$

A.3 Proof of Lemma 2

Proof. We first observe that, with the non-summable diminishing updating step size $\pi^{(k)}$, the heavy ball algorithm is guaranteed to converge to optimal solutions of the studied dual minimization problem based on [117]. With the convergence of the heavy ball algorithm at iteration K^* , market clearance violations tend to be 0, meaning that $\theta^{(K^*)} = \rho^{(K^*)}$, $\varphi^{(K^*)} = \delta^{(K^*)}$ and $\varepsilon^{(K^*)} = \beta^{(K^*)}$, with $\theta^{(K^*)}$, $\rho^{(K^*)}$, $\varphi^{(K^*)}$, $\delta^{(K^*)}$, $\varepsilon^{(K^*)}$, and $\beta^{(K^*)}$ being, respectively, the optimal solutions of payoff maximization problems (2.24)-(2.28) at iteration K^* . Thus, at the convergence of the heavy ball algorithm, a Walrasian equilibrium in the competitive market is reached.

Also, as the constraints are all linear equalities and inequalities, and the feasible region defined by these constraints are assumed to have at least one interior point, the strongly duality between the primal total payoff maximization problem (4.10) and its dual problem holds based on the Slater's condition [116]. The Slater's condition claims that, for any feasible linear optimization problem, the optimal solution of its dual problem is the optimal solution of its primal optimization problem, or in other words, strong duality holds. This means that the proposed solution can find the optimal solution for an optimization problem defined in the form of (4.10), if feasible, upon its convergence to the optimal solution of (4.10)'s dual problem. This completes the proof. $\square$

Appendix B

Appendices for Chapter 4

B.1 Proof of Proposition 2

Proof. To prove the convergence of the proposed VD-RL algorithm, we only need to prove that the proposed VD-RL algorithm satisfies the following conditions in [137]:

1. Value function $V(\mathbf{S}_k)$ converges to a local minimum value of $A^2(\mathbf{a}_k, \mathbf{S}_k)$, that is

$$2 \sum_{k=1}^K A(\mathbf{a}_k, \mathbf{S}_k) \nabla_{\theta_c} A(\mathbf{a}_k, \mathbf{S}_k) = 0. \quad (\text{B.1})$$

2. Policy function parameters are updated as in

$$\theta_a^{(i+1)} = \theta_a^{(i)} + \alpha_a^{(i)} \sum_{k=1}^K A(\mathbf{a}_k, \mathbf{S}_k) \nabla_{\theta_a} \log \pi_{\theta_a}(\mathbf{a}_k | \mathbf{S}_k). \quad (\text{B.2})$$

3. $\max_{\mathbf{a}_{n,k} \in \mathcal{C} \cup \{O\}, \mathbf{s}_{n,k} \in \mathcal{S}} \left| \frac{\partial \pi_{\theta_a}(\mathbf{a}_k | \mathbf{S}_k)}{\partial \theta_i \partial \theta_j} \right| < \infty$, for any elements θ_i and θ_j in the policy function parameter vector θ_a .

4. $\lim_{i \rightarrow \infty} \alpha_c^{(i)} = 0$, $\sum_{i=1}^{\infty} \alpha_c^{(i)} = \infty$, $\lim_{i \rightarrow \infty} \alpha_a^{(i)} = 0$, and $\sum_{i=1}^{\infty} \alpha_a^{(i)} = \infty$.

Next, we prove that the proposed VD-RL algorithm satisfies condition 1). When value function $\tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k})$ converges to a local minimum, such that

$$2A(\mathbf{a}_k, \mathbf{S}_k) \nabla_{\theta_{c,n}} \tilde{V}_{\theta_{c,n}}(\mathbf{s}_{n,k}) = 2A(\mathbf{a}_k, \mathbf{S}_k) \nabla_{\theta_{c,n}} \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) = 0. \quad (\text{B.3})$$

with $A(\mathbf{a}_k, \mathbf{S}_k) \neq 0$, $\nabla_{\theta_{c,n}} \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) = 0$. Thus, $2 \sum_{k=1}^K \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) \nabla_{\theta_{c,n}} \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) = 0$. Thus, VD-RL algorithm satisfies condition 1).

Next, we can see that the VD-RL's distributed update on policy function parameters satisfies condition 2), as each DBS n updates its policy function parameters in the form of $\theta_{a,n}^{(i+1)} =$

$\theta_{a,n}^{(i)} + \alpha_a^{(i)} \sum_{k=1}^K \tilde{A}_n(a_{n,k}, \mathbf{s}_{n,k}) \nabla_{\theta_{a,n}} \log \pi_{\theta_{a,n}}(a_{n,k} | \mathbf{s}_{n,k})$. Thus, the distributed update on policy parameters $\theta_{a,n}$ satisfies condition 2). Condition 3) can be satisfied by properly setting neural network of the policy functions in the proposed VD-RL algorithm (e.g. setting activation function). Meanwhile, condition 4) can be satisfied by adjusting the step sizes of the policy and value functions. In consequence, the proposed algorithm implemented at each DBS n satisfies all conditions from 1) to 4). In other words, each DBS n is guaranteed to converge to an optimal strategy π_n^* that yields a local maximal team utility $\bar{G}(\pi_n^*, \pi_{-n})$, with DBSs $n' \in \mathcal{N} \setminus n$ following strategies in $\pi_{-n} = [\pi_{n'}]_{n' \in \mathcal{N} \setminus n}$. In summery, by updating policies at each DBS in the system, the proposed VD-RL algorithm solves problem (6.5) step by step in the form of

$$\begin{aligned} \max_{\pi_1} \dots \max_{\pi_N} \sum_{\xi \in \mathcal{E}} G(\xi) \prod_{k=1}^K \pi_1(\xi | \xi_{1,k}, \tau_{1,k}) \prod_{k=1}^K \pi_2(\xi | \xi_{2,k}, \tau_{2,k}) \\ \dots \prod_{k=1}^K \pi_N(\xi | \xi_{N,k}, \tau_{N,k}) \end{aligned} \quad (\text{B.4})$$

$$\text{s. t.} \quad \sum_{\xi \in \mathcal{E}} \prod_{n=1}^N \prod_{k=1}^K \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) = 1, \quad (\text{B.4a})$$

$$\sum_{a_{n,k} \in \mathcal{A} \cup \{O\}} \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) = 1, \forall n \in \mathcal{N}, \xi \in \mathcal{E}, k \in \mathcal{K}, \quad (\text{B.4b})$$

$$0 \leq \pi_n(\xi | \xi_{n,k}, \tau_{n,k}) \leq 1, \forall n \in \mathcal{N}, \xi \in \mathcal{E}, k \in \mathcal{K}, \quad (\text{B.4c})$$

As the DBSs' strategies are independent, the local optimal strategy at each DBS n constructs a local optimal strategy of the non-convex problem (6.5). This completes the proof. $\square$

Appendix C

Appendices for Chapter 5

C.1 Proof of Proposition 3

Proof. Based on (6.7), the policy update direction at iteration i can be expressed as

$$\begin{aligned}
 & \nabla_{\boldsymbol{\theta}^{(i)}} J \left(\eta^{(i)}, \boldsymbol{\theta}^{(i)} \right) \\
 &= \sum_{k=1}^K \sum_{\kappa=k}^K \left(\eta^{(i)} \right)^{\kappa-k} r \left(a_{\kappa} | \mathbf{s}_{\kappa} \right) \nabla_{\boldsymbol{\theta}^{(i)}} \log \pi_{\boldsymbol{\theta}^{(i)}} \left(a_k | \mathbf{s}_k \right) \\
 &= \sum_{k=1}^K \sum_{\kappa=k}^K \left(\eta^{(i)} \right)^{\kappa-k} r \left(a_{\kappa} | \mathbf{s}_{\kappa} \right) \frac{\nabla_{\boldsymbol{\theta}^{(i)}} \pi_{\boldsymbol{\theta}^{(i)}} \left(a_k | \mathbf{s}_k \right)}{\pi_{\boldsymbol{\theta}^{(i)}} \left(a_k | \mathbf{s}_k \right)}. \tag{C.1}
 \end{aligned}$$

Also, to tune hyper-parameters η in the direction that achieves the best service success rate in the studied network, we measure η by cross validating the updated parameter $\boldsymbol{\theta}^{(i+1)}$ on second trajectory e' . In particular, the tuning direction at step i is given in the form of

$$\nabla_{\eta^{(i)}} \tilde{J} \left(\tilde{\eta}, \boldsymbol{\theta}^{(i+1)} \right) = \frac{\partial \tilde{J} \left(\tilde{\eta}, \boldsymbol{\theta}^{(i+1)} \right)}{\partial \boldsymbol{\theta}^{(i+1)}} \frac{d \boldsymbol{\theta}^{(i+1)}}{d \eta^{(i)}}, \tag{C.2}$$

where

$$\frac{\partial \tilde{J} \left(\tilde{\eta}, \boldsymbol{\theta}^{(i+1)} \right)}{\partial \boldsymbol{\theta}^{(i+1)}} = \sum_{k=1}^K \sum_{\kappa=k}^K r \left(a_{\kappa} | \mathbf{s}_{\kappa} \right) \frac{\nabla_{\boldsymbol{\theta}^{(i+1)}} \pi_{\boldsymbol{\theta}^{(i+1)}} \left(a_k | \mathbf{s}_k \right)}{\pi_{\boldsymbol{\theta}^{(i+1)}} \left(a_k | \mathbf{s}_k \right)}, \tag{C.3}$$

Meanwhile, $\frac{d\theta^{(i+1)}}{d\eta^{(i)}}$ can be expressed by

$$\begin{aligned}
\frac{d\theta^{(i+1)}}{d\eta^{(i)}} &= \frac{d \left[\theta^{(i)} + \alpha \nabla_{\theta^{(i)}} J \left(\eta^{(i)}, \theta^{(i)} \right) \right]}{d\eta^{(i)}} \\
&= \frac{d\theta^{(i)}}{d\eta^{(i)}} + \frac{\alpha d \sum_{k=1}^K \sum_{\kappa=k}^K \left(\eta^{(i)} \right)^{\kappa-k} r(a_{\kappa} | \mathbf{s}_{\kappa}) \frac{\nabla_{\theta^{(i)}} \pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}{\pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}}{d\eta^{(i)}} \\
&= \frac{d\theta^{(i)}}{d\eta^{(i)}} + \alpha \sum_{k=1}^K \sum_{\kappa=k}^K (\kappa - k) \left(\eta^{(i)} \right)^{\kappa-k-1} r(a_{\kappa} | \mathbf{s}_{\kappa}) \frac{\nabla_{\theta^{(i)}} \pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}{\pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}. \tag{C.4}
\end{aligned}$$

Here, we decay the first term in (C.4), to effectively tuning η [175]. In such a case, the first term is rewritten as $\mu \frac{d\theta^{(i)}}{d\eta^{(i)}}$, where $\mu \in [0, 1]$ is the decay factor. In practice, we approximation the first term by setting $\mu = 0$ to reduce computation complexity. In essence, we only consider hyper-parameter η 's effect on a single policy update. In summary, the tuning direction is finally approximated as

$$\nabla_{\eta^{(i)}} \tilde{J} \left(\tilde{\eta}, \theta^{(i+1)} \right) \approx \alpha \sum_{k'=1}^K \sum_{k=1}^K A_{k'} B_k x_{k'} y_k, \tag{C.5}$$

where $A_{k'} = \sum_{\kappa=k'}^K r(a_{\kappa} | \mathbf{s}_{\kappa})$, $x_{k'} = \frac{\nabla_{\theta^{(i+1)}} \pi_{\theta^{(i+1)}}(a_k | \mathbf{s}_k)}{\pi_{\theta^{(i+1)}}(a_k | \mathbf{s}_k)}$, $B_k = \sum_{\kappa=1}^K (\kappa - k) \left(\eta^{(i)} \right)^{\kappa-k-1} r(a_{\kappa} | \mathbf{s}_{\kappa})$, and $y_k = \frac{\nabla_{\theta^{(i)}} \pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}{\pi_{\theta^{(i)}}(a_k | \mathbf{s}_k)}$. This completes the proof. $\square$

Appendix D

Appendices for Chapter 7

D.1 Proof of Theorem 1

At history $h^x \in \mathcal{H}^x$, the soldier chooses its action from $\{a_1, a_2\}$, while the attacker chooses its action from $\{b_1, b_2\}$. We represent the soldier's and attacker's payoffs when the soldier takes action a_n and the attacker takes action b_m , where $n, m \in \{1, 2\}$, by $\pi_{n,m}$ and $\pi'_{n,m}$, respectively. Here, $\pi_{n,m}$ and $\pi'_{n,m}$ include the instantaneous payoffs the soldier and attacker receive when taking their action pair at history h^x as well as future expected payoffs at the following histories. As such, under each combination of the soldier's and attacker's pure strategies at h^x , the soldier's possible payoffs are represented by $\pi_{1,1}$, $\pi_{1,2}$, $\pi_{2,1}$ and $\pi_{2,2}$, while the attacker's payoffs are represented by $\pi'_{1,1}$, $\pi'_{1,2}$, $\pi'_{2,1}$ and $\pi'_{2,2}$. Note that, here, we consider that the following inequalities hold: $\pi_{1,2} > \pi_{2,2} \geq \pi_{1,1}$, and $\pi_{2,1} > \pi_{2,2} \geq \pi_{1,1}$. Indeed, $\pi_{1,2} > \pi_{2,2}$ reflects the gain that the soldier receives from communicating with the IoBT network without being jammed by the attacker, while $\pi_{2,2} \geq \pi_{1,1}$ reflects the loss the attacker incurs from attempting to communicate with a jammed IoBT network. In addition, $\pi_{2,1} > \pi_{2,2}$ reflects the gain the soldier will receive in future steps due to the attacker wasting some of its jamming power when the soldier had not attempted to communicate with the IoBT network. Similarly, we also consider the following inequalities to hold: $\pi'_{1,1} \geq \pi'_{2,2} > \pi'_{1,2}$, $\pi'_{1,1} \geq \pi'_{2,2} > \pi'_{2,1}$. These inequalities correspond to considering that: i) evaded attacks at history h^x yield higher expected material payoffs for the soldier and lower expected material payoffs for the attacker, at the current and future histories, and ii) a successful (unjammed) communication at history h^x yields a higher expected material payoff for the soldier and a lower expected material payoff for the attacker, at the current and future histories. In this respect, we can compute the psychological payoff of the soldier and the attacker under each combination of these pure strategies. In this regard, we consider α' to be the attacker's belief representing the probability with which the attacker believes that the soldier will choose action a_1 . In addition, we consider β' to be the belief that the soldier has, representing the probability with which the soldier believes that the attacker will choose action b_1 . Under the correctness of beliefs defined in (7.29) and (7.30) of the PE, these probabilities also reflect the second-order beliefs of the players

as well as the actual strategies chosen by each of the players. Starting from the soldier's side, the soldier's psychological payoffs at the strategy pairs $\{a_1, b_1\}$ and $\{a_1, b_2\}$ are, respectively, $\pi_{1,1}$ and $\pi_{1,2} + \omega_s (1 - \alpha') (\pi'_{2,2} - \pi'_{1,2})$. In addition, psychological payoffs of the soldier at the pure strategy pairs $\{a_2, b_1\}$ and $\{a_2, b_2\}$ are, respectively, $\pi_{2,1} + \omega_s \alpha' (\pi'_{1,1} - \pi'_{2,1})$ and $\pi_{2,2}$. On the other hand, the attacker's psychological payoffs at strategy pairs $\{a_1, b_1\}$ and $\{a_1, b_2\}$ are, respectively, $\pi'_{1,1} + \omega_a (1 - \beta') (\pi_{1,2} - \pi_{1,1})$ and $\pi'_{1,2}$. In addition, the attacker's psychological payoffs at strategy pairs $\{a_2, b_1\}$ and $\{a_2, b_2\}$ are, respectively, $\pi'_{2,1}$ and $\pi'_{2,2} + \omega_a \beta' (\pi_{2,1} - \pi_{2,2})$.

In the conventional dynamic game, in which the frustrations of the soldier and attacker are not considered in their opponent's utility functions, we denote the soldier's and attacker's strategies by $[\alpha, 1 - \alpha]$ and $[\beta, 1 - \beta]$, respectively. By using the indifference principle, we can compute the NE strategy of the soldier, which results in $\alpha = \frac{\pi'_{2,2} - \pi'_{2,1}}{\pi'_{1,1} + \pi'_{2,2} - \pi'_{1,2} - \pi'_{2,1}}$, and the NE strategy of the attacker, which results in $\beta = \frac{\pi_{2,2} - \pi_{1,2}}{\pi_{1,1} + \pi_{2,2} - \pi_{1,2} - \pi_{2,1}}$. Here, we note that this computed NE is unique since it can be shown that no NE exists in pure strategies, under our considered set of inequalities defined at the start of the proof, and the solution of the equations resulting from the indifference principle results in unique mixed-strategies α and β .

Now, considering the psychological game, we denote the soldier's and the attacker's strategies at the PE by $[\alpha', 1 - \alpha']$ and $[\beta', 1 - \beta']$, respectively. By using the indifference principle, with the soldier and the attacker holding correct (i.e. error-free) beliefs on one another, we can compute the PE strategies as follows:

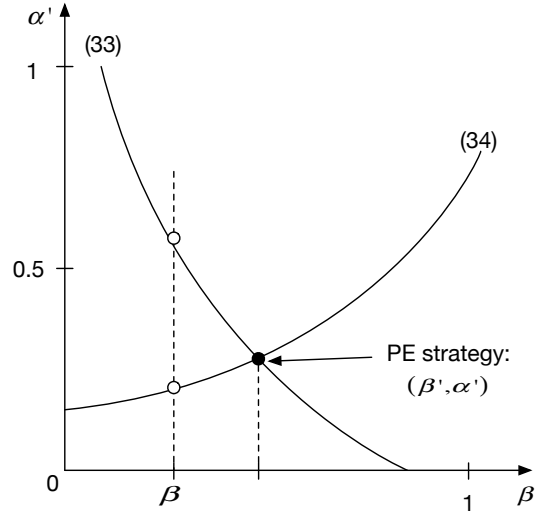
$$\alpha' = \frac{\omega_a \beta' (\pi_{2,1} - \pi_{2,2}) + (\pi'_{2,2} - \pi'_{2,1})}{D' + \omega_a [\beta' (\pi_{2,1} - \pi_{2,2}) + (1 - \beta') (\pi_{1,2} - \pi_{1,1})]}, \quad (D.1)$$

$$\beta' = \frac{\omega_s (1 - \alpha') (\pi'_{2,2} - \pi'_{1,2}) + (\pi_{1,2} - \pi_{2,2})}{D + \omega_s [(1 - \alpha') (\pi'_{2,2} - \pi'_{1,2}) + \alpha' (\pi'_{1,1} - \pi'_{2,1})]}. \quad (D.2)$$

where $D = \pi_{1,2} + \pi_{2,1} - \pi_{1,1} - \pi_{2,2}$, $D' = \pi'_{1,1} + \pi'_{2,2} - \pi'_{1,2} - \pi'_{2,1}$. Note that, in (D.1), when $\beta' = 0$, $\alpha' < \alpha$, when $\beta' = 1$, $\alpha' > \alpha$. At the same time, in (D.2), when $\beta' = 0$, $\alpha' = 1 + \frac{\pi_{1,2} - \pi_{2,2}}{\omega_s (\pi'_{2,2} - \pi'_{1,2})} > \alpha$, when $\beta' = 1$, $\alpha' = \frac{\pi_{1,1} - \pi_{2,1}}{\omega_s (\pi'_{1,1} - \pi'_{2,1})} < \alpha$. Hereinafter, we rewrite (D.1) as:

$$\alpha' = \frac{F_3 \beta' + F_4}{F_1 \beta' + F_2}, \quad (D.3)$$

where $F_1 = \omega_a (\pi_{2,1} + \pi_{1,1} - \pi_{1,2} - \pi_{2,2})$, $F_2 = \pi'_{2,2} - \pi'_{2,1} + \pi'_{1,1} - \pi'_{1,2} + \omega_a (\pi_{1,2} - \pi_{1,1}) > 0$, $F_3 = \omega_a (\pi_{2,1} - \pi_{2,2}) > 0$, $F_4 = \pi'_{2,2} - \pi'_{2,1} > 0$. Meanwhile, in (D.3), $\frac{\partial \alpha'}{\partial \beta'} = \frac{F_2 F_3 - F_1 F_4}{(F_1 \beta' + F_2)^2}$. Note that, when $\pi'_{2,2} - \pi'_{2,1} \geq \theta_1 (\pi_{2,1} - \pi_{2,2})$, we can prove that $F_2 F_3 - F_1 F_4 > 0$, which implies that, in (D.1), α' increases with an increase in $\beta' \in [0, 1]$.

Figure D.1: α' versus β' .

Meanwhile, we rewrite equation (D.2) with:

$$\alpha' = \frac{F'_3\beta' + F'_4}{F'_1\beta' + F'_2}, \quad (\text{D.4})$$

where $F'_1 = \omega_s (\pi'_{1,2} + \pi'_{1,1} - \pi'_{2,1} - \pi'_{2,2})$, $F'_2 = \omega_s (\pi'_{2,2} - \pi'_{1,2}) > 0$, $F'_3 = -D - \omega_s (\pi'_{2,2} - \pi'_{1,2}) < 0$, $F'_4 = \pi_{1,2} - \pi_{2,2} + \omega_s (\pi'_{2,2} - \pi'_{1,2}) > 0$. In (D.4), $\frac{\partial \alpha'}{\partial \beta'} = \frac{F'_2 F'_3 - F'_1 F'_4}{(F'_1 \beta' + F'_2)^2}$. Here, $F'_2 F'_3 - F'_1 F'_4 \leq \omega_s (\pi'_{2,2} - \pi'_{1,2}) (\pi_{1,1} - \pi_{2,1}) < 0$. Thus $\frac{\partial \alpha'}{\partial \beta'} < 0$, which implies that, in (D.2), α' decreases with an increase in $\beta' \in [0, 1]$.

In conclusion, when $\beta' = 0$, α' in (D.1) is smaller than α' in (D.2), while when $\beta' = 1$, α' in (D.1) is larger than α' in (D.2). In (D.1), α' is strictly increasing in β' , while, in (D.2), α' is strictly decreasing in β' . As such, for $0 \leq \beta' \leq 1$, (D.1) and (D.2) has 1 intersection point. This implies that the solution obtained from the indifference principle is unique and there is a unique PE in mixed strategies. In addition, given the inequalities stated at the beginning of the proof, it can be readily shown that no PE exists in pure strategies. Therefore, under the considered set of inequalities, the PE of the game is unique. This completes the proof.

D.2 Proof of Theorem 2

In (D.2), $\alpha' = \frac{\omega_s (\pi'_{2,2} - \pi'_{1,2}) (\pi_{2,1} - \pi_{1,1})}{\omega_s [(\pi'_{2,2} - \pi'_{1,2}) (\pi_{2,1} - \pi_{1,1}) + (\pi'_{1,1} - \pi'_{2,1}) (\pi_{1,2} - \pi_{2,2})]} \geq \frac{1}{2}$ when $\beta' = \beta$, since $\pi'_{1,1} - \pi'_{2,1} \leq \theta_1 (\pi_{2,1} - \pi_{1,1})$, $\pi'_{2,2} - \pi'_{1,2} \geq \theta_1 (\pi_{1,2} - \pi_{2,2})$. In (D.1), when $\beta' = \beta$, $\alpha' < \frac{1}{2}$ if $\pi_{2,1} < \pi_{1,2}$. As such, if $\pi_{2,1} < \pi_{1,2}$ (i.e. if the current IoBT device exhibits the best channel as compared to all

future IoBT devices), $\beta' > \beta$, at the intersection of (D.1) and (D.2), as shown in Fig.11. Indeed, $\pi_{1,2} > \pi_{2,1}$ reflects the gain that the soldier receives from communicating with the current IoBT device is larger than the gain it may potentially receive if it successfully communicates with the IoBT network in future steps. Thus, when there is no remaining IoBT devices in future steps, which have a better channel as compared to the current device, the attacker will be more likely to attack the current device. This completes the proof.

Bibliography

- [1] G. Bianchi, “Performance analysis of the IEEE 802.11 distributed coordination function,” *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, Mar. 2000.
- [2] A. Al-Hourani, S. Kandeepan, and A. Jamalipour, “Modeling air-to-ground path loss for low altitude platforms in urban environments,” in *IEEE Global Communications Conference*, Austin, USA, Dec 2014, pp. 2898–2904.
- [3] M. Alouini and A. J. Goldsmith, “Adaptive modulation over nakagami fading channels,” *Wireless Personal Communications*, vol. 13, no. 1-2, pp. 119–143, 2000.
- [4] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, Jun. 2015.
- [5] W. Rafique, L. Qi, I. Yaqoob, M. Imran, R. U. Rasool, and W. Dou, “Complementing iot services through software defined networking and edge computing: A comprehensive survey,” *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1761–1804, May 2020.
- [6] D. Ma, G. Lan, M. Hassan, W. Hu, and S. K. Das, “Sensing, computing, and communications for energy harvesting IoTs: A survey,” *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 1222–1250, Dec. 2019.
- [7] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Mobile unmanned aerial vehicles (UAVs) for energy-efficient internet of things communications,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7574–7589, Nov. 2017.
- [8] Y. Zhou, F. Zhou, H. Zhou, D. W. K. Ng, and R. Q. Hu, “Robust trajectory and transmit power optimization for secure UAV-enabled cognitive radio networks,” *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4022–4034, Jul 2020.
- [9] M. Chen, M. Mozaffari, W. Saad, C. Yin, M. Debbah, and C. S. Hong, “Caching in the sky: Proactive deployment of cache-enabled unmanned aerial vehicles for optimized quality-of-experience,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1046–1061, 2017.

- [10] Z. Xiao, P. Xia, and X. Xia, "Enabling UAV cellular with millimeter-wave communication: potentials and approaches," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 66–73, May 2016.
- [11] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, May 2016.
- [12] R. I. Bor-Yaliniz, A. El-Keyi, and H. Yanikomeroglu, "Efficient 3-D placement of an aerial base station in next generation cellular networks," in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, Jul. 2016, pp. 1–5.
- [13] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Unmanned aerial vehicle with underlaid device-to-device communications: Performance and tradeoffs," *IEEE Transactions on Wireless Communications*, vol. 15, no. 6, pp. 3949–3963, Feb. 2016.
- [14] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, Jul. 2014.
- [15] L. You, K. X. Li, J. Wang, X. Gao, X. G. Xia, and B. Ottersten, "Massive MIMO transmission for LEO satellite communications," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1851–1865, Jun. 2020.
- [16] R. Radhakrishnan, W. W. Edmonson, F. Afghah, R. M. Rodriguez-Orsorio, F. Pinto, and S. C. Burleigh, "Survey of inter-satellite communication for small satellite systems: Physical layer to network layer view," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2442–2473, May 2016.
- [17] C. Niephaus, M. Kretschmer, and G. Ghinea, "QoS provisioning in converged satellite and terrestrial networks: A survey of the state-of-the-art," *IEEE Communications Surveys Tutorials*, vol. 18, no. 4, pp. 2415–2441, Apr. 2016.
- [18] C. Liu, W. Feng, Y. Chen, C. Wang, and N. Ge, "Cell-free satellite-UAV networks for 6G wide-area internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1116–1131, Apr. 2021.
- [19] J. Zhao, F. Gao, Q. Wu, S. Jin, Y. Wu, and W. Jia, "Beam tracking for UAV mounted SatCom on-the-move with massive antenna array," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 363–375, Feb. 2018.
- [20] Z. Jia, M. Sheng, J. Li, D. Niyato, and Z. Han, "LEO satellite-assisted UAV: Joint trajectory and data collection for internet of remote things in 6G aerial access networks," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9814–9826, Jun. 2020.
- [21] M. Li, Y. Hong, C. Zeng, Y. Song, and X. Zhang, "Investigation on the UAV-to-satellite optical communication systems," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2128–2138, Aug. 2018.

- [22] J. Yu, X. Liu, Y. Gao, and X. Shen, "3D channel tracking for UAV-satellite communications in space-air-ground integrated networks," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 12, pp. 2810–2823, Dec. 2020.
- [23] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio, and R. Prasad, "Satellite communications supporting internet of remote things," *IEEE Internet of Things Journal*, vol. 3, no. 1, pp. 113–123, Oct. 2016.
- [24] Q. Wang, G. Ren, S. Gao, and K. Wu, "A framework of non-orthogonal slotted aloha (NOSA) protocol for TDMA-based random multiple access in IoT-oriented satellite networks," *IEEE Access*, vol. 6, pp. 77 542–77 553, Nov. 2018.
- [25] T. 38.811, "Study on new radio (NR) to support non-terrestrial networks," *VI5.4.0*, Oct. 2020.
- [26] T. 38.821, "Solutions for NR to support non-terrestrial networks (NTN)," *VI6.0.0*, 2020.
- [27] T. 36.763, "Study on narrow-band internet of things (NB-IoT) / enhanced machine type communication (eMTC) support for non-terrestrial networks (NTN)," *V0.0.1*, Jan. 2021.
- [28] T. 22.822, "Study on using satellite access in 5G," *VI6.0.0*, Aug. 2018.
- [29] T. 23.737, "Study on architecture aspects for using satellite access in 5G," *VI7.1.0*, Jul. 2020.
- [30] T. 28.808, "Study on management and orchestration aspects of integrated satellite components in a 5G network," *VI.1.0*, Jan. 2021.
- [31] T. 22.926, "Guidelines for extra-territorial 5G systems (5GS)," *V0.2.0*, Dec. 2020.
- [32] T. 24.821, "Study on PLMN selection for satellite access," *V0.3.0*, Feb. 2021.
- [33] N. Neshenko, E. Bou-Harb, J. Crichigno, G. Kaddoum, and N. Ghani, "Demystifying IoT security: An exhaustive survey on iot vulnerabilities and a first empirical look on internet-scale iot exploitations," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2702–2733, Apr. 2019.
- [34] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (IoT) security," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1646–1685, Apr. 2020.
- [35] I. Bor-Yaliniz and H. Yanikomeroglu, "The new frontier in RAN heterogeneity: Multi-tier drone-cells," *IEEE Communications Magazine*, vol. 54, no. 11, pp. 48–55, Nov. 2016.
- [36] J. KoÄmerl and A. Vilhar, "Base stations placement optimization in wireless networks for emergency communications," in *International Conference on Communications Workshops (ICC)*, Sydney, NSW, Australia, Aug. 2014, pp. 200–205.

- [37] A. Merwaday and I. Guvenc, "UAV assisted heterogeneous networks for public safety communications," in *Wireless Communications and Networking Conference Workshops (WCNCW)*, New Orleans, LA, USA, Jun. 2015, pp. 329–334.
- [38] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proceedings of the IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.
- [39] L. Zhu, J. Zhang, Z. Xiao, X. Cao, X. G. Xia, and R. Schober, "Millimeter-wave full-duplex UAV relay: Joint positioning, beamforming, and power control," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2057–2073, Jun. 2020.
- [40] T. Perry. (2017, Apr.) Facebook aims to remake telecom with millimeter waves and tether-tennas. [Online]. Available: <http://spectrum.ieee.org/view-from-the-valley/telecom/wireless/facebook-aims-to-remake-the-telecom-infrastructure-with-millimeter-waves-andtethertennas>.
- [41] M. Zuckerberg, "Connecting the world from the sky," 2014.
- [42] (2016, Oct.) F-cell technology from nokia bell labs revolutionizes small cell deployment by cutting wires costs and time. [Online]. Available: <https://tex.stackexchange.com/questions/3587/how-can-i-use-bibtex-to-cite-a-web-page>
- [43] (2019, Apr.) Erc perfume project. [Online]. Available: <http://www.ercperfume.org/about/>
- [44] B. Di, H. Zhang, L. Song, Y. Li, and G. Y. Li, "Ultra-dense LEO: Integrating terrestrial-satellite networks into 5G and beyond for data offloading," *IEEE Transactions on Wireless Communications*, Dec. 2019.
- [45] Z. Wang and P. T. Mathiopoulos, "On the performance analysis of dynamic channel allocation with FIFO handover queuing in LEO-MSS," *IEEE Transactions on Communications*, vol. 53, no. 9, pp. 1443–1446, Dec. 2005.
- [46] R. Deng, B. Di, S. Chen, S. Sun, and L. Song, "Ultra-dense LEO satellite offloading for terrestrial networks: How much to pay the satellite operator?" *IEEE Transactions on Wireless Communications*, vol. 19, no. 10, pp. 6240–6254, Jun. 2020.
- [47] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband leo satellite communications: Architectures and key technologies," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 55–61, Apr. 2019.
- [48] A. Botta and A. Pescape, "On the performance of new generation satellite broadband internet services," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 202–209, Jun. 2014.
- [49] Wikipedia contributors, "Starlink — Wikipedia, the free encyclopedia," 2020, [Online; accessed 25-October-2020]. [Online]. Available: <https://en.wikipedia.org/w/index.php?title=Starlink&oldid=985268118>

- [50] M. Zolanvari, R. Jain, and T. Salman, "Potential data link candidates for civilian unmanned aircraft systems: A survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 292–319, Dec. 2019.
- [51] S. Lien, K. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, Apr. 2011.
- [52] Y. Yao, Y. Sun, C. Phillips, and Y. Cao, "Movement-aware relay selection for delay-tolerant information dissemination in wildlife tracking and monitoring applications," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 3079–3090, Apr. 2018.
- [53] H. S. Dhillon, H. Huang, and H. Viswanathan, "Wide-area wireless communication challenges for the internet of things," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 168–174, Feb. 2017.
- [54] Z. Xue, J. Wang, G. Ding, H. Zhou, and Q. Wu, "Maximization of data dissemination in UAV-supported internet of things," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 185–188, Aug. 2019.
- [55] —, "Cooperative data dissemination in air-ground integrated networks," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 209–212, Aug. 2019.
- [56] G. Li, J. He, S. Peng, W. Jia, C. Wang, J. Niu, and S. Yu, "Energy efficient data collection in large-scale internet of things via computation offloading," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4176–4187, Oct. 2019.
- [57] H. Hu, K. Xiong, G. Qu, Q. Ni, P. Fan, and K. B. Letaief, "AoI-minimal trajectory planning and data collection in UAV-assisted wireless powered iot networks," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 1211–1223, Jan. 2021.
- [58] M. Tortonesi, A. Morelli, M. Govoni, J. Michaelis, N. Suri, C. Stefanelli, and S. Russell, "Leveraging internet of things within the military network environment challenges and solutions," in *Proc. of IEEE World Forum on Internet of Things (WF-IoT)*, Reston, USA, Dec. 2016.
- [59] P. P. R., "Towards an internet of things based architectural framework for defence," in *Proc. of International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT)*, Kumaracoil, India, Dec. 2015, pp. 411–416.
- [60] N. Abuzainab and W. Saad, "Dynamic connectivity game for adversarial internet of battlefield things systems," *IEEE Internet of Things Journal*, vol. 5, no. 1, pp. 378–390, Dec 2017.
- [61] N. B. Gaikwad, H. Ugale, A. Keskar, and N. C. Shivaprakash, "The internet of battlefield things (IoBT) based enemy localization using soldiers location and gunshot direction," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 725–11 734, Dec. 2020.

- [62] M. J. Farooq and Q. Zhu, "On the secure and reconfigurable multi-layer network design for critical information dissemination in the internet of battlefield things (IoBT)," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2618–2632, Feb. 2018.
- [63] J. L. Burbank, P. F. Chimento, B. K. Haberman, and W. T. Kasch, "Key challenges of military tactical networking and the elusive promise of manet technology," *IEEE Communications Magazine*, vol. 44, no. 11, pp. 39–45, Nov. 2006.
- [64] F. T. Johnsen, T. HafsÅe, A. Eggen, C. Griwodz, and P. Halvorsen, "Web services discovery across heterogeneous military networks," *IEEE Communications Magazine*, vol. 48, no. 10, pp. 84–90, Oct. 2010.
- [65] I. Agadacos, G. F. Ciocarlie, B. Copos, J. George, N. Leslie, and J. Michaelis, "Security for resilient iobt systems: Emerging research directions," in *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Paris, France, Apr. 2019, pp. 1–6.
- [66] E. D. Buenrostro, A. O. G. Rivera, D. Tosh, J. C. Acosta, and L. Njilla, "Evaluating usability of permissioned blockchain for internet-of-battlefield things security," in *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*, Norfolk, VA, USA, Mar. 2019, pp. 841–846.
- [67] D. Orfanus, E. P. de Freitas, and F. Eliassen, "Self-organization as a supporting paradigm for military uav relay networks," *IEEE Communications Letters*, vol. 20, no. 4, pp. 804–807, Feb. 2016.
- [68] V. Weerackody and L. Gonzalez, "Mobile small aperture satellite terminals for military communications," *IEEE Communications Magazine*, vol. 45, no. 10, pp. 70–75, Oct. 2007.
- [69] Y. Huang, J. Mo, X. and Xu, L. Qiu, and Y. Zeng, "Online maneuver design for uav-enabled noma systems via reinforcement learning," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.
- [70] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2059–2070, Aug. 2018.
- [71] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 18, no. 4, pp. 2125–2140, Mar 2019.
- [72] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Transactions on Communications*, vol. 68, no. 11, pp. 6807–6821, Nov. 2020.

- [73] J. Cui, Y. Liu, and A. Nallanathan, “Multi-agent reinforcement learning-based resource allocation for UAV networks,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 729–743, Aug. 2020.
- [74] X. Liu, Y. Liu, Y. Chen, and L. Hanzo, “Trajectory design and power control for multi-UAV assisted wireless networks: A machine learning approach,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7957–7969, May 2019.
- [75] M. Chen, D. Gündüz, K. Huang, W. Saad, M. Bennis, A. V. Feljan, and H. V. Poor, “Distributed learning in wireless networks: Recent progress and future challenges,” *IEEE Journal on Selected Areas in Communications*, to appear, 2021.
- [76] C. E. Palazzi, C. Roseti, M. Luglio, M. Gerla, M. Y. Sanadidi, and J. Stepanek, “Satellite coverage in urban areas using unmanned airborne vehicles (UAVs),” in *Vehicular Technology Conference(VTC)*, Milan, Italy, May 2004.
- [77] C. E. Palazzi, C. Roseti, M. Luglio, M. Gerla, M. Sanadidi, and J. Stepanek, “Enhancing transport layer capability in HAPS–satellite integrated architecture,” *Wireless Personal Communications*, vol. 32, no. 3-4, pp. 339–356, Feb. 2005.
- [78] X. Li, W. Feng, J. Wang, Y. Chen, N. Ge, and C. Wang, “Enabling 5G on the ocean: A hybrid satellite-UAV-terrestrial network solution,” *IEEE Wireless Communications*, vol. 27, no. 6, pp. 116–121, Sep. 2020.
- [79] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, “Joint UAV hovering altitude and power control for space-air-ground IoT networks,” *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1741–1753, Oct. 2019.
- [80] J. A. Fraire, G. Nies, C. Gerstacker, H. Hermanns, K. Bay, and M. Bisgaard, “Battery-aware contact plan design for leo satellite constellations: The ulloriaq case study,” *IEEE Transactions on Green Communications and Networking*, vol. 4, no. 1, pp. 236–245, Mar. 2020.
- [81] D. Zhou, M. Sheng, X. Wang, C. Xu, R. Liu, and J. Li, “Mission aware contact plan design in resource-limited small satellite networks,” *IEEE Transactions on Communications*, vol. 65, no. 6, pp. 2451–2466, Jun. 2017.
- [82] W. Shi, D. Gao, H. Zhou, Q. Xu, and C. Foh, “Traffic aware inter-layer contact selection for multi-layer satellite terrestrial network,” in *IEEE Global Communications Conference (GLOBECOM)*, Singapore, Singapore, Dec. 2017.
- [83] D. Zhou, M. Sheng, R. Liu, Y. Wang, and J. Li, “Channel-aware mission scheduling in broadband data relay satellite networks,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 1052–1064, May 2018.

- [84] C. Caini and R. Firrincieli, "Application of contact graph routing to LEO satellite DTN communications," in *IEEE International Conference on Communications (ICC)*, Ottawa, Canada, Jun. 2012.
- [85] X. Liu, "Analysis in big data of satellite communication network based on machine learning algorithms," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 7, p. e3861, Jan. 2021.
- [86] A. Gharanjik, M. R. B. Shankar, F. Zimmer, and B. Ottersten, "Centralized rainfall estimation using carrier to noise of satellite communication links," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 5, pp. 1065–1073, May 2018.
- [87] P. V. R. Ferreira, R. Paffenroth, A. M. Wyglinski, T. M. Hackett, S. G. Bilen, R. C. Reinhart, and D. J. Mortensen, "Reinforcement learning for satellite communications: From LEO to deep space operations," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 70–75, May 2019.
- [88] F. Pacheco, E. Exposito, and M. Gineste, "A framework to classify heterogeneous internet traffic with machine learning and deep learning techniques for satellite communications," *Computer Networks*, vol. 173, p. 107213, May 2020.
- [89] Z. Na, Z. Pan, X. Liu, Z. Deng, Z. Gao, and Q. Guo, "Distributed routing strategy based on machine learning for LEO satellite network," *Wireless Communications and Mobile Computing*, vol. 2018, Jan. 2018.
- [90] B. Deng, C. Jiang, H. Yao, S. Guo, and S. Zhao, "The next generation heterogeneous satellite communication networks: Integration of resource management and deep reinforcement learning," *IEEE Wireless Communications*, vol. 27, no. 2, pp. 105–111, Nov. 2020.
- [91] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Distributed multi-agent meta learning for trajectory design in wireless drone networks," *IEEE Journal on Selected Areas in Communications*, to appear, 2021.
- [92] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.
- [93] Y. Sun, D. Xu, D. W. K. Ng, L. Dai, and R. Schober, "Optimal 3D-trajectory design and resource allocation for solar-powered UAV communication systems," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4281–4298, Feb. 2019.
- [94] Z. M. Fadlullah, D. Takaishi, H. Nishiyama, N. Kato, and R. Miura, "A dynamic trajectory control algorithm for improving the communication throughput and delay in UAV-aided networks," *IEEE Network*, vol. 30, no. 1, pp. 100–105, Jan. 2016.
- [95] M. Hua, A. L. Swindlehurst, C. Li, and L. Yang, "UAV-aided backscatter networks: Joint UAV trajectory and protocol design," in *Proc Global Communications Conference (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019.

- [96] M. M. U. Chowdhury, S. J. Maeng, E. Bulut, and I. Guvenc, “3D trajectory optimization in UAV-assisted cellular networks considering antenna radiation pattern and backhaul constraint,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 56, no. 5, pp. 3735–3750, Mar. 2020.
- [97] Q. Wu, Y. Zeng, and R. Zhang, “Joint trajectory and communication design for multi-UAV enabled wireless networks,” *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 2109–2121, Jan. 2018.
- [98] A. R. Hota, S. Garg, and S. Sundaram, “Fragility of the commons under prospect-theoretic risk attitudes,” *Games and Economic Behavior*, vol. 98, pp. 135–164, Jul. 2016.
- [99] A. Sanjab, W. Saad, and T. Başar, “Prospect theory for enhanced cyber-physical security of drone delivery systems: A network interdiction game,” in *Proc. of IEEE International Conference on Communications (ICC)*, Paris, France, Jul. 2017.
- [100] L. Xiao, D. Xu, C. Xie, N. B. Mandayam, and H. V. Poor, “Cloud storage defense against advanced persistent threats: A prospect theoretic study,” *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 3, pp. 534–544, Jan. 2017.
- [101] A. Sanjab and W. Saad, “On bounded rationality in cyber-physical systems security: Game-theoretic analysis with application to smart grid protection,” in *Joint Workshop on Cyber-Physical Security and Resilience in Smart Grids (CPSR-SG)*, Oct 2016, pp. 1–6.
- [102] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, “Counterfactual multi-agent policy gradients,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
- [103] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, “Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.
- [104] P. Battigalli and M. Dufwenberg, “Dynamic psychological games,” *Journal of Economic Theory*, vol. 144, no. 1, pp. 1–35, Jan. 2009.
- [105] M. Li, Y. Hong, C. Zeng, Y. Song, and X. Zhang, “Investigation on the UAV-to-satellite optical communication systems,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2128–2138, Sep. 2018.
- [106] C. E. Palazzi, C. Roseti, M. Luglio, M. Gerla, M. Y. Sanadidi, and J. Stepanek, “Satellite coverage in urban areas using unmanned airborne vehicles (UAVs),” in *Proc. of IEEE Vehicular Technology Conference*, Milan, Italy, May 2004.
- [107] J. Zhao, F. Gao, Q. Wu, S. Jin, Y. Wu, and W. Jia, “Beam tracking for uav mounted satcom on-the-move with massive antenna array,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 363–375, Feb. 2018.

- [108] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, “Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8052–8066, Sep. 2017.
- [109] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez Jr, “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE access*, vol. 1, no. 1, pp. 335–349, May 2013.
- [110] O. Semiari, W. Saad, and M. Bennis, “Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4802–4816, Jul. 2017.
- [111] A. D. Panagopoulos, P. Arapoglou, and P. G. Cottis, “Satellite communications at Ku, Ka, and V bands: Propagation impairments and mitigation techniques,” *IEEE Communications Surveys & Tutorials*, vol. 6, no. 3, pp. 2–14, Oct. 2004.
- [112] “Real-time latency: Rethink possibilities with remote networks,” https://www.telesat.com/sites/default/files/telesat/brochures/telesat_leo_-_real-time_latency_rethink_the_possibilities_with_remote_networks.pdf, Jun. 2018.
- [113] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, “Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks,” *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.
- [114] Y. Zeng, R. Zhang, and T. Lim, “Wireless communications with unmanned aerial vehicles: Opportunities and challenges,” *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, May 2016.
- [115] K. J. Arrow, “General economic equilibrium: purpose, analytic techniques, collective choice,” *The American Economic Review*, vol. 64, no. 3, pp. 253–272, Jun. 1974.
- [116] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.
- [117] P. Camerini, L. Fratta, and F. Maffioli, “On improving relaxation methods by modified gradient techniques,” in *Nondifferentiable Optimization*. Springer, 1975, pp. 26–34.
- [118] A. S. Alfa, B. T. Maharaj, S. Lall, and S. Pal, “Mixed-integer programming based techniques for resource allocation in underlay cognitive radio networks: A survey,” *Journal of Communications and Networks*, vol. 18, no. 5, pp. 744–761, Oct. 2016.
- [119] M. Gapeyenko, V. Petrov, D. Moltchanov, S. Andreev, N. Himayat, and Y. Koucheryavy, “Flexible and reliable UAV-assisted backhaul operation in 5G mmwave cellular networks,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2486–2496, Nov 2018.

- [120] M. Chen, W. Saad, and C. Yin, "Liquid state machine learning for resource and cache management in LTE-U unmanned aerial vehicle (UAV) networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 3, pp. 1504–1517, Mar. 2019.
- [121] T. M. Nguyen, W. Ajib, and C. Assi, "A novel cooperative NOMA for designing UAV-assisted wireless backhaul networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2497–2507, Nov. 2018.
- [122] R. Mochaourab and E. Jorswieck, "Walrasian equilibrium in two-user multiple-input single-output interference channels," in *IEEE International Conference on Communications Workshops (ICC)*, Kyoto, Japan, Jun. 2011.
- [123] X. Duan, C. Zhao, S. He, P. Cheng, and J. Zhang, "Distributed algorithms to compute walrasian equilibrium in mobile crowdsensing," *IEEE Trans. Industrial Electronics*, vol. 64, no. 5, pp. 4048–4057, Dec. 2017.
- [124] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, Mar. 2019.
- [125] D. B. West, *Introduction to graph theory*. Prentice hall Upper Saddle River, 2001, vol. 2.
- [126] G. Maral, M. Bousquet, and Z. Sun, *Satellite communications systems: systems, techniques and technology*. John Wiley & Sons, 2020.
- [127] M. Sheng, Y. Wang, J. Li, R. Liu, D. Zhou, and L. He, "Toward a flexible and reconfigurable broadband satellite network: Resource management architecture and strategies," *IEEE Wireless Communications*, vol. 24, no. 4, pp. 127–133, Aug. 2017.
- [128] O. Semiari, W. Saad, and M. Bennis, "Joint millimeter wave and microwave resources allocation in cellular networks with dual-mode base stations," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4802–4816, Jul. 2017.
- [129] A. Golkar and I. L. i Cruz, "The federated satellite systems paradigm: Concept and business case evaluation," *Acta Astronautica*, vol. 111, pp. 230–248, Jun. 2015.
- [130] S. Tani, K. Motoyoshi, H. Sano, A. Okamura, H. Nishiyama, and N. Kato, "An adaptive beam control technique for diversity gain maximization in leo satellite to ground transmissions," in *IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.
- [131] M. Thurai, E. Deguchi, K. Okamoto, and E. Salonen, "Rain height variability in the tropics," *IEE Proceedings-Microwaves, Antennas and Propagation*, vol. 152, no. 1, pp. 17–23, Feb. 2005.
- [132] P. Series, "Propagation data and prediction methods required for the design of earth-space telecommunication systems," *Recommendation ITU-R*, pp. 618–12, 2015.

- [133] S. Tani, K. Motoyoshi, H. Sano, A. Okamura, H. Nishiyama, and N. Kato, "An adaptive beam control technique for diversity gain maximization in leo satellite to ground transmissions," in *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–5.
- [134] J. R. Marden and J. S. Shamma, "Game theory and control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 105–134, May 2018.
- [135] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.
- [136] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, Jan. 2021.
- [137] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Denver, USA, Dec. 2000.
- [138] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls *et al.*, "Value-decomposition networks for cooperative multi-agent learning," *arXiv preprint arXiv:1706.05296*, 2017.
- [139] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. of International Conference on Machine Learning (ICML)*, Sydney, Australia, Aug. 2017.
- [140] L. C. Baird III, "Advantage updating," WRIGHT LAB WRIGHT-PATTERSON AFB OH, Tech. Rep., 1993.
- [141] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 4, pp. 3039–3071, Fourthquarter 2019.
- [142] F. Jiang and A. L. Swindlehurst, "Optimization of UAV heading for the ground-to-air up-link," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 5, pp. 993–1005, Jan. 2012.
- [143] M. Mozaffari, A. Taleb Zadeh Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3D wireless cellular network," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 357–372, Jan. 2019.
- [144] 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 on Communications, Special Issue on Age of Information in Real-time Systems and Networks*, vol. 39, no. 5, pp. 1250–1265, May 2021.

- [145] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, "Meta-reinforcement learning for trajectory design in wireless UAV networks," in *Proc Global Communications Conference (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.
- [146] R. Houthoofd, Y. Chen, P. Isola, B. Stadie, F. Wolski, O. J. Ho, and P. Abbeel, "Evolved policy gradients," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, Dec. 2018, pp. 5400–5409.
- [147] S. Ritter, J. X. Wang, Z. Kurth-Nelson, S. M. Jayakumar, C. Blundell, R. Pascanu, and M. Botvinick, "Been there, done that: Meta-learning with episodic recall," in *International Conference on Machine Learning*, Jul. 2018, pp. 4354–4363.
- [148] Z. Xu, H. P. van Hasselt, and D. Silver, "Meta-gradient reinforcement learning," in *Proc. of Advances in Neural Information Processing Systems (NIPS)*, Montreal, Canada, Dec. 2018.
- [149] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in OFDMA systems with large numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3292–3304, Jul. 2012.
- [150] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, Jul. 1966.
- [151] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Apr. 2012.
- [152] Q. Zhang, W. Saad, M. Bennis, X. Lu, M. Debbah, and W. Zuo, "Predictive deployment of uav base stations in wireless networks: Machine learning meets contract theory," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 637–652, Jan. 2021.
- [153] J. Vanschoren, "Meta-learning: A survey," *arXiv preprint arXiv:1810.03548*, 2018.
- [154] S. Park, O. Simeone, and J. Kang, "Meta-learning to communicate: Fast end-to-end training for fading channels," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 5075–5079.
- [155] H. Mao, H. Lu, Y. Lu, and D. Zhu, "Roemnet: Robust meta learning based channel estimation in OFDM systems," in *Proc. of International Conference on Communications (ICC)*, Shanghai, China, May 2019.
- [156] H. Wu, Z. Zhang, C. Jiao, C. Li, and T. Q. S. Quek, "Learn to sense: A meta-learning-based sensing and fusion framework for wireless sensor networks," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8215–8227, May 2019.
- [157] N. Abuzainab and W. Saad, "Misinformation control in the internet of battlefield things: A multiclass mean-field game," in *IEEE Global Communications Conference (GLOBECOM)*, Abu Dhabi, UAE, Dec. 2018.

- [158] Y. Yang and R. Kravets, "Achieving delay guarantees in ad hoc networks by adapting ieee 802.11 contention windows," in *Proceedings of IEEE Infocom*, vol. 2006, Oct 2006, p. 18.
- [159] A. M. Colman, "Cooperation, psychological game theory, and limitations of rationality in social interaction," *Behavioral and brain sciences*, vol. 26, no. 2, pp. 139–153, Apr. 2003.
- [160] E. Alpaydin, *Introduction to machine learning*. MIT press, Mar. 2009.
- [161] J. Jaffray, "Bayesian updating and belief functions," *IEEE transactions on systems, man, and cybernetics*, vol. 22, no. 5, pp. 1144–1152, Sep. 1992.
- [162] E. Kalai and E. Lehrer, "Rational learning leads to nash equilibrium," *Econometrica: Journal of the Econometric Society*, vol. 61, no. 5, pp. 1019–1045, Sep. 1993.
- [163] D. Fudenberg and D. M. Kreps, "Learning in extensive-form games i. self-confirming equilibria," *Games and Economic Behavior*, vol. 8, no. 1, pp. 20–55, Jan. 1995.
- [164] N. Suri, M. Tortonesi, J. Michaelis, P. Budulas, G. Benincasa, S. Russell, C. Stefanelli, and R. Winkler, "Analyzing the applicability of internet of things to the battlefield environment," in *Proc. of Military Communications and Information Systems (ICMCIS)*, Brussels, Belgium, May 2016.
- [165] T. Park, N. Abuzainab, and W. Saad, "Learning how to communicate in the internet of things: Finite resources and heterogeneity," *IEEE Access, Special Issue on Optimization for Emerging Wireless Networks: IoT, 5G and Smart Grid Communication Networks*, vol. 4, Nov. 2016.
- [166] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 120–128, Feb. 2017.
- [167] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Machine learning for wireless networks with artificial intelligence: A tutorial on neural networks," *IEEE Communications Surveys Tutorials*, vol. 21, no. 4, pp. 3039–3071, Fourthquarter 2019.
- [168] P. Battigalli, M. Dufwenberg, and A. Smith, "Frustration and anger in games," *CESifo Working Paper Series*, Aug. 2015.
- [169] G. Rossi, A. Tcheukam, and H. Tembine, "How much does users' psychology matter in engineering mean-field-type games," *arXiv preprint arXiv:1702.05355*, 2017.
- [170] B. Powers, M. Smyrnakis, and H. Tembine, "Empathy in bimatrix games," *arXiv available online arXiv:1708.01910*, Aug. 2017.
- [171] M. K. Simon and M. S. Alouini, *Digital communication over fading channels*, 2005, vol. 95.

- [172] R. K. Wood, “Deterministic network interdiction,” *Mathematical and Computer Modelling*, vol. 17, no. 2, pp. 1–18, 1993.
- [173] “LTE System Toolbox,” <https://www.mathworks.com/help/lte/>.
- [174] W. Cook, A. Gerards, A. Schrijver, and É. Tardos, “Sensitivity theorems in integer linear programming,” *Mathematical Programming*, vol. 34, no. 3, pp. 251–264, Apr. 1986.
- [175] N. N. Schraudolph, “Local gain adaptation in stochastic gradient descent,” in *Proc. of International Conference on Artificial Neural Networks.*, Edinburgh, UK, Sep. 1999.