EFFICIENT AND SECURE RESOURCE ALLOCATION IN MOBILE EDGE COMPUTING ENABLED WIRELESS NETWORKS

by

Qun Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Electrical Engineering

Approved:

Rose Qingyang Hu, Ph.D. Major Professor	Don Cripps, Ph.D. Committee Member
Chris Winstead, Ph.D. Committee Member	Zhen Zhang, Ph.D. Committee Member
Haitao Wang, Ph.D. Committee Member	D. Richard Cutler, Ph.D. Interim Vice Provost of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

Efficient and Secure Resource Allocation in Mobile Edge Computing Enabled Wireless Networks

by

Qun Wang, Doctor of Philosophy

Utah State University, 2022

Major Professor: Rose Qingyang Hu, Ph.D.
Department: Electrical and Computer Engineering

Next-generation wireless communications aim to provide highly reliable communication services with low latency to critical applications such as the Industrial Internet of Things (IoT), autonomous driving, virtual reality, and augmented reality. However, explosive data generated by massive and densely connected devices will overwhelm the access and backbone networks if using centralized cloud computing services to process. Mobile Edge Computing (MEC) framework moves the computing services close to the user end and provides fast response data processing services with low latency. However, deploying multiple servers with decentralized distribution will raise energy consumption and transmission security concerns for MEC networks. Numerous techniques can be employed to improve energy efficiency (EE) and physical layer security (PLS) for data offloading and local computing in MEC networks. Non-orthogonal multiple access (NOMA) allows various users to transmit simultaneously. Multiple input multiple outputs (MIMO) with beamforming design enable the transceiver to achieve a higher channel capacity through multiple paths and space diversity. Intelligent reflecting surface (IRS) can supplement the severe pathloss and the lack of line of sight (LOS) link in the mmWave communication system, thus increasing the achievable rate on the server-side with the same transmit power on the user side. Combining those techniques

can improve the EE and PLS performance. Exploiting the machine learning algorithms to capture the trends of data traffic loads enables the system to coordinate its computing resources proactively, which can further improve the system EE. In this paper, we will explore all these techniques for improving the EE and PLS of the MEC networks. We have built several system models and formed corresponding optimization problems. By applying convex optimization theories, we proposed different algorithms to find resource allocation solutions. Our simulation results verified the effectiveness of the proposed approaches.

(161 pages)

PUBLIC ABSTRACT

Efficient and Secure Resource Allocation in Mobile Edge Computing Enabled Wireless Networks

Qun Wang

To support emerging applications such as autonomous vehicles and smart homes and to build an intelligent society, the next-generation internet of things (IoT) is calling for up to 50 billion devices connected world wide. Massive devices connection, explosive data circulation, and colossal data processing demand are driving both the industry and academia to explore new solutions.

Uploading this vast amount of data to the cloud center for processing will significantly increase the load on backbone networks and cause relatively long latency to time-sensitive applications. A practical solution is to deploy the computing resource closer to end-users to process the distributed data. Hence, Mobile Edge Computing (MEC) emerged as a promising solution to providing high-speed data processing service with low latency.

However, the implementation of MEC networks is handicapped by various challenges. For one thing, to serve massive IoT devices, dense deployment of edge servers will consume much more energy. For another, uploading sensitive user data through a wireless link introduces potential risks, especially for those size-limited IoT devices that cannot implement complicated encryption techniques. This dissertation investigates problems related to Energy Efficiency (EE) and Physical Layer Security (PLS) in MEC-enabled IoT networks and how Non-Orthogonal Multiple Access (NOMA), prediction-based server coordination, and Intelligent Reflecting Surface (IRS) can be used to mitigate them.

Employing a new spectrum access method can help achieve greater speed with less power consumption, therefore increasing system EE. We first investigated NOMA-assisted MEC networks and verified that the EE performance could be significantly improved. Idle

servers can consume unnecessary power. Proactive server coordination can help relieve the tension of increased energy consumption in MEC systems. Our next step was to employ advanced machine learning algorithms to predict data workload at the server end and adaptively adjust the system configuration over time, thus reducing the accumulated system cost. We then introduced the PLS to our system and investigated the long-term secure EE performance of the MEC-enabled IoT network with NOMA assistance. It has shown that NOMA can improve both EE and PLS for the network. Finally, we switch from the single antenna scenario to a multiple-input single-output (MISO) system to exploit space diversity and beamforming techniques in mmWave communication. IRS can be used simultaneously to help relieve the pathloss and reconfigure multi-path links. In the final part, we first investigated the secure EE performance of IRS-assisted MISO networks and introduced a friendly jammer to block the eavesdroppers and improve the PLS rate. We then combined the IRS with the NOMA in the MEC network and showed that the IRS can further enhance the system EE.

To my family.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude and appreciation to my Ph.D. advisor, Rose Qingyang Hu, for her countless time and effort in helping me during my entire Ph.D. program. She is a great advisor and also an inspiring role model that gives me tremendous instructions and motivations to be a professional and productive researcher. I have learned greatly from her solid technical knowledge, great sense for research direction, excellent writing and presentation skills.

I also want to express my sincere appreciations to my committee members, Prof. Don Cripps, Prof. Chris Winstead, Prof. Zhen Zhang, and Prof. Haitao Wang, for their valuable comments and suggestions that help me immensely with my study, proposal and dissertation. Additionally, I am very grateful to Prof. Todd Moon and Prof. Jacob Gunther for their classes and helpful discussions.

Next, I would like to thank Prof.Yi Qian, Dr. Fuhui Zhou, Dr. Han Hu, Dr. Le Thanh Tan. I have learned a lot from them while collaborating and discussing with them. For my colleagues and friends, I want to thank Dr.Haijian Sun, Dr. Zekun Zhang, Dr. Xuan Xie, Dr. Kuan Huang, Shan He, Xiang Ma. They helped me with multiple research discussions and other aspects.

Lastly, my father Xiaoliang Wang and mother Yingli Weng, brother Li Wang and sister Shu Pan, my nephew Zimo Wang, and grandparents Wenxuan He and Junlan Liu, gave me the most support through these Ph.D. days and provided me endless encouragement and love.

Qun Wang

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

IoT	Internet of Things
eMBB	Enhanced Mobile Broadband
mMTC	Massive Machine Type Communication
URLLC	Ultra Reliability and Low Latency Communication
MEC	Mobile Edge Computing
NOMA	Non-Orthogonal Multiple Access
FDMA	Frequency Division Multiple Access
IRS	Intelligent Reflecting Surfaces
MISO	Multiple Input Single Output
UE	User Equipment
EE	Energy Efficiency
PLS	Physical Layer Security
SIC	Successive Interference Cancellation
SCA	Successive Convex Approximation
SDR	Semi-Definite Relaxation
SE	Spectral Efficiency
OMA	Orthogonal Multiple Access
CSI	Channel State Information
mmWave	Millimeter Wave
Eve	Eavesdropper
FJ	Friendly Jammer

CHAPTER 1

INTRODUCTION

The development of 5G communication networks and smart services lead to an explosive increment of Internet of Things (IoT) devices, which will generate massive data for communication and computing. 5G is expected to provide new applications such as autonomous drive, virtual reality, smart homes, industrial IoT, and smart healthcare with high reliability, low latency, fast response services, and allow massive number of devices to connect to the Internet. However, limited wireless bandwidth and low user computation ability have gradually become a bottleneck to realize the use cases such as enhanced mobile broadband (eMBB), Massive Machine Type Communication (mMTC), and Ultra Reliability and Low Latency Communication (URLLC) and to meet the quality of service (QoS) requirements [1] [2]. Furthermore, transferring large amounts of data for communication and computing may expose networks to possible attacks, creating a need for effective security and privacy protection schemes.

In order to ensure ultra-low latency, localization of service processing, and data storage, it is critical to design a new solution that can provide high-speed communication in latency-sensitive scenarios and execute tasks and return results in a timely manner. Motivated by the increasing computational capacity of wireless local devices as well as the ever-increasing privacy and security concerns of sharing data, next-generation wireless communication networks have been encountering a paradigm shift from conventional cloud computing to Mobile Edge Computing (MEC), which largely deploys computing resource to the network edges/fog nodes to meet the needs of applications that demand very high computations and low latency. In wireless networks, edge nodes, such as base stations and edge routers, can be equipped with high computational and storage capabilities. Therefore, MEC enables user equipment (UE) to offload its tasks to nearby edge servers for processing and has the potential to provide location-aware, real-time, and low-cost services to support

emerging computation-intensive applications.

However, complex wireless environment, limited UE power supply, and potential malicious attacks all render it very challenging to design efficient and secure MEC networks. Therefore, our research aims to address two specific issues related to the MEC networks:

1. **Energy Efficiency:** Energy Efficiency (EE) is particularly crucial in the MEC networks. Since most UEs employ batteries as the power source, their limited energy capacities impose restrictions on system stability for long-term operations. Furthermore, the potential to deploy MEC systems in areas without a reliable energy supply calls for new methods of managing energy consumption not only in UEs but also through the entire network. On the other hand, to serve the diverse energy-intensive communication services from massive users, reducing greenhouse gas emissions caused by overall system energy consumption is imperative. As shown in Fig. 1.1, EE is usually defined as the tradeoff between achievable rate and consumed energy. Therefore, to improve the EE, the solution can be either seeking the new techniques that allow the system to achieve a higher rate with the same power consumption or using a lower power to achieve the same communication and computing performance. The hierarchical structure of MEC networks provides multiple opportunities for improving EE, such as advanced wireless offloading, intelligent server coordination, and efficient resource allocation strategies.

2. **Security:** Despite the possible privacy protection provided by MEC networks, the broadcast nature of wireless links and the magnitude and sensitivity of the information maintained by UEs necessitate secure mechanisms when performing task offloading. Specifically, malicious eavesdroppers can access data without being detected. Techniques that safeguard the security of communication and privacy of user information fall into two major categories: traditional cryptographic techniques and physical layer security (PLS).

   PLS has received tremendous attention in recent years. As illustrated in Fig. 1.2, PLS is defined as the difference between the achievable rate at legal receivers and

Fig. 1.1: Illustration of energy efficiency.

eavesdroppers. Security can be achieved by limiting the amount of information leakage to an acceptable level. PLS enable secure communications without extra overhead caused by protecting the security key. However, in PLS, the secrecy rate achieved by the mutual information difference between the legitimate receiver and the eavesdropper can be limited as it depends on the difference between the channel condition from the user to the legitimate receiver and that from the user to the eavesdroppers. More advanced techniques for improving the security rate need to be investigated.

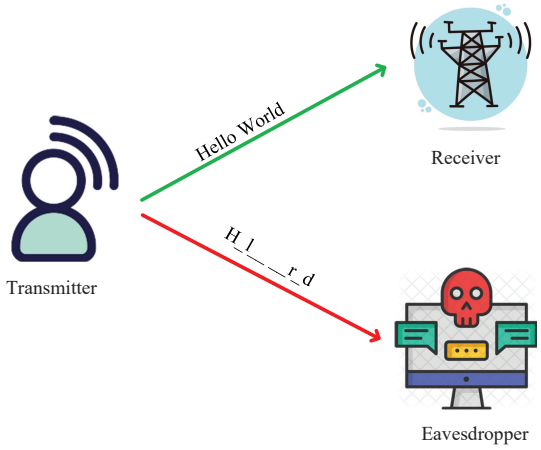

Fig. 1.2: Illustration of physical layer security.

The emergence of new wireless communication techniques provides the MEC networks with powerful tools for improving their EE and security performance. Specifically, Non-Orthogonal Multiple Access (NOMA) can help achieve higher efficiency and security rates. Advanced machine learning algorithms further enable the system to coordinate the edge servers proactively. A new technique known as Intelligent Reflecting Surface (IRS) could potentially improve the EE and PLS performance of the NOMA-enabled MEC network. The purpose of the research described in this work is to investigate techniques for increasing both system EE and PLS in MEC networks.

1.1 NOMA Enabled MEC Networks

In this work, we will first study the NOMA assisted MEC network for EE improvement. By exploiting superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, NOMA significantly changes traditional multiple access mechanisms. As shown in Fig. 1.3, NOMA allows multiple users to share the same radio bandwidth in either power domain or code domain to increase spectral efficiency with a relatively higher receiver complexity [3]. Therefore, applying NOMA to MEC-enabled networks has recently received extensive attention due to its performance gain in both spectrum efficiency and EE [4–7]. Some existing works on NOMA-assisted MEC with eavesdroppers also show that NOMA is an effective way to enhance the PLS [8–10].

Applying NOMA in MEC can improve the computational performance and user connectivity in ultra-dense IoT networks [5–7]. The successive interference cancellation (SIC) order and computation resource allocation have been jointly optimized in [5], which can minimize the maximum task execution latency for IoT devices under the limitation of computational resources. Sun *et al.* [6] proposed the NOMA communication method with the wireless energy supply for the IoT system. Pan *et al.* [7] studied the MEC system, which exploits the NOMA for computational task uploading and results downloading. By optimizing the transmit powers, transmission time allocation, and task offloading partitions, the minimization of total energy consumption was achieved by this work. It was demonstrated that the NOMA method can significantly improve EE compared with the OMA method.

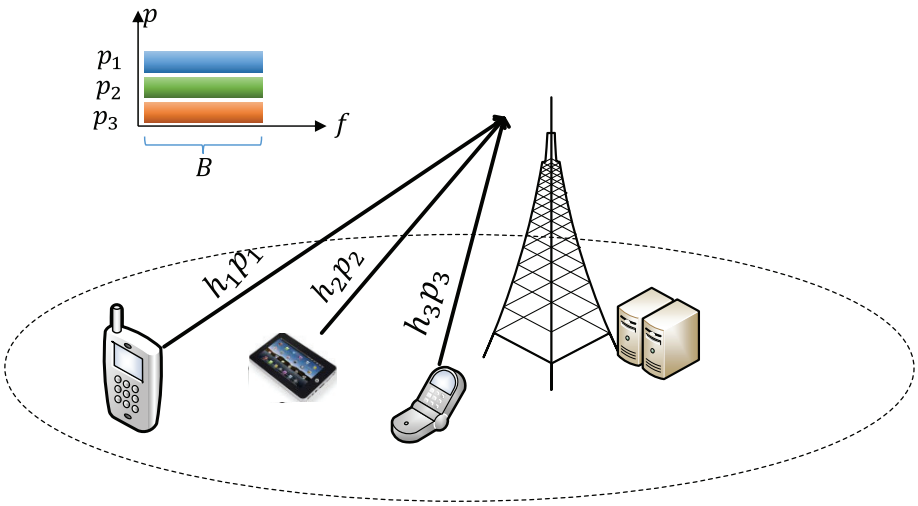

Fig. 1.3: Illustration of NOMA.

The PLS in NOMA-assisted MEC networks has garnered much interest in research [11]. The joint consideration of PLS in the NOMA assisted MEC network was studied in [8–10]. In [8], an iterative algorithm was proposed to maximize the minimum anti-eavesdropping ability in a MEC network with uplink NOMA. The authors in [9] proposed a bisection searching algorithm to minimize the maximum task completion time subject to the worst-case secrecy rate. Instead of only considering the power consumption or computing rate performance above, [10] studied the EE maximization problem for a NOMA enabled MEC network with eavesdroppers.

To achieve higher performance, effective resource allocation should incorporate various QoS metrics. One of the most important metrics is to guarantee fairness among different users. However, most of the existing works on NOMA-assisted MEC networks did not consider the fairness between users. Including fairness utility functions will increase the complexity of resource allocation designs in NOMA-enabled MEC networks, an unaddressed need that calls for effective solutions. Moreover, the existing works on NOMA-assisted MEC with external eavesdroppers typically focus on performance evaluation in scenarios where either channel conditions or required tasks remain constant. Such an assumption makes analysis on computation offloading and resource allocation more tractable. However,

in a dynamic environment, the dynamic behaviors of the workload arrivals and fading channels impact the overall system performance. Thus a system design that focuses on short-term performance may not be as effective from a long-term perspective. To this end, the stochastic task offloading models and resource allocation strategies need to be investigated over long timescales.

In this work, we will look at the EE and PLS performances of MEC networks with NOMA offloading techniques.

1.2 Server Coordination Techniques in MEC Networks With NOMA

Most existing works did not consider the computation capacity or energy consumption of edge servers. However, in order to support the ultra-densely deployed IoT devices and their massive tasks, edge servers' energy consumption and operation cost increase sharply. The edge servers should support an energy-saving mode that activate/deactivate the servers either manually or automatically based on demands. If this is to be done, efficient server coordination is essential. User's behavior and environment conditions are constantly fluctuating, but according to [12], the user's traffic behavior follows certain predictable patterns. For example, network activities activity decreases significantly during the night. Therefore, embedded workload prediction can help the system capture the peak and valley trends of workloads, which the system can then use to coordinate servers in a dynamic way. It is well known that servers consume energy even when they are not actively processing any tasks as many server units consume considerable energy even when idle. By letting idle servers to switch into sleeping mode or turning the sleeping servers back to active status when necessary can greatly increase the networks' energy efficiency. Thus, accurate prediction algorithms and effective server coordination methods are important. It can enable the network to flexibly adjust and allocate server resources, as shown in Fig. 1.4 .

As the number of UE devices grows, the edge computing service must also increase its scalability to guarantee a latency limit and quality threshold. Several works focus on server coordination in MEC networks.

In [13], an algorithm that utilizes virtual machine migration and transmission power

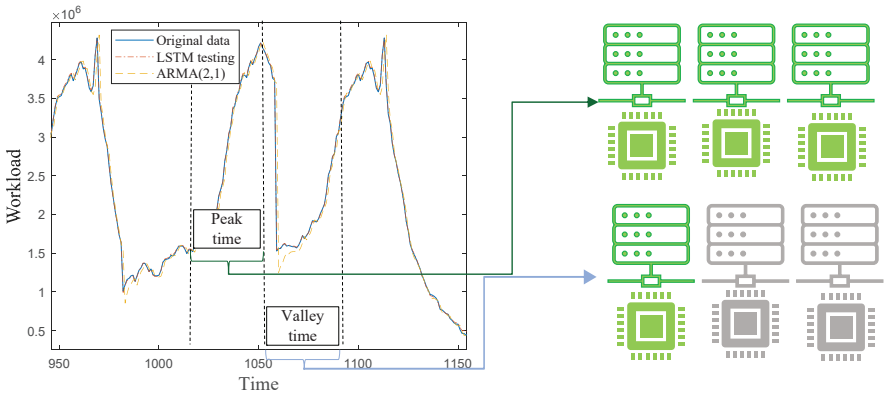

Fig. 1.4: Illustration of server coordination.

control was proposed to coordinate edge nodes and consequently maximize cost-effectiveness. Prediction functions can empower the MEC systems to proactively balance the users' workload and efficiently coordinate their computing resources. In [14], the authors proposed a task execution time prediction algorithm to solve a task offloading optimization problem in MEC networks. In [15], a computation offloading and task migration algorithm based on task prediction was proposed, which joint considered the data size of computation task and the performance features of edge nodes. A computation offloading strategy based on task prediction and task migration for the edge cloud scheduling scheme was used to assist in optimizing the edge computing offloading model. To reduce the energy cost of MEC networks, the authors in [16] formulated the energy-saving problem in MEC networks by dynamically switching on/off edge servers according to the variation of UEs' distribution. A dynamic server switching algorithm and a lightweight UE distribution prediction mechanism were proposed to solve the problem. In [17], a proactive approach to dynamic edge server provisioning for real-time IoT data streaming across edge nodes was proposed that adjusts server provisioning ahead of time based on predictions about the upcoming workload.

Most of the existing works employed a relative sample workload prediction model, which cannot accurately capture the workload trends and may lead to QoS degradation. Moreover, frequently switching edge servers will decrease their lifetime. A more flexible method that encompasses a more comprehensive time scale is needed. Finally, the existing

works did not combine server coordination methods with NOMA transmission techniques, which can further increase the system EE performance. This work aims to explore accurate prediction models and effective server coordination algorithms in improving overall system EE.

1.3 IRS Enabled MEC Networks

IRS has attracted significant attentions from the research community due to its potential of simultaneously improving EE and achieving secure communications [18]. As shown in Fig. 1.5, IRS consists of a large number of low-cost passive reflecting elements with adjustable phase shifts. By properly controlling the phase shifts of the IRS's elements, their reflected signals can be combined with those from other paths coherently to enhance the link achievable rate at the receiver and to decrease it at the eavesdropper [19]. Moreover, since IRS does not employ any transmit radio frequency (RF) chains, energy consumption only comes from reflective elements phase adjustment and is usually very low [20]. Thus, IRS is deemed a promising technology to increase EE of wireless communica tion networks and to improve system security [21]. The IRS-assisted multiple input single output (MISO) secure network has attracted increasingly elevated attention. The beamforming and phase shift matrix design schemes for different objectives were proposed in [22–28].

To determine the secrecy rate gain brought by IRS, in [22], Yu *et al.* considered an IRS-assisted secure MISO wireless system. To maximize the secrecy rate, both the beamformer and the IRS phase shift matrix were jointly optimized based on the block coordinate descent (BCD) and minimization maximization techniques. By combining the artificial noise (AN) technique, in [23], Xu *et al.* studied resource allocation design to maximize the system sum secrecy rate. By jointly optimized the phase shift matrix, the beamforming vectors, and the AN covariance matrix, the authors developed an efficient suboptimal algorithm based on alternating optimization, successive convex approximation, semidefinite programming relaxation (SDR), and manifold optimization. In [24], by jointly optimizing the beamformers at the BS and reflecting coefficients at the IRS, the authors formulated a minimum-secrecy-rate maximization problem under various practical constraints that captured the scenarios of

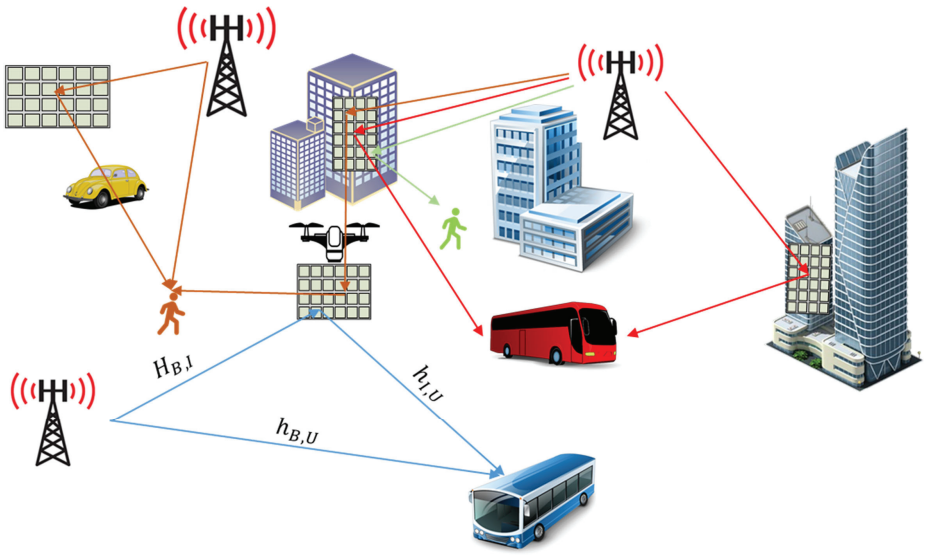

Fig. 1.5: Illustration of IRS.

both continuous and discrete reflecting coefficients of the reflecting elements. Since IRS can not only help increase the secrecy rate but also save more energy for the network, the joint optimization of rate and power was also studied. By considering the power consumption, in [25], the authors focused on maximizing the system secrecy rate subject to the source transmission power constraint and the unit modulus constraints imposed on phase shifts at the IRS. Furthermore, in [26], the authors proposed a power-efficient scheme to optimize the secure transmit power allocation and the surface reflecting phase shift to minimize the transmit power subject to the secrecy rate constraint. In [27], the authors proposed different methods to minimize the system's energy consumption in cases of rank-one and full-rank access point (AP)-IRS links. In [28], secure wireless information and power transfer with the IRS was proposed for a MISO system. Under the secrecy rate and the reflecting phase shifts of IRS constraints, the secure transmit beamforming at the access point and phase shifts at IRS were jointly optimized to maximize the harvested power of the energy harvesting receiver.

Although beamforming design problems in IRS-enabled secure communication systems have been investigated, few studies have been conducted for beamforming, friendly jamming,

and phase shift matrix design in IRS assisted wireless MISO networks. Moreover, there are no investigations that have studied energy-efficient design in secure IRS-assisted MISO networks. Motivated by the above-mentioned facts, an IRS assisted MISO network with cooperative jamming needs to be studied to pave the path for applying IRS into the MEC system.

In this work, we will first investigate the potential benefits of exploiting IRS for transmission with respect to efficiency and security. Then, the IRS-assisted MEC networks will be designed, and the corresponding performance amelioration will be illustrated.

1.4 Dissertation Outline

The focus of this dissertation is the improvement of EE and PLS with regards to the various challenges inherent to MEC networks mentioned above. We first investigate the EE performance of the NOMA-enabled MEC network. Based on this framework, a hierarchical MEC network is proposed and MEC server coordination across different layers is examined to further increase the system EE. The PLS and EE are then investigated for NOMA-enabled MEC networks. To explore the potential advancement of EE and secure rate when applying IRS, we first study the IRS-assisted wireless communication system, then launch the research for apply IRS into MEC based on this achieved research.

In Chapter 2, we first investigate how NOMA can help MEC achieve a higher EE. In order to improve the fairness and resource efficiency among IoT users, we consider a static setting and form a α fairness utility-based resource allocation optimization problem for ultra-dense MEC-enabled IoT networks with NOMA. An iterative algorithm based on successive convex approximation techniques is proposed to solve those challenging non-convex problems under three fairness use cases.

Chapter 3 extends the model in Chapter 2 and develops a dynamic long-term energy and computation optimization paradigm in a hierarchical MEC network with server coordination functionality. The tasks collected at local IoT devices can be computed at edge facilities. Both NOMA and frequency-division multiple access (FDMA) are used for computation offloading. The system model considers both long-term and short-term system

behaviors and makes the best decisions for energy consumption and computation efficiency. The LSTM network is applied to predict the long-term workload, based on which the number of active process units in the edge layer is optimized. In the short-term model, a resource optimization problem is formulated. Due to the dynamic arrival workload and nonconvex features of the problem, the Lyapunov optimization approach and successive convex approximation for the low-complexity method are applied to solve this problem.

In addition to EE, security is also a critical issues for MEC networks. With stochastic task arrivals, time-varying dynamic environment, and passive existing attackers, it is very challenging to offload computation tasks securely and efficiently. In Chapter 4, we study the task offloading and resource allocation problem in a NOMA assisted MEC network with security and energy efficiency considerations. To tackle the problem, a dynamic secure task offloading and resource allocation algorithm is proposed based on Lyapunov optimization theory. A stochastic non-convex problem is formulated to jointly optimize the local-CPU frequency and transmit power, aiming at maximizing the network energy efficiency, which is defined as the ratio of the long-term average secure rate to the long-term average power consumption of all users. The formulated problem is decomposed into the deterministic sub-problems in each time slot. The optimal local CPU-cycle and the transmit power of each user can be given in the closed-from.

In Chapter 5, we consider that the EE achieved by using PLS can be limited by the channel conditions. In order to tackle this problem, an intelligent reflecting surface (IRS) assisted multiple input single output (MISO) network with independent cooperative jamming is studied. The EE is maximized by jointly designing the transmit and jamming beamforming and IRS phase-shift matrix under both the perfect channel state information (CSI) and the imperfect CSI. In order to tackle the challenging non-convex fractional problems, an algorithm based on semidefinite programming (SDP) relaxation is proposed for solving energy efficiency maximization problem under the perfect CSI case while an alternate optimization algorithm based on S -procedure is used for solving the problem under the imperfect CSI case.

In Chapter 6, based on the works in previous chapters, an IRS assisted MEC network with NOMA is studied. The EE is maximized by jointly optimizing the offloading power, local computing frequency, receiving beamforming, and IRS phase-shift matrix. The problem is challenging to solve due to the non-convex fractional objective functions and the coupling among the variables. A semidefinite programming relaxation (SDR) based alternating algorithm is developed.

Chapter 7 concludes this dissertation and proposes future directions.

CHAPTER 2

Fair Resource Allocation in an MEC-Enabled Ultra-Dense IoT Network with NOMA

2.1 Introduction

In this chapter, we consider the resource allocation for MEC networks in a single time slot. We aim to investigate the benefits brought by NOMA to the system's EE. Considering that the channel qualities of different users can be varied, we also incorporate the fairness function to guarantee the quality of service for every user.

Communication and computation resources allocation optimization for different objects has been studied in MEC networks with orthogonal multiple access (OMA) [29–33]. Specifically, in [29], by combining local computing and data offloading, a weighted sum user computation efficiency optimization method based on time division multiple access (TDMA) was proposed. In [30], the maximal delay of the mobile devices was minimized by jointly optimizing sub-carrier and power allocation in MEC networks with orthogonal frequency division multiple access (OFDMA). The authors of [31] proposed a single-leader-multi-user Stackelberg game model to optimize the energy efficiency and computation capacity of the mobile users and the edge cloud. In OFDMA-enabled cloud radio access network (C-RAN) with an integrated MEC server, the joint sub-carrier power allocation and tasks partition problem were studied to minimize the user delay in [32]. The researchers in [33] took a critical look at the resource allocation for TDMA and OFDMA based multiuser MEC systems in order to minimize the weighted sum of mobile energy consumption. They demonstrated that the power allocation has a threshold-based structure with respect to a derived offloading priority. All these works showed that the combination of offloading and local computation outperforms models that only considers the offloading process.

As the need to enhance user connectivity and provide more users with MEC services has grown, NOMA has received great research attention lately. It is envisioned that the

application of MEC and NOMA into ultra-dense IoT networks can greatly improve the computation performance of users and enhance user connectivity. Recently, MEC-enabled IoT networks with NOMA have been studied in [34–36]. In [34], to maximize the harvesting power, a NOMA cognitive radio network with simultaneous wireless information and power transfer was considered. In [35], the weighted sum of the energy consumption of all users was minimized for a multi-user partial offloading MEC system with NOMA under the computation latency constraints. It was shown that NOMA method can significantly improve energy efficiency compared with OMA. In [36], the authors analyzed the performance of the spectral and energy efficiency of a multiple-user wireless communication system with the fairness consideration.

However, the works in [34], [35] did not consider the fairness as a performance target among users, which may result in unfairness, especially when there exist massive IoT devices all with very limited computation capability. Although the authors in [36] considered the fairness among users, the computation efficiency, an important metric in the IoT network, was not considered in this work. In order to achieve the optimal system efficiency in an IoT network, efficient resource allocation schemes are needed. An MEC-enabled ultra-dense IoT network with NOMA can effectively balance the computation efficiency and energy efficiency.

Towards that end, in this chapter, in order to achieve a better trade-off between computation efficiency and energy efficiency, we propose a new method for the resource allocation in an MEC-enabled ultra-dense IoT network with NOMA. Moreover, in order to address fairness among massive IoT users, we introduce fairness index into the utility function of the proposed scheme. The formulated problem is solved by using the successive convex approximation (SCA) method and successive convex approximation for low complexity (SCALE) method. The simulation results demonstrate that the proposed scheme achieves desirable performance by comparing with two other schemes.

2.2 System Model

An MEC-enabled ultra-dense IoT network as shown in Fig. 2.1 is considered. The

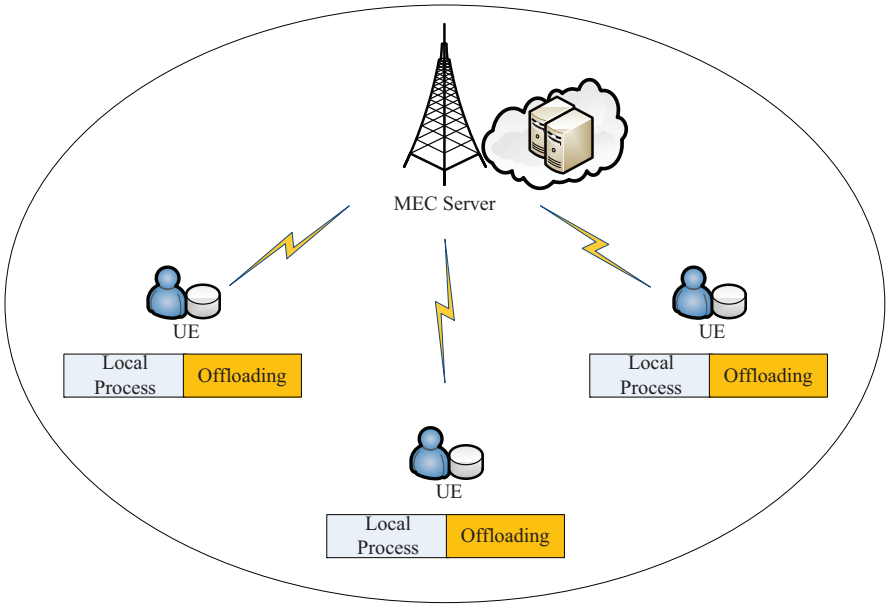

Fig. 2.1: System model of MEC-enabled IoT network with NOMA.

system consists of N IoT devices denoted as user equipment (UE) that need to execute computation-intensive yet delay-sensitive tasks and one MEC server that can provide MEC service for those UEs. Partial computation offloading mode is supported, in which the computation task can be partitioned into two parts, one for local computing and one for offloading to the MEC server for computing. Moreover, NOMA is applied so that multiple devices can offload their tasks simultaneously by using the same physical radio resource. Each IoT device can perform local computing and computation task offloading at the same time since the offloading unit and local computing unit are separated [37].

2.2.1 Data Offloading

In data offloading, users offload their partial computation tasks to the MEC server. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the set of UEs. Let g_n and p_n respectively represent the channel gain and the transmission power between the MEC server and UE n. At the beginning of each transmission frame, the MEC server ranks the UEs by their channel quality, i.e., $g_1^2 \leq g_2^2 \leq \cdots \leq g_N^2$. All the N users can send the offloading data to the MEC simultaneously on the same radio resource. Successive interference cancellation (SIC) technique is applied

at the receiving side, which is the MEC server in this case, to decode the signal for each user successfully [38]. Specifically, the MEC server starts the decoding process with the user that has the best received signal power by treating the signals from all other users as interference. The decoded signal is removed from the composite received signal and then decoding proceeds to the next best received signal. The process repeats untill all the user signals are decoded. To summarize, in order to decode the nth user's signal, the signals from user i, $1 \leq i \leq n-1$, are treated as interference and the signals from user m, $n \leq m \leq N$, are all removed from the composite received signal. Thus, the offloading rate for the nth user can be expressed as

$$r_n^{\text{off}} = B \log_2(1 + \frac{p_n g_n^2}{\sum_{i=1}^{n-1} p_i g_i^2 + \sigma^2}), \tag{2.1}$$

where σ^2 is the power of the noise and B is the bandwidth shared by N NOMA users. The corresponding power consumption for UE n under the offloading mode can be expressed as

$$p_n^{\text{off}} = \zeta p_n + p_r, \tag{2.2}$$

where ζ denotes the amplifier coefficient, and the first part p_n denotes the information transmission power consumption, and p_r denotes the constant circuit power consumed for signal processing and it is assumed to be the same for all UEs [39].

2.2.2 Local Computing

Let C_n be the number of computation cycles required to process one bit of data for UE n locally. Each UE can compute the data throughout the entire transmission duration. Let f_n denote the computing speed of the processor (cycles per second). Therefore, the local computing rate of the nth UE can be expressed as

$$r_n^{local} = \frac{f_n}{C_n}. \tag{2.3}$$

The power consumption of local computing is modeled as a function of processor speed f_n. It can be given as

$$p_n^{local} = \epsilon f_n^3, \tag{2.4}$$

where ϵ is the effective capacitance coefficient of the processor's chip [40].

2.2.3 Utility Function

Combining data offloading and local computing, the total computation rate R_n for UE n can be expressed as $R_n = r_n^{off} + r_n^{local}$. The total power consumption P_n of UE n is $P_n = p_n^{off} + p_n^{local}$. Moreover, in order to consider the computation rate fairness among users, the following utility $U_\alpha(R_n)$ is defined [41], where α is the fairness index.

$$U_\alpha(R_n) = \begin{cases} \dfrac{R_n^{1-\alpha}}{1-\alpha} & \text{if } \alpha \geq 0,\ \alpha \neq 1, \\ \ln(R_n) & \alpha = 1. \end{cases} \tag{2.5}$$

Accordingly, the fairness degree can range from zero to infinity. There are three special cases corresponding to three fairness degrees, namely $\alpha = 0$, 1, and ∞. When $\alpha = 0$, the utility function is the sum computation rate for all UEs; when $\alpha = 1$, the utility function is the sum logarithmic function of UE rates, which normally provides proportional fairness; when $\alpha = \infty$, the utility function is the minimum rate among all UEs, which corresponds to the max-min fairness.

2.3 Fair Resource Allocation

In this section, with respect to the value of fairness index α, the optimal trade-off between the number of computation bits and energy efficiency is studied. The optimization problem aims to minimize the total power consumption as well as maximize the achievable data rate utility. We exploit the weighted sum to tackle this multi-objective problem [36].

Thus, the problem is formulated as

$$\mathbf{P}_{2.1}: \quad \max_{f_n, p_n} \ \Phi_{SE} \sum_{n\in\mathcal{N}} U_\alpha(R_n) - \Phi_{EE} \sum_{n\in\mathcal{N}} P_n \tag{2.6a}$$

$$s.t. \ \ C1: \ p_n \geq 0, \ \ \forall n \in \mathcal{N}, \tag{2.6b}$$

$$C2: \ R_n \geq R_n^{th}, \ \ \forall n \in \mathcal{N}, \tag{2.6c}$$

$$C3: \ P_n \leq P_n^{th}, \ \ \forall n \in \mathcal{N}, \tag{2.6d}$$

$$C4: \ f_n^{min} \leq f_n \leq f_n^{max}, \ \ \forall n \in \mathcal{N}. \tag{2.6e}$$

$\mathbf{P}_{2.1}$ is a resource allocation problem that optimizes the offloading power p_n, local computing chip frequency f_n. Φ_{SE} and Φ_{EE} are the weighting factors that can be used to prioritize different computation service requirements of UEs. $C1$ states that the transmit power levels of UEs are greater than 0. In C_2, R_n^{th} denotes the minimum computing data rate of UE n. P_n^{th} in C_3 is the total power available for UE n. C_4 defines the minimum and maximum computation capacity of each UE. $\mathbf{P}_{2.1}$ is extremely challenging to solve due to the complex objective function and the non-convex constraints. In the following sections, we will investigate three different fairness cases, i.e., $\alpha = 1, \infty, 0$.

2.3.1 Proportional Fairness $\alpha = 1$

When $\alpha = 1$, the system utility is the sum logarithmic function of UE's computation rate. In this case, the proportional fairness can be achieved. The original problem $\mathbf{P}_{2.1}$ under this case can be expressed as

$$\begin{aligned}\mathbf{P}_{2.2}: \quad & \max_{f_n, p_n} \ \Phi_{SE} \sum_{n\in\mathcal{N}} \ln(B\log_2(1+\frac{p_n g_n^2}{\sum_{i=1}^{n-1} p_i g_i^2 + \sigma^2}) + \frac{f_n}{C_n}) - \Phi_{EE} \sum_{n\in\mathcal{N}} (\zeta p_n + p_r + \epsilon f_n^3) \\ & s.t. \ \ C1 - C4.\end{aligned} \tag{2.7}$$

Theorem 2.1: $\mathbf{P}_{2.2}$ is a non-convex optimization problem.

Proof: The auxiliary variables a_n are introduced to meet the following constraints:

$$\exp(a_n) \leq B\log_2(1+\frac{p_n g_n^2}{\sum_{i=1}^{n-1} p_i g_i^2+\sigma^2})+\frac{f_n}{C_n}. \tag{2.8}$$

Auxiliary variables ρ_n are introduced to meet $\rho_n = \ln(p_n)$. Thus, eq. (2.8) is equivalent to

$$\ln(2^{\frac{\exp(a_n)-\frac{f_n}{C_n}}{B}}-1)+\ln(\sum_{i=1}^{n-1}\frac{g_i^2}{g_n^2}e^{\rho_i-\rho_n}+\frac{\sigma^2}{g_n^2}e^{-\rho_n})\leq 0. \tag{2.9}$$

Let $\mathbf{a}=[a_1,\cdots,a_N]^T$, $\mathbf{P}_{2.2}$ can be transformed into

$$\begin{aligned}
\mathbf{P}_{2.3}: \max_{f_n,\rho_n,a_n}\ & \Phi_{SE}\sum_{n\in\mathcal{N}} a_n - \Phi_{EE}\sum_{n\in\mathcal{N}}(\zeta\exp(\rho_n)+p_r+\epsilon f_n^3) \\
s.t.\ \ & a_n \geq \ln(R_n^{th}),\ \ \forall n\in\mathcal{N}, && \text{(2.10a)}\\
& \zeta\exp(\rho_n)+p_r+\epsilon f_n^3 \leq P_n^{th},\ \forall n\in\mathcal{N}, && \text{(2.10b)}\\
& f_n^{min}\leq f_n\leq f_n^{max},\ \ \forall n\in\mathcal{N}, && \text{(2.10c)}\\
& \ln(2^{\frac{\exp(a_n)-\frac{f_n}{C_n}}{B}}-1)+\ln(\sum_{i=1}^{n-1}\frac{g_i^2}{g_n^2}e^{\rho_i-\rho_n}+\frac{\sigma^2}{g_n^2}e^{-\rho_n})\leq 0,\ \ \forall n\in\mathcal{N}. && \text{(2.10d)}
\end{aligned}$$

The objective function is jointly concave with respect to a_n and ρ_n due to the subtraction of a linear term and convex term. However, the first part of the last constraint (2.10d) is non-convex. Thus, the problem is still non-convex. In order to tackle it, the SCA method is applied. By introducing the auxiliary variables x_n, we have $\exp(a_n)-\frac{f_n}{Cn}\leq\exp(x_n)$ and

$\exp(x_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2})$. Then the problem $\mathbf{P}_{2.3}$ becomes:

$$\mathbf{P}_{2.4}: \max_{f_n,\rho_n,a_n,x_n} \Phi_{SE}\sum_{n\in\mathcal{N}} a_n - \Phi_{EE}\sum_{n\in\mathcal{N}}(\zeta\exp(\rho_n)+p_r+\epsilon f_n^3)$$

$$s.t.\ \ a_n \geq \ln(R_n^{th}),\ \forall n\in\mathcal{N}, \tag{2.11a}$$

$$\zeta\exp(\rho_n)+p_r+\epsilon f_n^3 \leq P_n^{th},\ \forall n\in\mathcal{N}, \tag{2.11b}$$

$$f_n^{min} \leq f_n \leq f_n^{max}, \forall n\in\mathcal{N}, \tag{2.11c}$$

$$\exp(a_n) - \frac{f_n}{Cn} \leq \exp(x_n), \tag{2.11d}$$

$$\exp(x_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}). \tag{2.11e}$$

Since the constraint (2.11d) is non-convex, by using the SCA technique, the first-order Taylor expansion is used to approximate the right part. Thus, $\mathbf{P}_{2.4}$ can be solved by iteratively solving the following approximate problem, given as

$$\mathbf{P}_{2.5}: \max_{f_n,\rho_n,a_n,x_n} \Phi_{SE}\sum_{n\in\mathcal{N}} a_n - \Phi_{EE}\sum_{n\in\mathcal{N}}(\zeta\exp(\rho_n)+p_r+\epsilon f_n^3) \tag{2.12a}$$

$$s.t.\ \ a_n \geq \ln(R_n^{th}),\ \forall n\in\mathcal{N}, \tag{2.12b}$$

$$\zeta\exp(\rho_n)+p_r+\epsilon f_n^3 \leq P_n^{th},\ \forall n\in\mathcal{N}, \tag{2.12c}$$

$$f_n^{min} \leq f_n \leq f_n^{max}, \forall n\in\mathcal{N}, \tag{2.12d}$$

$$\exp(a_n) - \frac{f_n}{Cn} \leq \exp(\overline{x}_n^k) + \exp(\overline{x}_n^k)(x_n-\overline{x}_n^k), \tag{2.12e}$$

$$\exp(x_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}), \tag{2.12f}$$

where $\overline{x}_n^k, n\in\mathcal{N}$ are the given local points at the kth iteration. The above problem is convex and can be readily solved by using the existing convex optimization tool.

2.3.2 Max-Min Fairness $\alpha = \infty$

In this case, the objective of the system is to maximize the minimum computation rate among all the users. For that purpose, the sum computation bit rate and energy efficiency may have to be comprised to achieve this Max-Min fairness. When $\alpha = \infty$, the original

optimization problem $\mathbf{P}_{2.1}$ becomes

$$\begin{aligned}\mathbf{P}_{2.6}: \quad & \max_{f_n,p_n} \Phi_{SE} \min_{n\in\mathcal{N}} B\log_2(1+\frac{p_n g_n^2}{\sum_{i=1}^{n-1} p_i g_i^2+\sigma^2}) \\ & + \frac{f_n}{C_n} - \Phi_{EE}\sum_{n\in\mathcal{N}} \zeta p_n + p_r + \epsilon f_n^3 \\ & s.t. \;\; C1-C4.\end{aligned} \tag{2.13}$$

It is difficult to directly solve the max-min problem $\mathbf{P}_{2.6}$. By introducing a new variable l as a lower bound and auxiliary variables $\rho_n = \log(p_n)$, the problem $\mathbf{P}_{2.6}$ can be transformed into

$$\max_{f_n,\rho_n,l} \Phi_{SE} * l - \Phi_{EE}\sum_{n\in\mathcal{N}} (\zeta\exp(\rho_n) + p_r + \epsilon f_n^3) \tag{2.14a}$$

$$s.t.\; l \geq R_n^{th},\; \forall n\in\mathcal{N}, \tag{2.14b}$$

$$\zeta\exp(\rho_n) + p_r + \epsilon f_n^3 \leq P_n^{th},\; \forall n\in\mathcal{N}, \tag{2.14c}$$

$$f_n^{min} \leq f_n \leq f_n^{max},\; \forall n\in\mathcal{N}, \tag{2.14d}$$

$$l \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}) + \frac{f_n}{C_n}. \tag{2.14e}$$

The objective function of the above problem is concave due to the subtraction of a linear term and a convex term, the first three constraints are convex. The final constraint (2.14e) can be expressed as

$$\ln(\frac{\sigma^2}{g_n^2}e^{-\rho_n} + \sum_{i=1}^{n-1}\frac{g_n^2}{g_i^2}e^{\rho_i-\rho_n}) + \ln((\exp(\ln 2\frac{l-\frac{f_n}{C_n}}{B})) - 1) \leq 0, \tag{2.15}$$

which is neither convex nor concave. By introducing the new auxiliary variable z_n,

$$\exp(z_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}), \tag{2.16}$$

the problem can be expressed as

$$\max_{f_n,\rho_n,l,z_n} -\Phi_{EE}\sum_{n\in\mathcal{N}}(\zeta\exp(\rho_n)+p_r+\epsilon f_n^3)+\Phi_{SE}*l \tag{2.17a}$$

$$s.t.\ l \geq R_n^{th},\ \forall n\in\mathcal{N}, \tag{2.17b}$$

$$\zeta\exp(\rho_n)+p_r+\epsilon f_n^3 \leq P_n^{th},\ \forall n\in\mathcal{N}, \tag{2.17c}$$

$$f_n^{min} \leq f_n \leq f_n^{max},\ \forall n\in\mathcal{N}, \tag{2.17d}$$

$$l \leq \exp(z_n)+\frac{f_n}{C_n}, \tag{2.17e}$$

$$\exp(z_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}). \tag{2.17f}$$

Since the constraint (2.17e) is not convex, similar to $\mathbf{P}_{2.4}$, the SCA method is used to solve the problem given by eq. (2.17a). In this case, eq. (2.17a) is solved by iteratively solving the approximate problem, given as

$$\mathbf{P}_{2.7}: \max_{f_n,\rho_n,l,z_n} -\Phi_{EE}\sum_{n\in\mathcal{N}}(\zeta\exp(\rho_n)+p_r+\epsilon f_n^3)+\Phi_{SE}*l \tag{2.18a}$$

$$s.t.\ \ l \geq R_n^{th},\ \forall n\in\mathcal{N}, \tag{2.18b}$$

$$\zeta\exp(\rho_n)+p_r+\epsilon f_n^3 \leq P_n^{th},\ \forall n\in\mathcal{N}, \tag{2.18c}$$

$$f_n^{min} \leq f_n \leq f_n^{max}, \forall n\in\mathcal{N}, \tag{2.18d}$$

$$l \leq \exp(\overline{z}_n^k)+\exp(\overline{z}_n^k)(z_n-\overline{z}_n^k)+\frac{f_n}{C_n}, \tag{2.18e}$$

$$\exp(z_n) \leq B\log_2(1+\frac{\exp(\rho_n)g_n^2}{\sum_{i=1}^{n-1}\exp(\rho_i)g_i^2+\sigma^2}), \tag{2.18f}$$

where $\overline{z}_n^k, n\in\mathcal{N}$ are the given local points at the kth iteration. It is not difficult to prove that the above problem is convex and can be readily solved by using the existing convex optimization tool. The algorithm for solving $\mathbf{P}_{2.5}$ and $\mathbf{P}_{2.7}$ is given in Algorithm 2.1.

Table 2.1: SCA Iteration Algorithm

Algorithm 2.1: The SCA iterative algorithm for $\mathbf{P}_{2.5}$ and $\mathbf{P}_{2.7}$

1) **Input settings:**
   the error tolerance $\xi > 0$, $R_n^{th} > 0$ and $P_n^{th} > 0$,
   the maximum iteration number K.
2) **Initialization:**
   $k = 0$, $f_n(0)$, $\rho_n(0)$,$a_n(0)$ $\overline{x}_n^0$;
3) **Optimization:**
   ⊵ **for k=1:K**
      solve $\mathbf{P}_{2.5}$/$\mathbf{P}_{2.7}$ by using the interior-point method;
      obtain the solution $\{f_n^{k*}, \rho_n^{k*}, a_n^{k*}, x_n^{k*}\}$ and the system efficiency H_k^*;
      if $\|H_k^* - H_{k-1}^*\| \leq \xi$;
         the maximum system efficiency H^* is obtained;
         break;
      else
         update $\overline{x}_n^{k+1} = x_n^{k*}$ and $k = k + 1$.
      end
   ⊵ **end**
4) **Output:**
   $\{f_n^*, \rho_n^*\}$ and system efficiency H^*

2.3.3 Fairness with $\alpha = 0$

In this case, the tradeoff between the weighted sum computation rate and the power consumption cost is considered. The original problem $\mathbf{P}_{2.1}$ can be expressed as

$$\mathbf{P}_{2.8}: \quad \max_{f_n, p_n} \Phi_{SE} \sum_{n\in\mathcal{N}} (B\log_2(1+\frac{p_n g_n^2}{\sum_{i=1}^{n-1} p_i g_i^2 + \sigma^2}) + \frac{f_n}{C_n}) - \Phi_{EE} \sum_{n\in\mathcal{N}} (\zeta p_n + p_r + \epsilon f_n^3)$$
$$s.t. \;\; C1 - C4. \tag{2.19}$$

Based on the proof of Theorem 1 and Theorem 3 in [42], the problem (2.19) is NP-hard. In order to solve it, we apply a SCALE method to approximate problem $\mathbf{P}_{2.8}$ into a sequence of convex programs and obtain the sub-optimal solution by the proposed algorithm [43], as follows:

$$a\log z + b \le \log_2(1+z). \tag{2.20}$$

That is tight at $z = z_0$ when the approximation constants are given as

$$a = \frac{z_0}{1+z_0}, \tag{2.21a}$$
$$b = \log_2(1+z_0) - \frac{z_0}{1+z_0}\log_2(z_0). \tag{2.21b}$$

By applying the SCALE method to the problem (2.19), and the logarithmic change of variables $\rho_n = \log(p_n)$, we can obtain the following problem as

$$\mathbf{P}_{2.9} \quad \max_{f_n, \rho_n} \Phi_{SE} \sum_{n\in\mathcal{N}} (B\overline{R}_n(\rho_n; a_n, b_n) + \frac{f_n}{C_n}) - \Phi_{EE} \sum_{n\in\mathcal{N}} (\zeta\exp(\rho_n) + p_r + \epsilon f_n^3), \tag{2.22a}$$
$$s.t. \;\; B\overline{R}_n(\rho_n; a_n, b_n) + \frac{f_n}{C_n} \ge R_n^{th}, \forall n \in \mathcal{N}, \tag{2.22b}$$
$$\zeta\exp(\rho_n) + p_r + \epsilon f_n^3 \le P_n^{th}, \forall n \in \mathcal{N}, \tag{2.22c}$$
$$f_n^{min} \le f_n \le f_n^{max}, \forall n \in \mathcal{N}, \tag{2.22d}$$
$$\tag{2.22e}$$

where

$$z_n = \frac{\exp(\rho_n) g_n^2}{\sum_{i=1}^{n-1} \exp(\rho_i) g_i^2 + \sigma^2}, \tag{2.23a}$$

$$a_n = \frac{z_n}{1 + z_n}, \tag{2.23b}$$

$$b_n = \log_2(1 + z_n) - \frac{z_n}{1 + z_n} \log_2(z_n), \tag{2.23c}$$

and

$$\begin{aligned} \overline{R}_n(\rho_n; a_n, b_n) &= a_n \log_2\left(\frac{\exp(\rho_n) g_n^2}{\sum_{i=1}^{n-1} \exp(\rho_i) g_i^2 + \sigma^2}\right) + b_n \\ &= a_n[\log_2(g_n^2) + \ln 2\rho_n - \log(\sum_{i=1}^{n-1} \exp(\rho_i) g_i^2 + \sigma^2)] + b_n. \end{aligned} \tag{2.24}$$

Here, we note the log-sum-exp is convex, thus $\overline{R}_n(\rho_n; a_n, b_n)$ is a concave function because it is the sum of linear and concave terms within the square brackets. Thus the problem (2.22a) is a standard concave maximization problem. The algorithm for solving $\mathbf{P}_{2.8}$ based on solving the convex relaxation problem $\mathbf{P}_{2.9}$ is given in Algorithm 2.2.

2.4 Performance Evaluation

In this section, we present the performance results of the proposed scheme. The parameters are set as follows [29]. The system bandwidth is $B = 2$ MHz, the number of total UEs is $N = 3$, the local data process capacity for one bit is $C_n = 10^3$ cycles. The computation energy efficiency coefficient is $\epsilon = 10^{-28}$, the power weight $\zeta = 2$. The channel between the MEC server and each UE is modeled as the joint effect of large-scale and small-scale fading, with $g_k^2/\sigma^2 = G_k h_k$, $G_1 = 7$, $G_2 = 40$, $G_3 = 144$. h_k is the unitary Gaussian random variable. The maximum and minimum computation capacity of each UE is set equally as f_n^{max} is 10^9 Hz and f_n^{min} is 10^6 Hz. The circuit power $p_r = 5$ dBm. The results are obtained by performing over different random channel realizations. Two benchmark schemes, namely Offloading Only + NOMA scheme and Local&Offloading + FDMA scheme, are considered for comparison. The proposed scheme is marked as Local&Offloading + NOMA.

In Fig. 2.2, the fairness case with $\alpha = 1$ is studied. In order to balance the number of

Table 2.2: SCALE Iteration Algorithm

Algorithm 2.2: The SCALE iterative algorithm for $\mathbf{P}_{2.9}$

1) **Input settings:**
   the error tolerance $\xi > 0$, $R_n^{th} > 0$ and $P_n^{th} > 0$, $a_n^1 = 1, b_n^1 = 0$,
   the maximum iteration number K.
2) **Initialization:**
   $k = 0$, $f_n(0)$, $\rho_n(0)$, $a_n(0)$ and $b_n(0)$;
3) **Optimization:**
   ⊵ for *k*=1:K
      solve $\mathbf{P}_{2.9}$ by using the interior-point method;
      obtain the solution $\{f_n^*, \rho_n^*\}$ and the rate approximation $\overline{R}_n^{k*}$;
      if $\|\overline{R}_n^{k*} - \overline{R}_n^{k-1*}\| \leq \xi$;
         the maximum system efficiency H^* is obtained;
         break;
      else
         update a_n^{k+1}, b_n^{k+1} by (2.21) and $k = k + 1$.
      end
   ⊵ end
4) **Output:**
   $\{f_n^*, \rho_n^*\}$ and system efficiency H^*

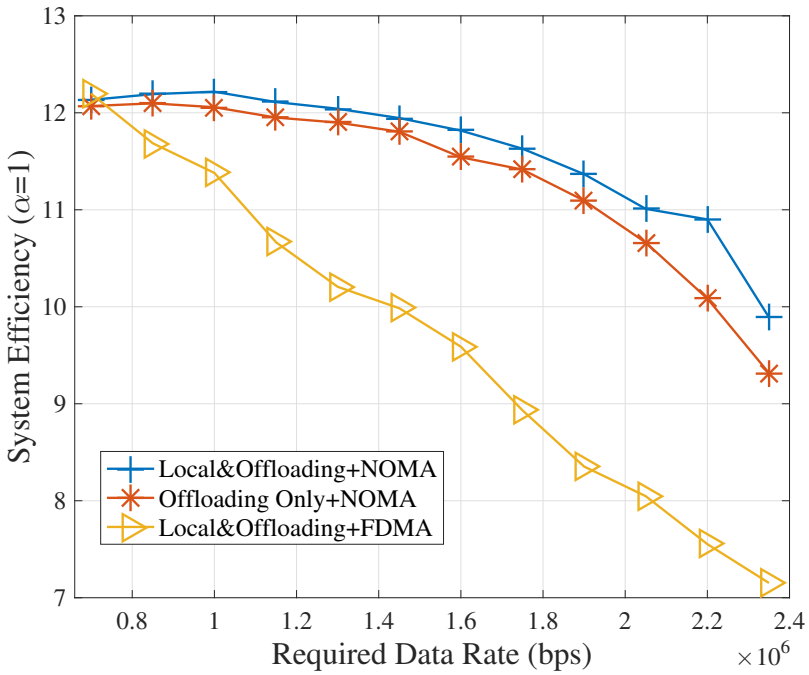

Fig. 2.2: System efficiency for $\alpha = 1$.

computation bits and energy efficiency, we choose $\Phi_{SE} = 0.3$ and $\Phi_{EE} = 0.7$. The results show that the proposed model has a higher system efficiency than two benchmark schemes. The system efficiency decreases with the increase of the required computation bits rate R_n^{th} because each UE needs to increase is transmit power as well as their local process rate in order to meet the data requirement. The increase of the transmitting power and local process rate both will elevate the energy cost, resulting in the decrease of the overall system efficiency. It can also be seen that using NOMA can achieve a higher system efficiency compared with using FDMA.

In Fig. 2.3, the fairness case with $\alpha = \infty$ is studied. $\Phi_{SE} = 10^{-6}$ and $\Phi_{EE} = 0.5$ are the weighting factors. It can be seen that the system efficiency decreases with the increase of R_n^{th}. The reason is the same as for Fig. 2.2. It can also be seen that the proposed scheme is still better than the two benchmark schemes. Moreover, the performance of the FDMA method decreases faster than that of the scheme with NOMA.

The system efficiency of the maximum sum rate case is presented in Fig. 2.4, where

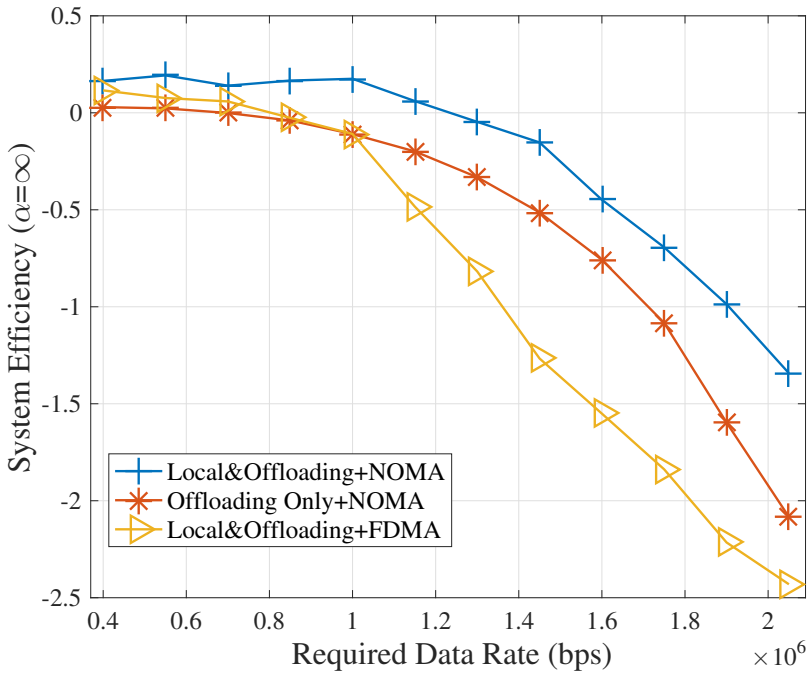

Fig. 2.3: System efficiency for $\alpha = \infty$.

$\alpha = 0$. The smaller the required data computation rate is, the smaller the difference in system efficiency among three schemes becomes. The reason is that when the system requires a small computation bits rate, it becomes less difficult for all the schemes to meet the requirement. However, as the required data rate keeps increasing, the energy cost for different schemes increases as well. The proposed scheme can allocate the resources in a more efficient way based on the combination of local computation and NOMA offloading, thus outperforming other benchmark schemes.

In Fig. 2.5, we set $\xi = 10^{-4}$. Only several iterations are required for our proposed algorithms to converge, showing the computation efficiency of the proposed algorithm.

2.5 Chapter Conclusion

In this chapter, we formulated a fairness resource allocation problems in an ultra-dense MEC-enabled IoT network with NOMA to improve the fairness among IoT users. We are considering three special fairness cases to meet different system goals. In each case,

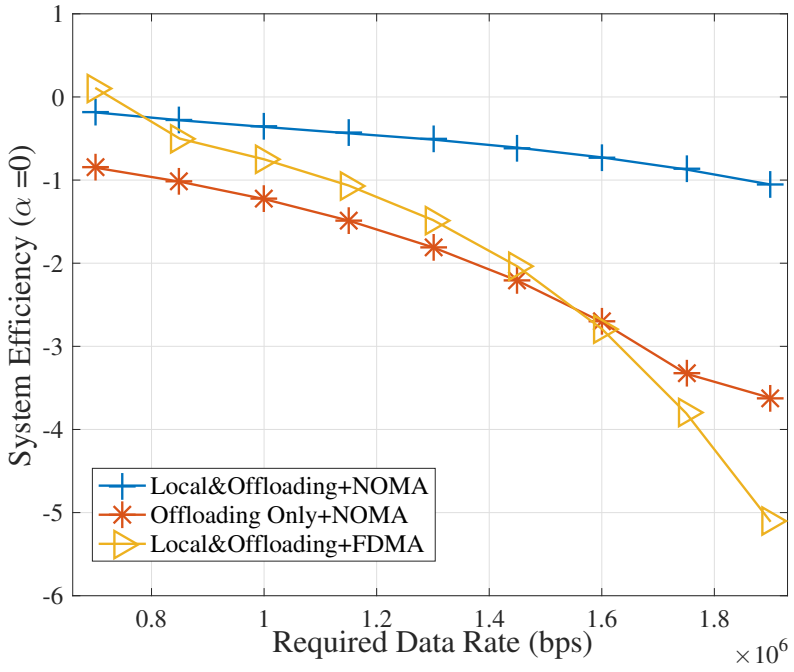

Fig. 2.4: System efficiency for $\alpha = 0$.

the resource allocation schemes were obtained by using the SCA or the SCALE method. Simulation results verified that our proposed scheme can achieve a better performance than two other benchmark schemes.

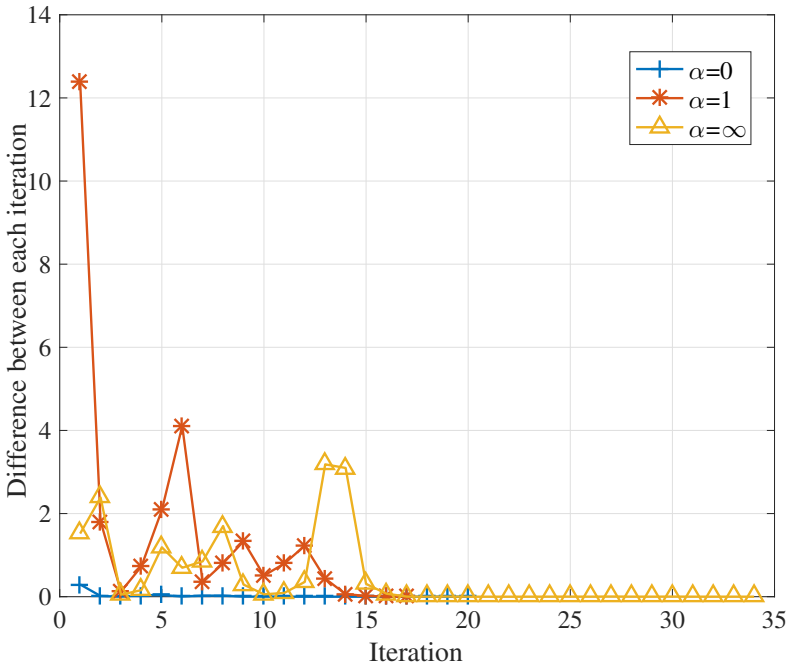

Fig. 2.5: Convergence with Iteration

CHAPTER 3

Hierarchical Energy Efficient Mobile Edge Computing in IoT Networks

3.1 Introduction

In the previous chapter, we investigated resource allocation for MEC enabled IoT networks with NOMA in a single time slot. In addition, we have verified that the combination of NOMA and MEC can significantly improve the system EE. In real applications, the users' behavior and environmental conditions change dynamically. The adaptive resource allocation method is required to capture the dynamics in the networks. Mao *et al.* [44] developed an online joint radio and computational management algorithm for multi user MEC systems, which aims for minimizing the long-term average weighted sum power consumption of the mobile devices and the MEC server. Lyu *et al.* [45] designed a perturbed Lyapunov function to stochastically maximize a network utility balancing throughput and fairness.

In this chapter, a flexible hierarchical edge computing architecture is used, in which edge nodes and cloud servers with diverse power and computation capabilities form two tiers to best serve end user needs. A hierarchical communication and computation framework for jointly optimizing energy consumption and computation rate is proposed. The hierarchical framework consists of three layers, i.e. sensor layer, edge layer, and cloud server layer. The accumulated computing power minimization and computing rate maximization trade-off optimization problem is formulated in this part. We will develop a prediction-based edge node turning on/off algorithm based on long-term data dynamics to reduce system operating cost, while we devise the dynamic resource allocation algorithm based on shot-term data dynamics. The LSTM network [46] for arrival tasks prediction mode is applied in long-term process unit status decision operation. Furthermore, the real-life user data will be employed to test and to validate our proposed algorithms. The optimization of offloading transmit

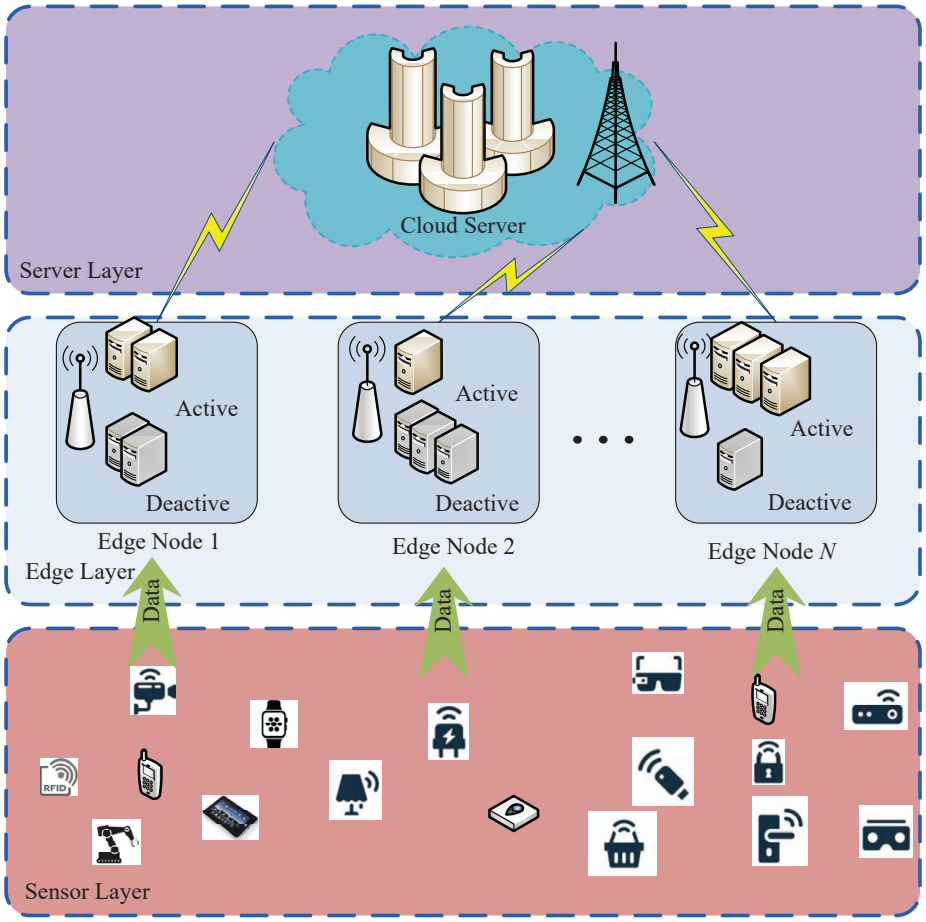

Fig. 3.1: System model for the three-layer IoT network.

power and local processing speed is determined based on Lyapunov optimization method.

3.2 System Model

We consider a three-layer IoT network as described in Fig. 3.1. The first layer is IoT senor layer, which consists of different IoT sensor devices such as smartphones, environmental sensors, and wearable devices. The second layer is edge layer consisting of mobile edge nodes, while the third layer is the server layer consisting of centralized or cloud servers. All the sensor devices are deployed around randomly distributed edge nodes. IoT sensor devices keep collecting and uploading the data to their associated edge nodes for data processing. There are N edge nodes in the system, which provide the data processing service for the IoT devices. After the massive raw data is received, each edge node can choose to process

Table 3.1: LIST OF SYMBOLS

Symbol	Definition
$A_n(t)$	The workload arriving at edge node.
$Q_n(t)$	The buffer size.
$R_n^{tot}(t)$	The total processing rate of edge node n.
$E_n^{tot}(t)$	The total energy cost of edge node n.
$C_{m,n}$	The number of computation cycles to compute 1 bit of data.
$f_{m,n}(t)$	The computing rate (cycles per second).
$g_n(t)$	The channel gain of edge node n.
$p_n(t)$	The transmit power of edge node n.
ζ	The amplifier coefficient.
$p_n(t)$	The transmission power consumption.
p_r	The constant circuit power.

the data locally, or to offload the data to the more powerful cloud server, or a combination of both. We furthermore assume that each edge node has M processing units (PUs), which can be turned on or turned off individually based on needs.

Let $A_n(t)$ denote the workload arriving at edge node n at time t. Note that the computing workload at each edge node dynamically changes from time to time but with a predictable pattern in many cases. We allow each edge node adaptively turn on/turn off a subset or all of its PUs to save energy. The operation is done based on the prediction of traffic patterns in a relatively long-term scale.

At t, $A_n(t)$ bits arrive at edge node n from the connected IoT sensor devices. The size of the buffered data at edge node n becomes

$$Q_n(t+1) = \max\{Q_n(t) + A_n(t) - R_n^{tot}(t)\tau, 0\}, \tag{3.1}$$

where $Q_n(t)$ is the buffer size at t, $R_n^{tot}(t)$ is the total processing rate of edge node n at time t, which include both local processing rate and the cloud processing rate achieved through offloading. We consider a partial offloading for each edge node so it can decide how to partition workload between itself and the cloud server. Table 3.1 lists the symbols and

their definitions.

3.2.1 Local Processing Mode

At t, edge node n computes the workload at the buffer. In particular, the workload for local processing at node n is furthermore divided into M parts. Let $C_{m,n}$ be the number of computation cycles needed to compute one bit of data at $PU(m,n)$. Each PU can compute the data in the entire transmission duration. Let $f_{m,n}(t)$ be the computing rate (cycles per second) at t for $PU(m,n)$ and $s_{m,n}$ represent the status of $PU(m,n)$, where $s_{m,n} = 1$ if $PU(m,n)$ is active and $s_{m,n} = 0$ otherwise. Therefore, the local computing rate of edge node n is calculated as

$$r_n^{local}(t) = \sum_{m=1}^{M} \frac{f_{m,n}(t)}{C_{m,n}} s_{m,n}(t). \tag{3.2}$$

The energy consumption of local computing is expressed as

$$e_n^{local}(t) = \sum_{m=1}^{M} [\epsilon_{m,n} f_{m,n}^3(t) s_{m,n}(t)\tau + p_{m,n}^{idle} s_{m,n}(t)\tau], \tag{3.3}$$

where $\epsilon_{m,n}$ is the energy efficiency coefficient for an active $PU(m,n)$ and $p_{m,n}^{idle}$ is the energy consumption of an idle $PU(m,n)$. τ is the duration of each time slot.

3.2.2 Data Offloading Mode

In the partial offloading, part of the data in each edge node can be offloaded to the cloud server. Two different offloading schemes are considered, i.e. FDMA and NOMA.

FDMA based offloading

Assume the total channel bandwidth between edge nodes and the cloud server is W, which is equally partitioned among N edge nodes by using FDMA. So the bandwidth of each channel is $B = \frac{W}{N}$. Let $g_n(t)$ and $p_n(t)$ represent the channel gain and transmit power for edge node n, respectively. The offloading rate for node n under the FDMA method can

be expressed as

$$r_n^{off}(t) = B\log_2(1 + \frac{p_n(t)g_n^2(t)}{\sigma_n^2}). \tag{3.4}$$

The corresponding energy consumption is

$$e_n^{off}(t) = (\zeta p_n(t) + p_r)\tau, \tag{3.5}$$

where ζ is the amplifier coefficient. $p_n(t)$ is the transmission power consumption and p_r is the constant circuit power.

NOMA based offloading

In NOMA, each edge node pairs with another edge node for transmission. For that, at time t, all the edge nodes are firstly ranked by their channel quality, i.e. $g_1(t) \leq g_2(t) \leq \cdots \leq g_N(t)$. Then, NOMA groups are formed according to the following rules. Node 1 pairs with node $K+1$, node 2 pairs with node $K+2$, $\cdots$, node K pairs with node N, where $K = N/2$. The two nodes in the same NOMA group can offload workload to the cloud server simultaneously on the same radio resource. Successive interference cancellation (SIC) technique is applied at the cloud server to decode the signals for each node [38]. Specifically, let $g_n(t), g_k(t), p_n(t)$, and $p_k(t)$ respectively represent the channel gains and transmit powers for both strong node n and weak node k in the same group, where $g_n(t) \geq g_k(t)$. The cloud server first decodes the signal of the strong node n, then subtracts the decoded signal of node n from the composite received signal and proceeds to decode the signal of the weak node k. When decoding the signal from node n, the signal from node k stays as interference. Compared with FDMA, the bandwidth allocated for each NOMA group is $2B$. Correspondingly, the offloading rates $(r_n^{off}(t), r_k^{off}(t))$ for the strong and weak nodes

(n, k) can be expressed as

$$r_n^{off}(t) = 2B \log_2(1 + \frac{p_n(t) g_n^2(t)}{p_k(t) g_k^2(t) + \sigma_n^2}), \quad \text{strong node}, \tag{3.6}$$

$$r_k^{off}(t) = 2B \log_2(1 + \frac{p_k(t) g_k^2(t)}{\sigma_k^2}), \quad \text{weak node}, \tag{3.7}$$

where σ_n^2 and σ_k^2 are the noise powers at the strong and weak nodes, respectively. The corresponding total energy consumption for nodes (n, k) using NOMA method is given as

$$e_{n,k}^{off}(t) = (\zeta_n p_n(t) + \zeta_k p_k(t) + 2p_r)\, \tau, \tag{3.8}$$

where (ζ_n, ζ_k) are the amplifier coefficients. The first two terms represent the transmission power consumption, while the third is the constant circuit power consumption. Moreover, p_r is assumed to be the same for all the edge nodes.

3.3 Problem Formulation

We aim to jointly design the data offloading and local computing in this work. The total computational throughput $R_n^{tot}(t)$ and the total energy consumption $E_n^{tot}(t)$ for node n at t are expressed as

$$R_n^{tot}(t) = r_n^{local}(t) + r_n^{off}(t), \tag{3.9}$$

$$E_n^{tot}(t) = e_n^{local}(t) + e_n^{off}(t). \tag{3.10}$$

Our goal is to achieve a high computational throughput as well as a high energy efficiency by minimizing the power consumption and maximizing the computed bits. These two performance metrics are normally two conflicting goals to optimize. We exploit the weighted sum to tackle this multi-objective problem and define the system cost as follows [47]:

$$F(\mathbf{x}(t), \mathbf{s}(t)) = \phi_{fe} E_{tot}(\mathbf{x}(t), \mathbf{s}(t)) - \phi_{fs} R_{tot}(\mathbf{x}(t), s(t)), \tag{3.11}$$

where $x(t) =: \{f_{m,n}(t), p_n(t)\}$, $E_{tot}(\mathbf{x}(t), \mathbf{s}(t)) = \sum_{n=1}^{N} E_n^{tot}(t)$ is the overall energy cost by all the edge nodes at Edge layer, while $R_{tot}(\mathbf{x}(t), s(t)) = \sum_{n=1}^{N} R_n^{tot}(t)$, is the overall system computation throughput. Furthermore, (ϕ_{fe}, ϕ_{fs}) are the energy and rate coefficients. The problem is formulated as

$$\begin{aligned}
\mathbf{P}_{3.1} \quad & \min_{p_n(t), f_{m,n}(t), s_{m,n}(t)} \limsup_{\mathcal{B}\to\infty} \frac{1}{\mathcal{B}} \sum_{t=0}^{\mathcal{B}-1} \mathbb{E}\left\{F(\mathbf{x}(t), \mathbf{s}(t))\right\} \\
& s.t. \quad C3.1: \quad \limsup_{\mathcal{B}\to\infty} \frac{1}{\mathcal{B}} \sum_{t=0}^{\mathcal{B}-1} \sum_{n=1}^{N} \mathbb{E}\left\{Q_n(t)\right\} < \infty, \\
& \qquad C3.2: \quad f_{m,n}^{min} \le f_{m,n}(t) \le f_{m,n}^{max}, \forall m, n, \\
& \qquad C3.3: \quad 0 \le p_n(t) \le P_n^{max}, \forall n, \\
& \qquad C3.4: \quad s_{m,n}(T) \in \{0, 1\}, \forall m, n.
\end{aligned} \tag{3.12}$$

The first constraint $C3.1$ is the queue stability constraint. Constraints $C3.2$ and $C3.3$ represent the ranges of edge node processing frequency and transmission power, while constraint $C3.4$ denotes active or de-active state for each PU. Each PU of an edge node can be turned on or turned off depending on the demands. In the following, we will use the system cost and system efficiency alternately, since minimizing the system cost defined in $\mathbf{P}_{3.1}$ is the same as maximizing the system efficiency.

Firstly, the problem $\mathbf{P}_{3.1}$ is an NP-Hard mixed integer nonlinear programming problem, which normally has a very high computational complexity and is very challenging to solve in a real-time manner. On the other hand, the PU turn-on/turn-off operation may not need to be made in real-time for most systems due to on-off overhead concerns and hardware constraints. To address both real time need to allocate computing and communication resources as well as non-real-time need to turn on/off processing units, we propose a two-timescale algorithm to solve this optimization problem. The small timescale problem is executed every time slot t, while the large timescale problem is executed every epoch with the duration of T time slots. Correspondingly, the original problem $\mathbf{P}_{3.1}$ can be decomposed into two sub-problems. The first sub-problem at large timescale decides how many PUs are needed for each edge node, while the second one at small timescale is the

computing/communication resource allocation problem. To solve the first sub-problem, we design the large timescale prediction scheme to estimate the arriving workload, based on which turn-on/turn-off decisions for the PUs at each edge node are made. As a result, the number of active PUs changes from one epoch to another. For the second sub-problem, we aim for minimizing the total cost of both the energy consumption and the delay by using efficient resource allocation.

3.3.1 Large Timescale Optimization Model

This sub-problem aims for minimizing the energy consumption from large timescale perspective. The status of PU $s_{m,n}(T)$, $m = 1, 2, \cdots, M$, at node n is determined in every T time slots. The sub-problem can be formulated as

$$\begin{aligned} \mathbf{P}_{3.2} \quad & \min_{\{s_{m,n}(T)\}} \sum_{T \in \mathbf{T}} F_1\left(\mathbf{x}^*, \mathbf{s}(T)\right) \\ & s.t. \quad s_{m,n}(T) \in \{0, 1\}, \forall m, n, \end{aligned} \tag{3.13}$$

where $F_1(\mathbf{x}^*, \mathbf{s}(T)) = \phi_{fe} E_{tot}(\mathbf{x}^*, \mathbf{s}(T)) - \phi_{fs} R_{tot}(\mathbf{x}^*, \mathbf{s}(T))$. $\mathbf{x}(t)$ is firstly set to $\mathbf{x}(T-1)^*$ in the objective function, which is the optimal resource allocation in the previous epoch. The decision is made based on both the prediction of the arriving workloads and also the efficiency of turn-on/off. The main idea for deciding turn-on/off of a PU is presented as follows.

The decision to turn on the inactive units depends on the workload state. If the arriving workload to an edge node keeps increasing, this edge node needs to turn on more PUs to support the arising workload. The turning-on operation is also performed at the beginning of each epoch. This operation is the coarse control in the long-term timescale model to protect the negotiated service level agreement (SLA). The finer control is implemented at the short timescale model to regulate the network parameters for the workload processing. This cooperative long-term and short-term timescale models would not only reduce the operating cost but also ensure the system's stability. Therefore, the proposed two-timescale framework is more efficient and flexible. The turning-on/off algorithm at each epoch is summarized

in Algorithm. 3.1. The remaining task is how to estimate the arriving workloads, which is presented in the sections 3.4.

Table 3.2: Prediction Based Coordination Algorithm

Algorithm 3.1: Prediction Based Coordination Algorithm

1) **Input settings:**
   Workload $A(t)$
2) **Initialization:**
   Set $\mathbf{s}(t) = \mathbf{1}$;
3) **Optimization:**
   Predict the arriving workload for each edge node at the beginning of each epoch T.
   At every epoch T, perform the following jobs according to estimated workload:
   if workload increases
   Turn on the PUs from de-active set until the available resources can serve the arriving workload.
   else if workload decreases
   Turn off the PUs from the active set of the edge nodes until reaching the balance of demand and resources.
4) **Output:**
   Set the processing unit status $\mathbf{s}(T) = \mathbf{s}^*(T)$.

3.3.2 Small Timescale Optimization Model

In this subsection, we aim for minimizing the system cost at each time slot, given that the optimal value of the status vector $\mathbf{s}_t = \mathbf{s}^*(T)$ is obtained in the large timescale model. The small timescale model directly uses $\mathbf{s}^*(T)$ to seek the optimal value for $\mathbf{x}(t)$. Thus, we formulate the second sub-problem as follows:

$$\begin{aligned} \mathbf{P}_{3.3}: \quad & \min_{\mathbf{x}(t)} \quad \sum_{t=t_0}^{t_0+T-1} F_2(\mathbf{x}(t), \mathbf{s}^*(T)) \\ & s.t. \qquad C3.1 - C3.3, \end{aligned} \tag{3.14}$$

where t_0 is the starting point of the current epoch and $F_2(\mathbf{x}(t), \mathbf{s}^*(T)) = \phi_{fe} E_{tot}(\mathbf{x}(t), \mathbf{s}^*(T)) - \phi_{fs} R_{tot}(\mathbf{x}(t), \mathbf{s}^*(T))$. The Lyapunov optimization method and SCALE method are used to

make the resource allocation decision in every time slot, which will be presented in Section 3.5.

3.4 Large Timescale Workload Prediction

In this section, we firstly present the overview of machine learning, which is applied to the prediction method. Next, we provide some constraints of this method. We then present the long short term memory network to overcome these constraints.

3.4.1 Overview of Machine Learning Based Prediction Method

Many stochastic mechanisms have been exploited to effectively predict the workload flows. These methods can be generally classified into two categories, linear prediction methods and nonlinear prediction methods. For the linear prediction methods, one of the best prediction mechanisms in the correlated time series category is the autoregressive–moving-average (ARMA) model [48, 49], while the most commonly used non-linear mechanism is neural network. ARMA is a typical parametric regression model, which assumes that the traffic condition is a stationary process. It implies that the mean, variance and autocorrelation all stay constant. However, the ARMA method cannot capture the rapid variational process underlying the traffic data due to it concentrates on the mean value of the past series data.

The neural network technique is able to model more complicated data by using the distributed and hierarchical feature representation. Recently, many deep learning models that can extract multilevel features have been developed. To train the network parameters, they employ the machine learning such as supervised/unsupervised/semi-supervised learning methods, reinforcement learning schemes and nature-inspired algorithms [50–54]. Workload prediction accuracy can be greatly improved. One of the common methods in deep neural network forecast is based on recurrent neural network (RNN) [52, 54], which is used in this paper. In particular, RNN works efficiently with time series and sequence modeling tasks, because it contains a self-recurrent loop that facilitates transporting information from one time slot to another. Note that the traditional neural network cannot

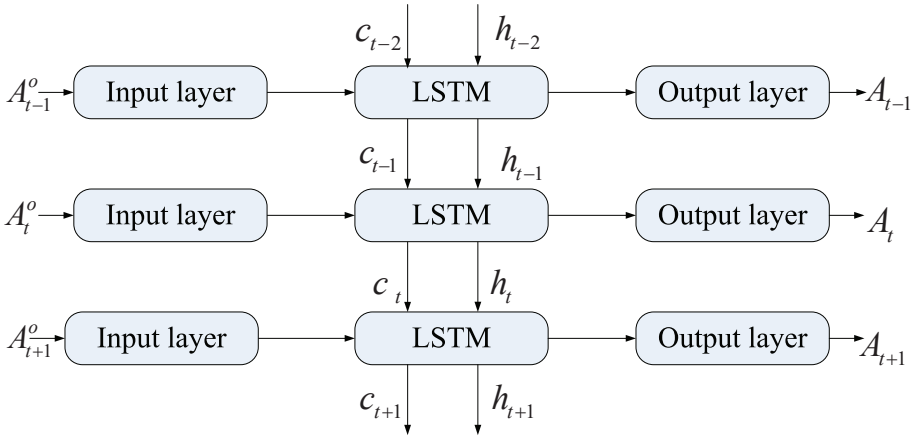

Fig. 3.2: LSTM network.

achieve the same level performance in the temporal-spatial problems as it does not have the interconnection between nodes in the same layer. RNN introduces hidden units that allows the current state to receive feedback from the previous state.

3.4.2 Long Short Term Memory Network

The original RNN only has one state, i.e. **h**. Therefore it is very sensitive to the short timescale input and has the gradient vanishing problem for large timescale forecasting. To tackle this issue, we use the long short term memory (LSTM) model, which is one of the specially designed RNN models. LSTM [46] does the advanced time series prediction with long temporal dependency. It can learn information with long time spans and determine the optimal time lags autonomously. The key component that makes LSTM work for the long-term dependencies is called memory block. Fig. 3.2 shows a typical LSTM network, which usually consists of one input layer, one output layer, and one recurrent hidden layer containing the memory block.

The memory block integrated in the LSTM network is illustrated in Fig. 3.3. Here, the memory block is the recurrently connected subnet, which contains functional modules such as memory cells and gates. The memory cells are used for memorizing the temporal states of the network while the gates are composed of sigmoid layers, which are responsible for controlling the amount of information flows. According to the corresponding functions,

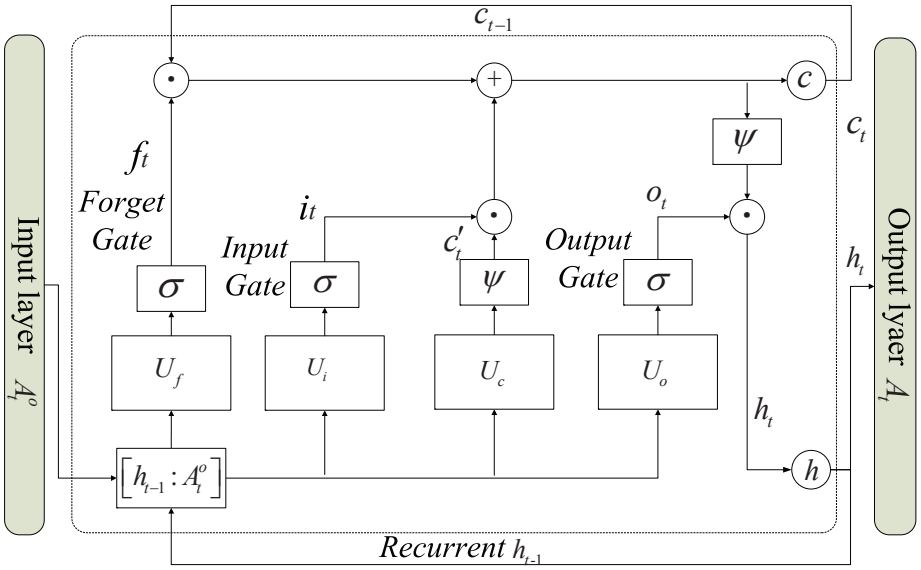

Fig. 3.3: LSTM memory block.

gates can be classified into input gates, output gates, and forget gates. The input gate controls how much new information flows into the memory cell and its weight matrix is defined as $\mathbf{U}_i$. The forget gate controls how much information remains in the current memory cell through a recurrent connection and its corresponding weight matrix is $\mathbf{U}_f$. An output gate controls how much information is used to compute the output activation of the memory block and how much information furthermore flows into the rest of the network. Its weight matrix is denoted by $\mathbf{U}_o$.

Detailed relationships of the entities in the LSTM network can be presented as follows. Recall that the input vector at the input layer at time t is $\mathbf{A}_t^o$, the hidden state vector at time t is $\mathbf{h}_t$, and the memory cell at time t is $\mathbf{c}_t$. Let the operations of dot product and summation of two vectors be denoted by $\cdot$ and $+$, respectively. Let $\sigma(\cdot)$ and $\psi(\cdot)$ denote the *sigmoid* function and the *hyperbolic tangent* function, respectively. So we have the definitions of these functions as

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{3.15}$$

$$\psi(x) = 2\sigma(2x) - 1. \tag{3.16}$$

Based on the above definitions, the output of the forget gate $\mathbf{f}_t$ can be written as $\mathbf{f}_t = \sigma(\mathbf{U}_{fh}\mathbf{h}_{t-1} + \mathbf{U}_{fa}\mathbf{A}_t^o + \mathbf{b}_f)$, where $\mathbf{U}_{fa}$ and $\mathbf{U}_{fh}$ are the weight matrices of $\mathbf{A}_t^o$ and $\mathbf{h}_{t-1}$, respectively. Furthermore, $\mathbf{b}_f$ is the bias of $\mathbf{f}_t$, while σ is a sigmoid function.

Similar to the input gate, the output vector $\mathbf{i}_t$ of the input gate can be given by $\mathbf{i}_t = \sigma(\mathbf{U}_{ih}\mathbf{h}_{t-1} + \mathbf{U}_{ia}\mathbf{A}_t^o + \mathbf{b}_i)$, where $\mathbf{U}_{ia}$ and $\mathbf{U}_{ih}$ are the weight matrix for $\mathbf{A}_t^o$ and $\mathbf{h}_{t-1}$, while $\mathbf{b}_i$ is the corresponding bias. The input activation vector $\mathbf{c'}_t$ of the memory cell can be calculated as $\mathbf{c'}_t = \psi(\mathbf{U}_{ch}\mathbf{h}_{t-1} + \mathbf{U}_{ca}\mathbf{A}_t^o + \mathbf{b}_c)$, where $\mathbf{U}_{ca}$ and $\mathbf{U}_{ch}$ are the weight matrix for $\mathbf{A}_t^o$ and $\mathbf{h}_{t-1}$, $\mathbf{U}_c = [\mathbf{U}_{ch}, \mathbf{U}_{ca}]$, $\mathbf{b}_c$ is the corresponding bias. Here, the memory cell state vector $\mathbf{c}_t$ can be calculated as $\mathbf{c}_t = \mathbf{f}_t \cdot \mathbf{c}_{t-1} + \mathbf{i}_t \cdot \mathbf{c'}_t$, which is the combination of the input activation vector $\mathbf{c'}_t$ and the long memory $\mathbf{c}_{t-1}$.

Now we consider the last gate, i.e. the output gate. We have the calculation of the vector $\mathbf{o}_t$ of the output gate as $\mathbf{o}_t = \sigma(\mathbf{U}_{oa}\mathbf{A}_t^o + \mathbf{U}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o)$, where $\mathbf{U}_{oa}$ and $\mathbf{U}_{oh}$ are the weight matrix for $\mathbf{A}_t^o$ and $\mathbf{h}_{t-1}$ respectively, and $\mathbf{b}_o$ is the corresponding bias. The output vector $\mathbf{h}_t$ of the hidden layer can be expressed as $\mathbf{h}_t = \mathbf{o}_t \cdot \psi(\mathbf{c}_t)$. So the output vector $\mathbf{A}_t$ of the output layer is calculated as

$$\mathbf{A}_t = g(\mathbf{U}_N\mathbf{h}_t), \tag{3.17}$$

where $\mathbf{U}_N$ is the weight matrix of the output layer, $g(\cdot)$ is active function.

Now, we obtain the predicted workload $\mathbf{A}_t$ for the large timescale operation. In Section 3.6.1, we provide detailed operations of LSTM networks, the configuration for the inputs and the networks as well as the performance evaluations. In particular, our predicted mechanism outperforms the benchmark of ARMA. In the next section, based on the large timescale optimization results, we further determine the small timescale resource allocation optimization for each edge node.

3.5 Small Timescale Optimization and Configuration Model

In this section, we aim to minimize the system cost at each time slot. Recall that the large timescale model is proposed to solve problem $\mathbf{P}_{3.2}$, where the optimal value of the

status vector $\mathbf{s}(T) = \mathbf{s}^*(T)$ is determined. Given the solution of $\mathbf{s}^*(T)$, the small timescale model seeks the optimal value $\mathbf{x}(t)$ for problem $\mathbf{P}_{3.3}$. Note that problem $\mathbf{P}_{3.3}$ is an NP-hard problem, so we develop the Lyapunov optimization method to solve this problem in the following.

3.5.1 Overview of Lyapunov Optimization

We firstly provide the brief description of Lyapunov optimization, interested readers can find detailed information in [55]. The dynamic change of the arrival workload in constraint $C3.1$ of problem $\mathbf{P}_{3.3}$ makes the objective function $F_2(x(t), s^*(T))$ hard to be solved. However, Lyapunov optimization method [55] can be used to deal with the dynamic resource allocation problem. In particular, the optimal solution would be obtained by solving a deterministic per-time slot problem with a much lower complexity.

We now formulate the Lyapunov function $L(t)$ as follow:

$$L(t) = \frac{1}{2}\sum_{n=1}^{N} Q_n^2(t). \tag{3.18}$$

The Lyapunov drift $\triangle(t)$ can be written as

$$\triangle(t) = L(t+1) - L(t) = \frac{1}{2}\sum_{n=1}^{N}[Q_n^2(t+1) - Q_n^2(t)]. \tag{3.19}$$

Accordingly, the Lyapunov drift-plus-penalty function can be expressed as

$$\triangle_V(t) = \triangle(t) + VF_2(t), \tag{3.20}$$

where V is a control parameter, while $F_2(t) = F_2(x(t), s^*(T))$ is the objective function of $\mathbf{P}_{3.3}$.

Based on the definition of $Q_n(t)$, we have

$$\begin{aligned} Q_n^2(t+1) \leq [Q_n(t) + A_n(t) - R_n^{tot}(t)\tau]^2 \leq Q_n^2(t) + A_n^2(t) + R_n^{tot\,2}(t)\tau^2 \\ + 2Q_n(t)A_n(t) - 2(Q_n(t) + A_n(t))R_n^{tot}(t)\tau. \end{aligned} \tag{3.21}$$

Combining (3.19) and (3.21) and furthermore employing some mathematical manipulations, we have

$$\triangle(t) \leq \sum_{n=1}^{N} \frac{1}{2}(A_n^2(t) + {R_n^{tot}}^2(t)\tau^2) - A_n(t)R_n^{tot}(t)\tau + Q_n(t)(A_n(t) - R_n^{tot}(t)\tau). \tag{3.22}$$

Substituting $R_n^{tot}(t)$ from (3.9). Since $\log_2(1+x) \leq \frac{x}{\ln 2}$ and $\log_2^2(1+x) \leq \frac{2x}{(\ln 2)^2}$, we have

$$\begin{aligned}\triangle(t) \leq & \sum_{n=1}^{N} \frac{1}{2}[(A_n^{max})^2 + (\sum_{m=1}^{M} \frac{f_{m,n}^{max}\tau}{C_{m,n}} + \frac{B\gamma_n^{max}\tau}{(\ln 2)})^2] - A_n^{max}(\sum_{m=1}^{M} \frac{f_{m,n}^{max}\tau}{C_{m,n}} + \frac{B\gamma_n^{max}\tau}{\ln 2}) \\ & + \sum_{n=1}^{N} Q_n(t)(A_n(t) - R_n^{tot}(t)\tau) \\ \leq & D_1 + \sum_{n=1}^{N} Q_n(t)(A_n(t) - R_n^{tot}(t)\tau)),\end{aligned} \tag{3.23}$$

where A_n^{max} is the maximum arrival workload at edge node n, while γ_n^{max} is the maximum SINR for the edge node n. So D_1 is defined as

$$D_1 = \sum_{n=1}^{N} \frac{1}{2}\left[(A_n^{max})^2 + \left(\sum_{m=1}^{M} \frac{f_{m,n}^{max}\tau}{C_{m,n}} + \frac{B\gamma_n^{max}\tau}{(\ln 2)}\right)^2\right] - A_n^{max}\left(\sum_{m=1}^{M} \frac{f_{m,n}^{max}\tau}{C_{m,n}} + \frac{B\gamma_n^{max}\tau}{\ln 2}\right). \tag{3.24}$$

By adding $VF_2(t)$ to both sides of the above inequality (3.23), we obtain

$$\triangle_V(t) \leq D_1 + VF_2(t) + \sum_{n=1}^{N} Q_n(t)(A_n(t) - R_n^{tot}(t)\tau). \tag{3.25}$$

The proposed resource allocation algorithm for the small timescale model mainly focuses on minimizing the upper bound of $\triangle_V(t)$ on the right side of (3.25) at each time slot t.

3.5.2 Problem Formulation for Computation and Communication

We now formulate the problem for both communication and computation in the small timescale model. In particular, we aim for performing the following goals: *the workload at buffer queue $Q_n(t)$ can be kept at a stable level, while the system cost can also be minimized for each edge node and the cloud server.* Using derivations in section 3.5.1, problem $\mathbf{P}_{3.3}$

can be transformed to

$$\begin{aligned}\mathbf{P}_{3.4} \quad & \min_{x(t)} D_1 + V[\phi_{fe} E_{tot}(x(t)) - \phi_{fs} R_{tot}(x(t))] \\ & + \sum_{n=1}^{N} Q_n(t)(A_n(t) - R_n^{tot}(x(t))\tau) \\ & s.t. \;\; C3.2, C3.3.\end{aligned} \tag{3.26}$$

It is worth mentioning that the objective function of $\mathbf{P}_{3.4}$ is from the right-hand side of (3.25) and all the constraints in $\mathbf{P}_{3.3}$ except the task buffer constraint $C3.1$ are retained in $\mathbf{P}_{3.4}$. We can observe that it is still difficult to determine the optimal solution for $\mathbf{P}_{3.4}$, where we must seek the optimal edge node local computation frequency and the optimal transmit power for offloading. Therefore, the problem $\mathbf{P}_{3.4}$ can be further divided into two sub-problems, i.e. problem for local process speed optimization and problem for offloading power optimization.

Local Process Speed Optimization

In this subsection, we determine the optimal solution for the processing rates of PUs at each edge node. To obtain the optimal local process frequency, we would solve the following sub-problem

$$\begin{aligned}\mathbf{P}_{3.41}: \quad & \min_{f_{m,n}(t)} \sum_{t=t_0}^{t_0+T-1} F_3(t) \\ & s.t. \;\; f_{m,n}^{min} \le f_{m,n}(t) \le f_{m,n}^{max},\end{aligned} \tag{3.27}$$

where $F_3(t) = V \sum_{n=1}^{N} [\phi_{fe} e_n^{local}(t)\tau - \phi_{fs} r_n^{local}(t)\tau] + \sum_{n=1}^{N} Q_n(t)(A_n(t) - r_n^{local}(t)\tau)$. Problem $\mathbf{P}_{3.41}$ is a convex problem as the objective function is convex and the constraints are linear. The optimal solution is given as

$$f_{m,n}^*(t) = \min(f_{m,n}^{max}, \max(\overline{f}_{m,n}(t), f_{m,n}^{min})), \tag{3.28}$$

where $\overline{f}_{m,n}(t) = \sqrt{\frac{V\phi_{fs} + Q_n(t)}{3\phi_{fe} V \epsilon_{m,n} C_{m,n}}}$.

We have the following observations of the relation between $f_{m,n}^*(t)$ and the network

parameters. In (3.28), it is readily observed that $f^*_{m,n}(t)$ is a strictly increasing function with respect to $Q_n(t)$. It means that when $Q_n(t)$ increases, the computational frequency $f^*_{m,n}(t)$ increases to keep the local computing buffer queue at the certain acceptable level. On the other hand, $f^*_{m,n}(t)$ decreases with the increases of $C_{m,n}, \epsilon_{m,n}$ and ϕ_{fe}. In particular, when $C_{m,n}$ and ϵ increase, the system needs a higher computational frequency and/or more energy consumption to process one bit of data. Thus, the processing unit m at edge node n must reduce its computational frequency to consume less power. With the increase of ϕ_{fe}, we have a higher priority on the energy consumption and therefore, the computational frequency $f^*_{m,n}(t)$ must be decreased. Of course, we also observe that $f^*_{m,n}(t)$ is a strictly increasing function of ϕ_{fs}

Let us consider the homogeneous scenario, where we have the same computational energy efficiency coefficient $\epsilon_{m,n}$ and the same number of computation cycles $C_{m,n}$ needed to process one bit of data for each processing unit m at node n. In this scenario, we have the same computational rate for all the units at each node. So the workload is equally assigned to each unit at node n.

FDMA Based Offloading Power Optimization

The task offloading can employ two methods, i.e. FDMA and NOMA. The optimization sub-problem of the FDMA based offloading is formulated in this subsection, while NOMA based offloading is presented in the next section. So the optimization sub-problem under the FDMA setting is given as

$$\begin{aligned} \mathbf{P}_{3.42}: \quad & \min_{p_n(t)} \sum_{t=t_0}^{t_0+T-1} F_4(t) \\ & s.t. \;\; 0 \le p_n(t) \le p_n^{max}, \end{aligned} \tag{3.29}$$

where $F_4(t)$ is defined as

$$F_4(t) = V \sum_{n=1}^{N} \left[\phi_{fe} e_n^{off}(t)\tau - \phi_{fs} r_n^{off}(t)\tau \right] + \sum_{n=1}^{N} Q_n(t) \left[A_n(t) - r_n^{off}(t)\tau \right]. \tag{3.30}$$

Recall that $e_n^{off}(t)$ and $r_n^{off}(t)$ are calculated at (3.4) and (3.5). By substituting $e_n^{off}(t)$ and $r_n^{off}(t)$ into the above equation, $F_4(t)$ can be rewritten as

$$\begin{aligned} F_4(t) = & V\tau\sum_{n=1}^{N}\phi_{fe}(\zeta p_n(t)+p_r)-\phi_{fs}B\log_2(1+\frac{p_n(t)g_n^2(t)}{\sigma_n^2}) \\ & +\sum_{n=1}^{N}Q_n(t)\left[A_n(t)-B\tau\log_2(1+\frac{p_n(t)g_n^2(t)}{\sigma_n^2})\right]. \end{aligned} \tag{3.31}$$

The problem $\mathbf{P}_{3.42}$ is a convex function as its objective function is a linear function of convex terms and the constraint is also linear. We can derive its optimal solution as

$$p_n^*(t) = \min(p_n^{max}, \max(\overline{p}_n(t), 0)), \tag{3.32}$$

where $\overline{p}_n(t) = \frac{(V\phi_s+Q_n(t))B}{\ln 2V\phi_{fe}\zeta} - \frac{\sigma_n^2}{g_n^2(t)}$.

We have the following observations on the optimal transmit power $p_n^*(t)$. The optimal transmit power $p_n^*(t)$ is non-decreasing with respect to queue size $Q_n(t)$. This indicates that when the queue size of node n increases, the offloading power increases in order to achieve a higher offloading rate. As a result, the workload accumulated in the queue is reduced. The transmit power $p_n^*(t)$ also increases with the increase of bandwidth B. Thus, the offloading rate is increased and we can offload more workloads. The transmit power is a decreasing function of the energy weight ϕ_{fe}. In particular, when ϕ_{fe} is higher, we set the higher priority for optimizing the energy consumption, i.e. energy consumption would be reduced.

Offloading Power Optimization for NOMA Method

With NOMA based offloading, the problem $\mathbf{P}_{3.42}$ is reformulated to $\mathbf{P}_{3.43}$. The detailed procedure is presented as follows. Under the NOMA setting, the optimization problem $\mathbf{P}_{3.43}$ can be further divided to K optimization problems with each one denoted as $\mathbf{P}_{3.43k}$ for each

NOMA group k. So problem $\mathbf{P}_{3.43k}$ is expressed as

$$\begin{aligned}\mathbf{P}_{3.43k}: \quad & \min_{p_{k,i}(t)} \sum_{t=t_0}^{t_0+T-1} F_5^k(t) \\ & s.t. \;\; 0 \le p_{k,i}(t) \le p_k^{max}, i = 1, 2,\end{aligned} \tag{3.33}$$

where $F_5^k(t)$ is the objective function, which is given as

$$\begin{aligned}F_5^k(t) = & V\phi_{fe}(\zeta p_{k,1}(t) + \zeta p_{k,2}(t) + 2p_r)\tau - V\phi_{fs}2B\tau[\log_2(1 + \frac{p_{k,1}(t)g_{k,1}^2(t)}{p_{k,2}(t)g_{k,2}^2(t) + \sigma_{k,1}^2}) \\ & + \log_2(1 + \frac{p_{k,2}(t)g_{k,2}^2(t)}{\sigma_{k,2}^2})] + Q_{k,1}(t)[A_{k,1}(t) - 2B\tau\log_2(1 + \frac{p_{k,1}(t)g_{k,1}^2(t)}{p_{k,2}(t)g_{k,2}^2(t) + \sigma_{k,1}^2})] \\ & + Q_{k,2}(t)[A_{k,2}(t) - 2B\tau\log_2(1 + \frac{p_{k,2}(t)g_{k,2}^2(t)}{\sigma_{k,2}^2})].\end{aligned} \tag{3.34}$$

Here, we have two kinds of nodes in a NOMA group, i.e. the strong node and the weak node, which are denoted by the subscripts $_{k,1}$ and $_{k,2}$, respectively. Recall that we use the calculations of the offloading rates and the energy consumption for both strong and weak nodes from (3.6), (3.7) and (3.8).

It is observed that the objective function $F_5^k(t)$ is not a convex function, therefore the optimization problem is a non-convex problem. We exploit the convex relaxation method called SCALE (Successive Convex Approximation for Low Complexity) [43]. In the SCALE method, we use the following observation

$$\alpha\log_2(z) + \beta \le \log_2(1 + z). \tag{3.35}$$

This is tight at $z = \overline{z} \ge 0$, when the approximation coefficients are given as

$$\begin{aligned}\alpha &= \frac{\overline{z}}{1 + \overline{z}}, \\ \beta &= \log_2(1 + \overline{z}) - \frac{\overline{z}}{1 + \overline{z}}\log_2(\overline{z}).\end{aligned} \tag{3.36}$$

We now apply the SCALE method, where we use a logarithmic change of variables $\rho_{k,i}(t) =$

$\ln(p_{k,i}(t))$. Furthermore, the terms $(\alpha_{k,1}, \beta_{k,1})$ and $(\alpha_{k,2}, \beta_{k,2})$ calculated by (3.36) are used for both the strong and weak nodes. After using some simple manipulations, we have the approximation of the $F_5^k(t)$ as

$$\begin{aligned}\tilde{F}_5^k(t) &= V\phi_{fe}[\zeta(\exp(\rho_{k,1}(t)) + \exp(\rho_{k,2}(t))) + 2p_r]\tau + Q_{k,1}(t)A_{k,1}(t) + Q_{k,2}(t)A_{k,2}(t) \\ &+ 2B\tau(V\phi_{fs} + Q_{k,1}(t))\{\alpha_{k,1}\log_2[\exp(-\rho_{k,1}(t))\frac{\sigma_{k,1}^2}{g_{k,1}^2(t)} + \frac{g_{k,2}^2(t)}{g_{k,1}^2(t)}(\exp(\rho_{k,2}(t) - \rho_{k,1}(t))] \\ &- \beta_{k,1}\} + 2B\tau(V\phi_{fs} + Q_{k,2}(t))\{\alpha_{k,2}\log_2[\exp(-\rho_{k,2}(t))\frac{\sigma_{k,2}^2}{g_{k,2}^2(t)}] - \beta_{k,2}\}.\end{aligned} \tag{3.37}$$

So the problem $\mathbf{P}_{3.43k}$ can be transformed into

$$\begin{aligned}\mathbf{P}_{3.44k}: \quad &\min_{\rho_{k,i}(t)} \ \tilde{F}_5^k(t) \\ &s.t.\ \rho_{k,i}(t) \leq \ln(p_k^{max}).\end{aligned} \tag{3.38}$$

Lemma 1. *The problem $\mathbf{P}_{3.44k}$ is a convex problem with the given $\alpha_{k,i}, \beta_{k,i}$.*

Proof. The first part of the objective function $\tilde{F}_5^k(t)$ is evidently convex. The last two terms of the objective function are also convex as they are the log-sum-exp functions. The remaining parts are constant. Therefore, the objective function $\tilde{F}_5^k(t)$ is the summation of all convex terms, which is also convex. Furthermore, the constraint is convex. As a result, the problem considered is convex. □

In the following, we utilize the Lagrangian duality technique to solve the problem $\mathbf{P}_{3.44k}$. We define the Lagrangian function $L(\rho)$ as $L(\rho) = \tilde{F}_5^k(t)$. We firstly solve the stationary condition, i.e. $\partial L(\rho)/\partial\rho_{k,1} = 0$, where $\partial L(\rho)/\partial\rho_{k,1}$ is calculated as

$$\frac{\partial L(\rho)}{\partial(\rho_{k,1})} = V\phi_{fe}\tau\zeta\exp(\rho_{k,1}) - \frac{2B}{\ln 2}\alpha_{k,1}\tau(V\phi_{fs} + Q_{k,1}(t)). \tag{3.39}$$

Then, we transform the result back to the original solution space after solving the fixed-point equation. So the optimal solution for the strong node in group k at time slot t is

given as

$$p_{k,1}^{*}(t) = \min(p_n^{max}, \max(\overline{p}_{k,1}(t), 0)), \tag{3.40}$$

where $\overline{p}_{k,1}(t) = \frac{2B\alpha_{k,1}(V\phi_{fs}+Q_{k,1}(t))}{\ln 2V\phi_{fe}\zeta}$.

We have the following observations for the optimal solution as follows. The optimal transmit power $p_{k,1}^{*}(t)$ for strong node increases with the increase of the rate coefficient ϕ_{fs} and the buffer queue $Q_{k,1}(t)$. This confirms that 1) when the weight of data process rate becomes higher, the edge node increases its offloading power to achieve a higher rate; 2) when the buffer queue $Q_{k,1}(t)$ becomes larger, the edge node also needs to improve its offloading rate to reduce the buffer queue by increasing the transmit power. The optimal transmit power $p_{k,1}^{*}(t)$ decreases with the increase of the weight of transmission power consumption. When ζ becomes larger, i.e. the system consumes more energy for offloading tasks, it needs to reduce the transmit power and allocate more tasks to the local processing than to offloading.

Similarly, we can determine the optimal solution for the weak node in the same manner. We firstly set $\partial L(\rho)/\partial\rho_{k,2} = 0$ and then use some manipulations to obtain the optimal transmit power. The calculations and derivations of the optimal solution are omitted because they can be done in the simple steps. Finally, the optimal transmit power for the weak node in group k at time slot t can be given as

$$p_{k,2}^{*}(t) = \min(p_n^{max}, \max(\overline{p}_{k,2}(t), 0)), \tag{3.41}$$

where $\overline{p}_{k,2}(t) = \frac{-1}{2}(d_3 - \sqrt{d_3^2 + 4d_4})$, $d_3 = \frac{\sigma_{k,1}^2}{g_{k,2}^2(t)} + \frac{d_1 - d_2}{V\phi_{fe}\zeta}$, $d_4 = \frac{d_2\sigma_{k,1}^2}{V\phi_{fe}\zeta g_{k,2}^2(t)}$, $d_1 = \frac{2B}{\ln 2}\alpha_{k,1}(V\phi_{fs} + Q_{k,1}(t))$ and $d_2 = \frac{2B}{\ln 2}\alpha_{k,2}(V\phi_{fs} + Q_{k,2}(t))$. We summarize the procedures of solving problem $\mathbf{P}_{3.44k}$ in Algorithm. 3.2.

Algorithm complexity analysis

For Algorithm. 3.1, the complexity comes from two parts. The first part comes from

estimating the workload at each E-node, while the second part comes from the turn on/off operation performed at each PU. Let N and M denote the number of users and the number of PUs of each user, respectively. Then, based on workload estimation and turn on/off operation, the complexity of Algorithm. 3.1 is $\mathcal{O}(MN)$.

For Algorithm. 3.2, the complexity also comes from two parts. The first part comes from updating the parameters α and β, while the second part comes from calculating the offloading power for each edge node. Let L be the number of iterations required to update the approximation parameters α and β and let N be the number of edge nodes. Then, the complexity of Algorithm. 3.2 is $O(LN)$.

Table 3.3: The SCALE Iterative Algorithm for $\mathbf{P}_{3.44k}$

Algorithm 3.2: The SCALE Iterative Algorithm for $\mathbf{P}_{3.44k}$

1) **Input settings:**
The error tolerance $\xi > 0$, $p_n^{max} > 0$ and ϕ_{fe},ϕ_{fs}, $\alpha_{k,1}^1 = 1$,$\alpha_{k,2}^1 = 1$, $\beta_{k,1}^1 = 0$, $\beta_{k,2}^1 = 0$ and the maximum iteration number I.
2) **Initialization:**
$i = 0$, $p_{k,1}^i(t)$ and $p_{k,2}^i(t)$.
3) **Optimization:**
$i = 1 : I$ obtain the solution $p_{k,1}^i(t)$ by (3.40) and solve $p_{k,2}^i(t)$ by (3.41)
if $p_{k,1}^i(t) - p_{k,1}^{i-1}(t) \leq \xi$ & &$p_{k,2}^i(t) - p_{k,2}^{i-1}(t) \leq \xi$
the optimal $p_{k,1}^*(t)$ and $p_{k,2}^*(t)$ are obtained.
else
update $\alpha_{k,1}^{i+1}$, $\alpha_{k,2}^{i+1}$, $\beta_{k,1}^{i+1}$ and $\beta_{k,2}^{i=1}$ by (3.36) and $i = i + 1$.
4) **Output:**
$\{p_{k,1}^*(t), p_{k,2}^*(t)\}$.

3.6 Numerical Results

3.6.1 Long-Short Term Memory Workload Prediction

The LSTM model performance on traffic prediction is first evaluated using the real traffic dataset [56], which records a total of 91065 user activities and their behaviors between

Jan. 2006 and Jan. 2009. These data are widely used in different cloud communication studies and used as the arrival workload in this study [57]- [58]. The original data contains many features such as user ID, user class, sequence number, etc. Here, the number of active users is used as the number of arrived workloads. Therefore, the raw data is transformed to the number of users arriving in every time slot. We assume that each user represents a workload with 1000-bit data that needs to be processed [44] [47].

For the efficient learning of LSTM, the original data first is normalized based on min-max normalization as follows:

$$\overline{A}_t^o = \frac{A_t^o - A_{min}^o}{A_{max}^o - A_{min}^o}, \tag{3.42}$$

where A_{min}^o and A_{max}^o are the minimum and maximum values of the original data [59]. The LSTM network adopted has one input layer with one input, one hidden layer with 4 LSTM blocks, and one output layer that makes a single-value prediction. We use the LSTM method to predict the data arriving at the edge nodes. Each dataset is divided into two parts, where 67% of the dataset are used for training the LSTM model, and the remaining 37% data are used for testing. We also compare the proposed method with ARMA$(2,1)$ model [60]. The mean absolute performance error (MAPE) is used in this paper for evaluating the prediction errors [61]. The MAPE is the ratio of the error and the true value, which is defined as

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T}\frac{|A_t - \overline{A}_t^o|}{\overline{A}_t^o}. \tag{3.43}$$

Figs. 3.4 and 3.5 illustrate the training outcomes with different methods. For a better observation, we shift the results of the ARMA model [48, 49] and the LSTM model with 1 time slot from the original data. From Fig. 3.4, both ARMA model and LSTM model can well capture the overall trend of the original data. However, Fig. 3.5 with finer granularity indicates that ARMA method does not follow the rapid change of the workload flow as well as LSTM. Thus, the ARMA method has larger prediction errors when comparing to the LSTM mechanism. Figs. 3.6 and 3.7 show the testing results for the different prediction method, where the LSTM method outperforms the ARMA(2,1) method. The MAPE

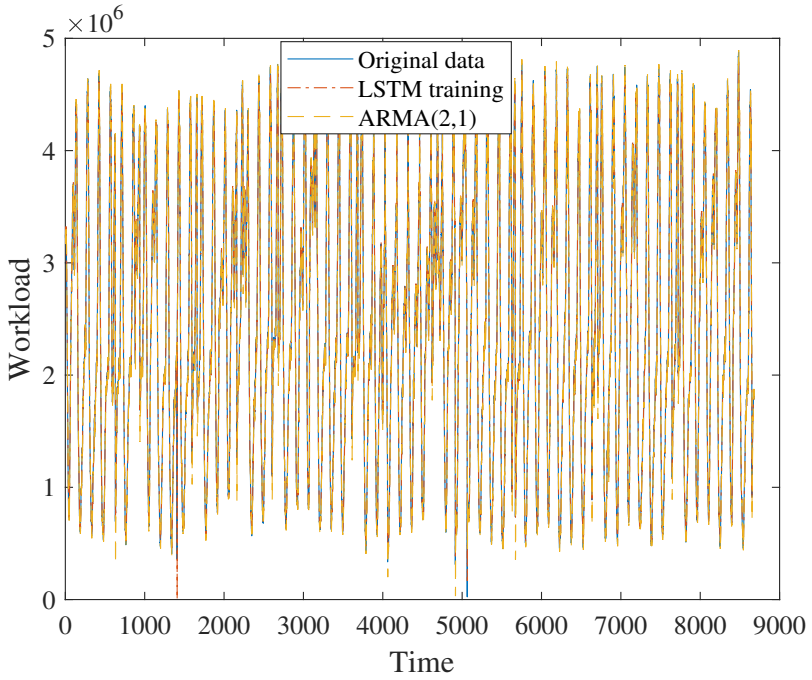

Fig. 3.4: Comparison between LSTM model and ARMA model in training.

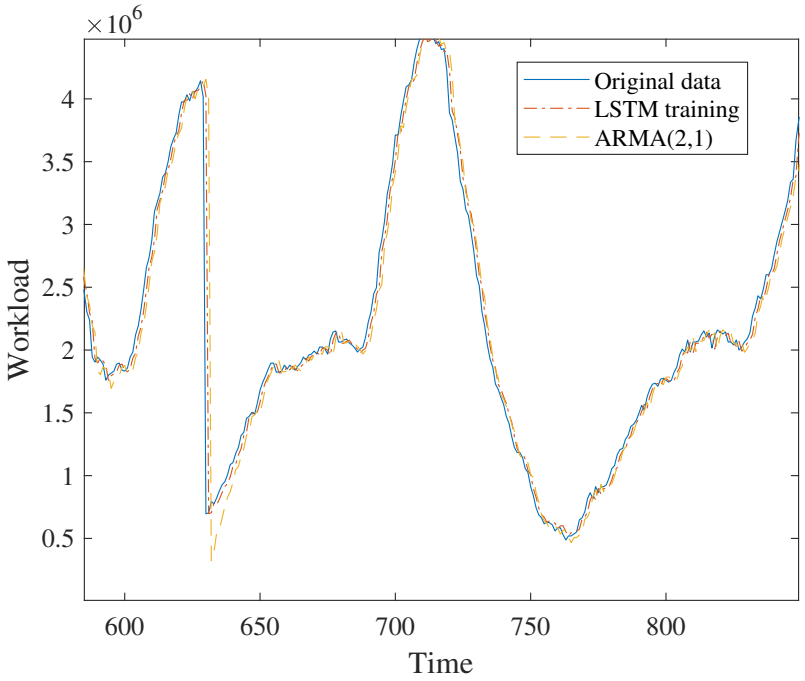

Fig. 3.5: Detailed training part of LSTM model and ARMA model.

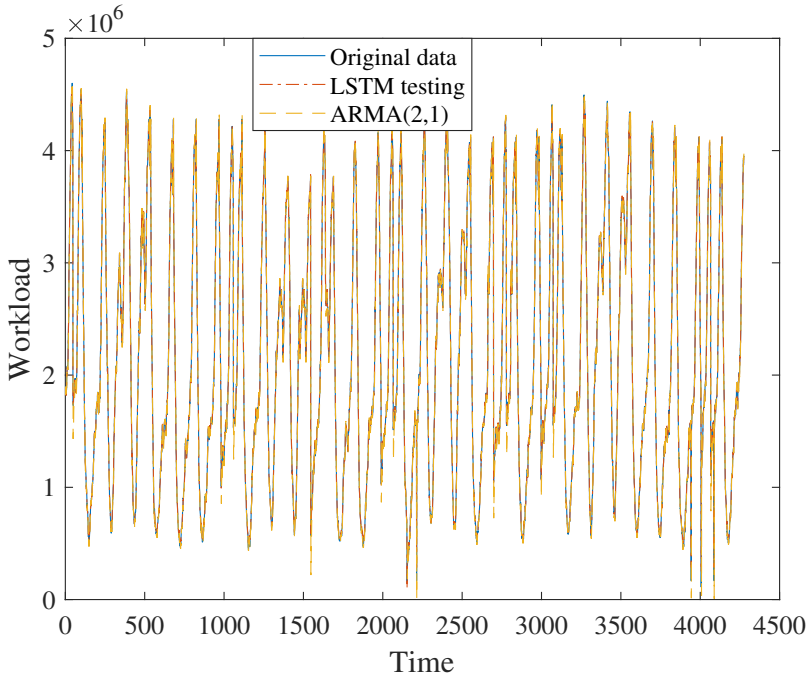

Fig. 3.6: Comparison between LSTM model and ARMA model in testing.

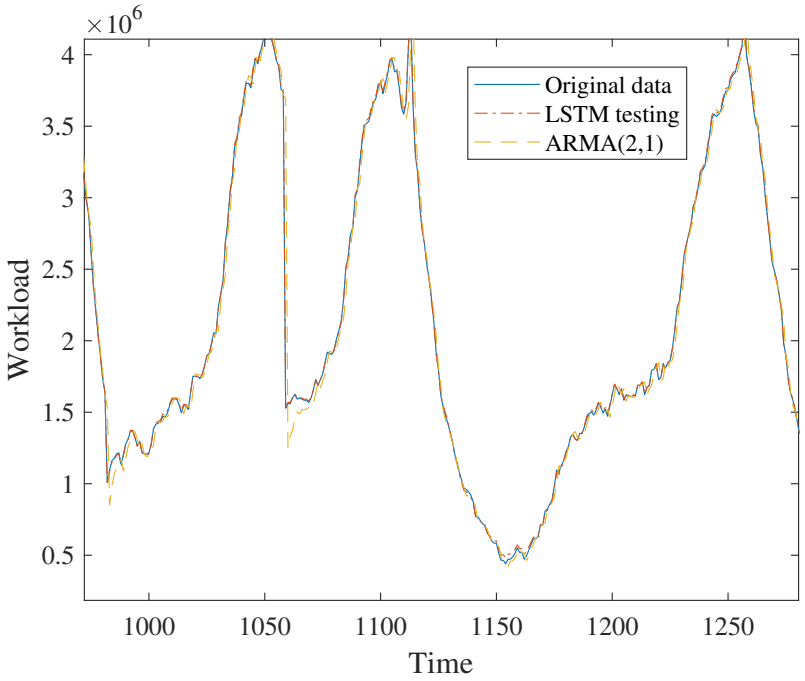

Fig. 3.7: Detailed testing part of LSTM model and ARMA model.

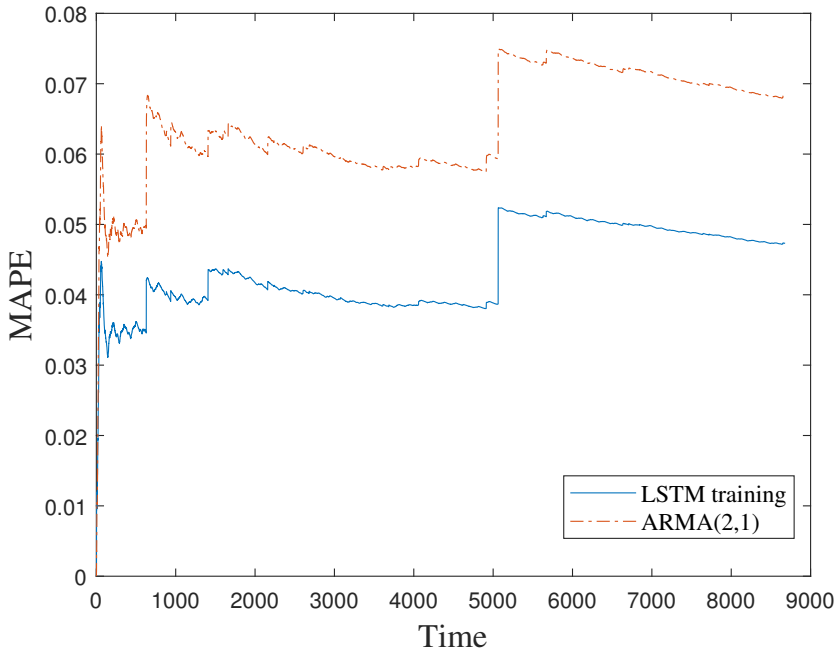

Fig. 3.8: Comparison of the performance of the mean absolute performance error for training part.

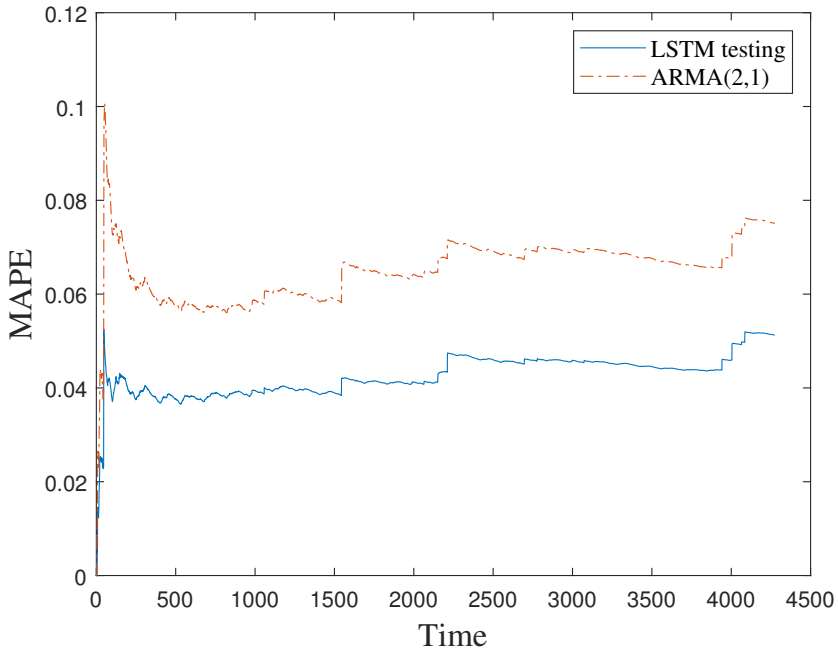

Fig. 3.9: Comparison of the performance of mean absolute error for testing part.

performances for both training part and testing part of the two methods are presented in Figs. 3.8 and 3.9, respectively. We can see that LSTM can achieve a lower error result. Therefore, the LSTM method is used for the prediction of workload flows in the following experiments.

3.6.2 System Cost Optimization

Based on the prediction results, this section gives the performance of the proposed methods. The simulation settings are based on the work in [29], [62]. All the parameters are chosen as follows unless stated otherwise. There are $N = 2$ edge nodes and one centralized server, where each edge node has $M = 10$ processing units. Since the mobility is not considered in this paper, the location for each edge node is fixed during the entire simulation. The offloading transmission power for the communication link between each edge node and the sever is in the range of $\left[p_n^{min}, p_n^{max}\right]$, where $p_n^{min} = 0$ W and $p_n^{max} = 2.5$ W. The channel bandwidth for FDMA is $B = 2$MHz, the local processing capacity for one bit is $C_{m,n} = 1 \times 10^3$ cycles/bit. The computational energy efficiency coefficient is $\epsilon = 1 \times 10^{-27}$, the power weight $\zeta = 2$. The channel between the edge node and the server is modeled as the joint effect of large-scale and small-scale fading, where the channel parameters are given as $g_k^2/\sigma^2 = G_k h_k$, $G_1 = 7$ and $G_2 = 3$. Note that h_k is the unitary Gaussian random variable [29]. The computational capacity of each edge node is set equally in the range of $\left[f_{m,n}^{min}, f_{m,n}^{max}\right]$, where $f_{m,n}^{max} = 10^9$ Hz and $f_{m,n}^{min} = 10^3$ Hz. The circuit power is $p_r = 1$ dBm. To balance the value of throughput and energy, the weights are selected as $\phi_{fe} = 10^6$ and $\phi_{fs} = 0.1$. The results are obtained in different random channel realizations.

The study considers four cases: (1) the proposed PU on/off scheme with FDMA offloading. In this scheme, the data offloading from the edge node to the cloud server adopts the FDMA method based on problem P4.2, thus, it marked as "On/Off + FDMA offloading". (2) The proposed method PU on/off with NOMA offloading "On/Off + NOMA offloading", where the offloading scheme between the edge node and the cloud server is the NOMA method based on the problem P4.3. (3) The benchmark scheme without PU on/off based on FDMA offloading. All the PUs keep in "on" status. The scheme is marked

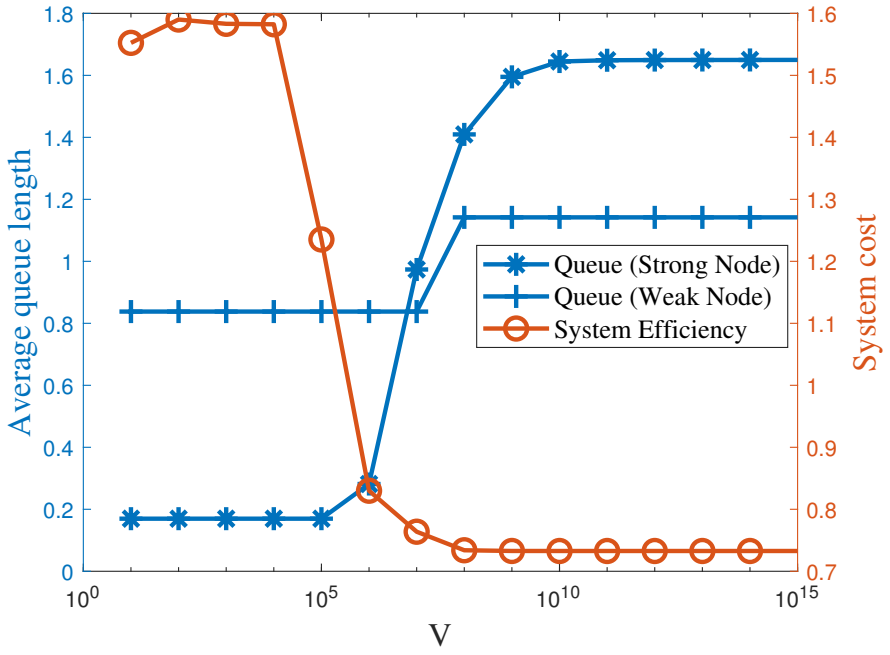

Fig. 3.10: Normalized system cost and average queue length per edge node vs the control parameter V.

as "FDMA offloading" in the figure. (4) The benchmark scheme without the cloud server assistance. So the system can only process the data locally at edge nodes but allows PUs to turn on/off. The method is marked as "On/Off + local computing only". We set the epoch duration at $T = 50\tau$. All the simulation results are averaged based on 100 independent runs. We note that only the last scheme is the local processing, while the first three schemes have both local processing and cloud processing. To avoid any confusion, we firstly confirm that the term "offloading" does not mean that all the tasks must be offloaded to the cloud. It only means that we use offloading mechanisms, like NOMA or FDMA, to offload partial, or complete, or no tasks to the cloud, while the remaining tasks can be still processed at the local PUs. So we always keep the PUs in "on" status in the third case of "without PU on/off" due to the following reason. In this case, we do not use the large timescale model to predict the workload flow as well as use the turning-on/off algorithm. So we keep the PUs "on" to serve the high demand of workload as we consider the dynamic change of the workload flow.

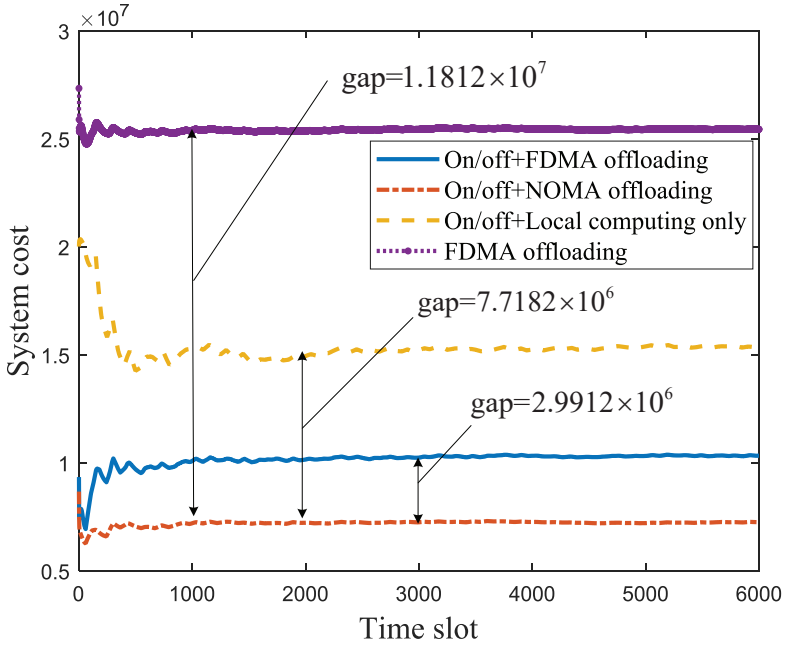

Fig. 3.11: Comparison of system cost over total duration.

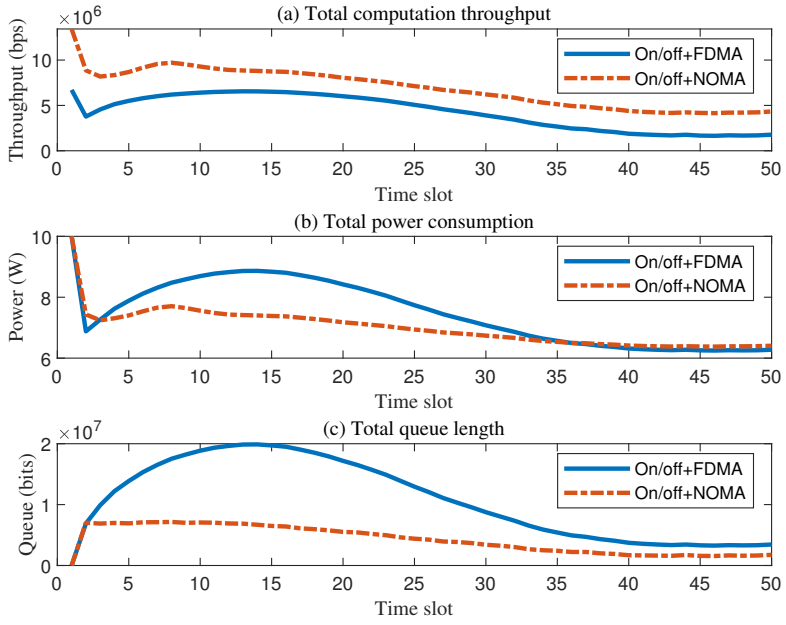

Fig. 3.12: "On/Off+FDMA" scheme vs. "On/Off+NOMA" scheme.

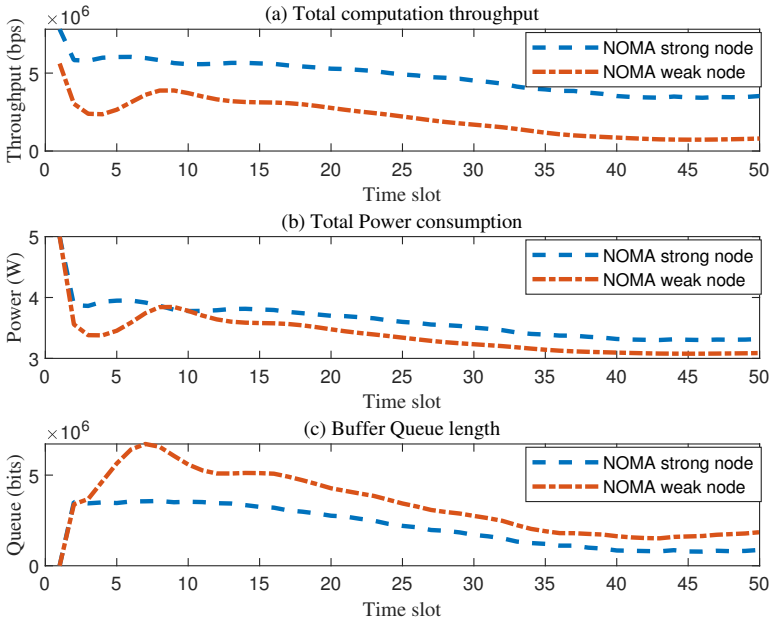

Fig. 3.13: Strong node vs. weak node in the "On/Off+NOMA" scheme.

The relationship between the system cost/average queue length of the task buffers and the control parameter V is presented in Fig. 3.10 for the proposed "On/Off + NOMA offloading" scheme. For a better illustration, the values of both system cost and queue length are normalized. The system cost firstly maintains at a high level when $V \leq 10^4$, it then decreases with the increase of V. When $V \geq 10^8$, the system cost keeps at a low level. On the other hand, the lengths of buffer queues for both the strong edge node and weak edge node are small when the system cost is high. It then increases when the system cost drops down. The reason is that the parameter V controls the tradeoff between the original cost function $F(t)$ and the buffer queue stability in the Lyapunov drift-plus-penalty function in (3.20). Therefore, by increasing the value of V, the system gives more priority to reduce the system cost and less priority to serve the buffered data. The optimal solution of V to achieve a good tradeoff between the system cost and buffer queue stability is around $V = 10^7$ based on the simulation setting. Thus, we choose the control parameter as $V = 10^7$ in the following simulation results.

We proceed to show the performance of different schemes by using the predicted workloads and parameters setting found above. Fig. 3.11 shows the system cost for four different schemes defined above. By properly turning on/off processing units, the proposed scheme achieves a much lower total system cost than all other schemes by integrating NOMA, computation offloading and PU On/Off altogether. The only scheme "FDMA offloading" that does not use PU On/Off has a significantly higher system cost than others. Furthermore, computation offloading considerably reduces the system cost. We also observe that all the schemes converge after 500 time slots. This observation confirms the stability and convergence of the proposed methods. Although the workload flows dynamically change over time, all the edge nodes effectively allocate the transmit power for communication and also adjust the processing unit frequency for computation. Therefore, the model can adaptively achieve the optimal and stable system efficiency even with the dynamic traffic behaviors.

We further compare the performance for the two best schemes, namely "On/Off + NOMA offloading" and "On/Off + FDMA offloading" in Fig. 3.12. Fig. 3.12(a) shows that the NOMA offloading based scheme achieves a higher computation throughput than the FDMA offloading based scheme. Further in Fig. 3.12(b), the NOMA offloading based scheme consumes less power than the FDMA offloading based scheme. Combining the two figures, it demonstrates that NOMA based scheme attains a much higher efficiency in energy usage by consuming less energy but gaining a higher computation rate. A higher computation throughput leads to less queued data in the buffer, which is verified in Fig. 3.12(c).

Another thing worth noticing is that the curves of the total computation throughput, the power consumption and the buffer queue length for the FDMA method in Fig. 3.12 first go up and then decline in the main observation. However, we would see the fluctuations at the very beginning. This transient behavior is explained as follows. The initial value is high based on the initial parameter setting so that the system can achieve high throughput and low power consumption. There are not too many workloads needed to be processed at the beginning of large timescale. Thus, the system only adjusts parameters for the throughput and power, which can help to keep the higher efficiency. We now explain the

system behavior in the main observation. At the beginning of each epoch, the workload is firstly buffered in the queue due to insufficient computation throughput of each node. Due to the increase of the queue length, the system adjusts both the offloading rate and local process speed so that the queued data can be served. At the end of the epoch, with the small queue length, the system can keep the computation throughput stabilized at a lower optimal level and maintains the minimum system cost. On the other hand, we observe that the curves of the NOMA based scheme keep decreasing. At the beginning of the epoch, the computation throughput is high enough to sufficiently serve the arrival and queued data so that queue size does not build up. Ultimately the NOMA and FDMA based schemes converge to the similar queue level and similar power consumption level.

The performance of the strong node and weak node for the NOMA based scheme is provided in Fig. 3.13. The weak edge node in the NOMA method has a lower throughput and also a lower energy consumption than the strong node. The proposed method aims to minimize the overall system cost. The transmit power for the nodes with different channel quality is adjusted by slightly increasing the power of the strong node and decreasing the weak one. We also have the transient duration at the beginning, which is similar to the observation in Fig. 3.12. Thus, the system can achieve an overall low system cost but still guarantee the weak node's performance. Therefore, the proposed method can dynamically adjust the resource allocated to each node to achieve the optimal overall system efficiency and meet the performance requirements of each node.

3.7 Chapter Conclusions

This chapter considers a hierarchical architecture that consists of IoT sensor layer, edge computing layer, and cloud server layer. A twin-timescale optimization model was developed to manage the workload offloading in the system to optimize the trade off between the power consumption and overall computation throughput. In the large timescale model, based on predicted workload, the scheme decides how to turn on or turn off processing unit in order to save energy. In the small timescale model, a Lyapunov optimization method was designed to allocate the offloading power and to determine the local process speed for each

processing unit. Simulation results reveal that the proposed method can greatly improve system performance by saving energy costs and achieving a high processing rate.

CHAPTER 4

Secure and Energy-Efficient Offloading and Resource Allocation in a NOMA-Based MEC Network

4.1 Introduction

In the previous chapters, we investigated EE improvement by NOMA and prediction-based server coordination methods from different time scales. Besides EE, secure communications are also critical in the 5G wireless networks as communication environments become increasingly complicated, and both the security and privacy of user information are put at risks [63]. To this end, PLS has received significant attention in recent years. This is because it can achieve secure communications without extra overhead caused by protecting the security key [50]. However, the secrecy rate achieved by the mutual information difference between the legitimate user and the eavesdropper is limited as it depends on the difference between the channel condition from the base station to the legitimate user and that from the base station to the eavesdroppers [64]. NOMA has been viewed as one potential mechanism for improving the achievable secure rate at the receiver side to make up for this oversight. For this reason, PLS in NOMA-assisted MEC networks has been the subject of considerable research [11]. In this chapter, we will begin to take a look at the security issues in MEC-enabled IoT networks with NOMA.

The joint consideration of PLS in the NOMA assisted MEC network was studied in [8]-[10]. Most of the existing works on NOMA-assisted MEC with external eavesdroppers typically focus on the performance evaluation in the scenarios where either channel conditions or required tasks remain constant. Such an assumption makes the analysis on the computation offloading and resource allocation more tractable. However, in a dynamic environment, the dynamic behaviors of the workload arrivals and fading channels impact the overall system performance. Thus the system design that focuses on the short term performance may not

work well from the long term perspective. Towards that, a stochastic task offloading model and resource allocation strategy should be adopted [65]. In this work, we integrate PLS and study the long-term EE performance in a NOMA-enabled MEC network. By incorporating the statistical behaviors of the channel states and task arrivals, we formulate a stochastic optimization problem to maximize the long-term average EE subject to multiple constraints including task queue stability, maximum available power, and peak CPU-cycle frequency. An energy-efficient offloading and resource allocation method based on Lyapunov optimization is proposed.

4.2 System Model

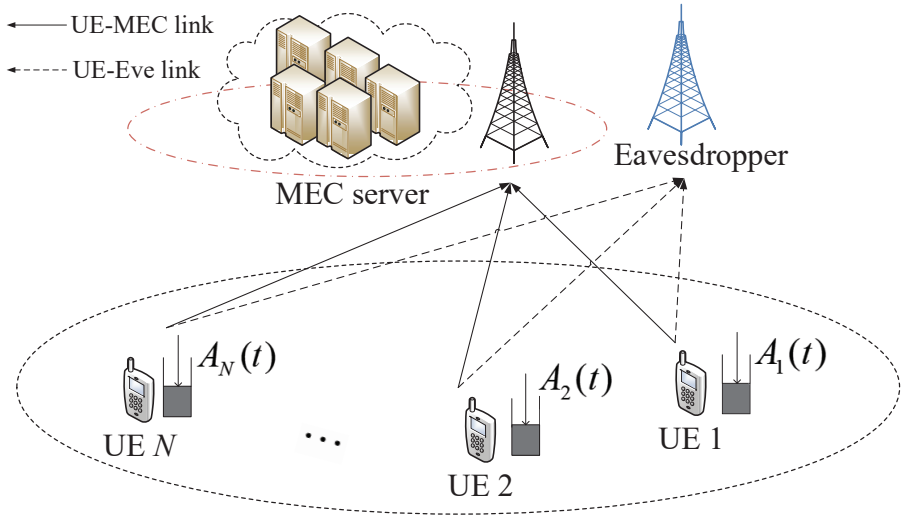

Fig. 4.1: System Model.

In Fig. 4.1, an uplink NOMA communication system is considered, which consists of N user equipments (UEs), one access point (AP) with the MEC server, and one external eavesdropper (Eve) near the AP. All the UEs can offload their computation tasks to the MEC while the external eavesdropper intends to intercept the confidential information. The arrival task of user n at time slot t is denoted as $A_n(t)$. Note that the prior statistical information of $A_n(t)$ is not required and it could be difficult to obtain in the practical systems.

We focus on a data-partition-oriented computation task model. A partial offloading scheme is used, i.e., part of the task is processed locally and the remaining part of the data can be offloaded to the remote server for processing. For each UE, local computing and task offloading can be executed simultaneously.

Assuming that each UE has buffering ability, where the arrived but not yet processed data can be queued for the next time slot. Let $Q_n(t)$ be the queue backlog of UE n, and its evolution equation can be expressed as

$$Q_n(t+1) = \max\{Q_n(t) - R_n^{tot}(t)\tau, 0\} + A_n(t), \tag{4.1}$$

where $R_n^{tot}(t) = R_n^{off}(t) + R_n^{loc}(t)$ is the total computing rate of UE n at time slot t, $R_n^{off}(t)$ and $R_n^{loc}(t)$ are secure offloading rate and local task processing rate, respectively. τ is time duration of each slot.

4.2.1 Local Computing Model

Let $f_n(t)$ denote the local CPU-cycle frequency of UE n, which cannot exceed its maximum value $f_{\max}$. Let C_n be the computation intensity (in CPU cycles per bit). Thus, the local task processing rate can be expressed as $R_n^{loc}(t) = f_n(t)/C_n$. We use the widely adopted model $P_n^{loc}(t) = \kappa_n f_n^3(t)$ to calculate the local computing power consumption of UE n, where κ_n is the energy coefficient and its value depends on the chip architecture [66].

4.2.2 Task Offloading Model

The independent and identically distributed (i.i.d) frequency-flat block fading channel model is adopted, i.e., the channel remains static within each time slot but varies across different time slots. The small-scale fading coefficients from UE n to the MEC server and to the Eve are denoted as $H_{b,n}(t)$ and $H_{e,n}(t)$, respectively. Both are assumed to be exponential distributed with unit mean [44]. Thus, the channel power gain from UE n to the MEC is given as $h_{i,n}(t) = H_{i,n}(t) g_0 (d_0/d_{i,n})^\theta$, $i \in \{b, e\}$, where g_0 is the path-loss constant, θ is the path-loss exponent, d_0 is the reference distance, and $d_{i,n}$ is the distance from UE n to

receiver. Furthermore, to improve the spectrum efficiency, NOMA is applied on the uplink access for offloading. We assume that $h_{b,1} \leq h_{b,2} \leq \cdots \leq h_{b,N}$ and $h_{e,1} \leq h_{e,2} \leq \cdots \leq h_{e,N}$. Using SIC at the receiver side, the achievable secure offloading rate at UE n can be given by

$$R_n^{off}(t) = [B\log_2(1+\gamma_{b,n}) - B\log_2(1+\gamma_{e,n})]^+, \tag{4.2}$$

where B is the bandwidth allocated to each UE, $\gamma_{b,n} = \frac{p_{n(t)}h_{b,n}(t)}{\sum_{i=1}^{n-1} p_i(t)h_{b,i}(t)+\sigma_{b,n}^2}$ and $\gamma_{e,n} = \frac{p_n(t)h_{e,n}(t)}{\sum_{i=1}^{n-1} p_i h_{e,i}(t)+\sigma_{e,n}^2}$ are the SINRs received by the MEC server and the Eve respectively. $p_n(t)$ is the transmit power of UE n, $\sigma_{b,n}^2$ and $\sigma_{e,n}^2$ are the background noise variances at the MEC and the Eve respectively. $[x]^+ = \max(x, 0)$. The power consumption for offloading can be expressed as

$$P_n^{off}(t) = \zeta p_n(t) + p_r, \tag{4.3}$$

where ζ is the amplifier coefficient and p_r is the constant circuit power consumption.

4.3 Dynamic Task Offloading and Resource Allocation

4.3.1 Problem Formulation

EE is defined as the ratio of the number of long term total computed bits achieved by all the UEs to the total energy consumption (in unit bits/Joule) [67],

$$\eta(t) = \frac{\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} R_{tot}(t)\tau\right]}{\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} P_{tot}(t)\tau\right]} = \frac{\overline{R}_{tot}\tau}{\overline{P}_{tot}\tau}, \tag{4.4}$$

where $R_{tot}(t) = \sum_{n=1}^{N} R_n^{tot}(t)$ and $P_{tot}(t) = \sum_{n=1}^{N} P_n^{off}(t) + P_n^{loc}(t)$ are the total achievable rate and consumed power by all the users at t.

This work aims to maximize the long-term average EE for all the UEs under the constraints of resource limitations while guaranteeing the average queuing length stability.

Therefore, the problem is formulated as

$$\mathbf{P}_{4.0} \quad \max_{f_n(t),p_n(t)} \eta$$

$$s.t. \quad P_n^{tot}(t) \leq P_{\max}, \tag{4.5a}$$

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}[|\ \overline{Q}_n(t)|] = 0, \tag{4.5b}$$

$$f_n(t) \leq f_{\max}, \tag{4.5c}$$

$$0 \leq p_n(t), \tag{4.5d}$$

where $\overline{Q}_n(t)$ is the average queue length of UE n. The constraint (4.5a) indicates that the total power consumed by UE at time slot t should not exceed the maximum allowable power $P_{\max}$. (4.5b) requires the task buffers to be mean rate stable, which also ensures that all the arrived computation tasks can be processed within a finite delay. (4.5c) is the range of local computing frequency, and (4.5d) denotes the transmit power of each UE should not be negative.

4.3.2 Problem Transformation Using Lyapunov Optimization

The problem $\mathbf{P}_{4.0}$ is a non-convex problem, which is difficult to be solved due the fractional structure of the objective function and the long term queue constraint (4.5b). By introducing a new parameter $\eta^*(t) = \frac{\sum_{i=0}^{t-1} R_{tot}(i)\tau}{\sum_{i=0}^{t-1} P_{tot}(i)\tau}$ [67], the problem can be transformed to $\mathbf{P}_{4.1}$, which can be solved in an alternating way.

$$\mathbf{P}_{4.1} : \max\nolimits_{f_n(t),p_n(t)} \overline{R}_{tot}(t)\tau - \eta^*(t)\overline{P}_{tot}(t)\tau$$

$$s.t. \quad (4.5a) - (4.5d).$$

Note that $\eta^*(t)$ is a parameter that depends on the resource allocation strategy before t-th time block [67]. In the following, the Lyapunov optimization is introduced to tackle the task queue stability constraint.

To stabilize the task queues, the quadratic Lyapunov function is first defined as $L(\mathbf{Q}(t))$

Algorithm 4 Dynamic Resource Allocation Algorithm

1: At the beginning of the tth time slot, obtain $\{Q_n(t)\}$, $\{A_n(t)\}$.
2: Determine $\mathbf{f}(t)$ and $\mathbf{p}(t)$ by solving

$$\mathbf{P}_{4.2} \max_{f_n(t),p_n(t)} \sum_{n=1}^{N}\{Q_n(t)(R_n^{tot}(t)\tau - A_n(t))\} + V\sum_{n=1}^{N}[R_n^{tot}(t)\tau - \eta^*(t)P_n^{tot}(t)\tau]$$
$$s.t. \quad (4.5a), (4.5c), (4.5d)$$

3: Update $\{Q_n(t)\}$ and set $t = t+1$. Go back to step 1.

$\triangleq \frac{1}{2}\sum_{n=1}^{N} Q_n^2(t)$ [55]. Next, the one-step conditional Lyapunov drift function is introduced to push the quadratic Lyapunov function towards a bounded level.

$$\Delta(\mathbf{Q}(t)) \triangleq \mathbb{E}[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t))|\mathbf{Q}(t)]. \tag{4.7}$$

By incorporating queue stability, the Lyapunov drift-plus-penalty function is defined as

$$\Delta_V(\mathbf{Q}(t)) = -\Delta(\mathbf{Q}(t)) + V[R_{tot}(t)\tau - \eta^*(t)P_{tot}(t)\tau], \tag{4.8}$$

where V is a control parameter to control the tradeoff between the queue length and system EE. The minus sign is used to maximize EE and to minimize the queue length bound. For an arbitrary feasible resource allocation decision that is applicable in all the time slots, the drift-plus-penalty function $\Delta_V(\mathbf{Q}(t))$ satisfies

$$\Delta_V(\mathbf{Q}(t)) \geq -C + \sum_{n=1}^{N}\mathbb{E}\{Q_n(t)(R_n^{tot}(t)\tau - A_n(t))\} + V\sum_{n=1}^{N}[R_n^{tot}(t)\tau - \eta^*(t)P_n^{tot}(t)\tau], \tag{4.9}$$

where $C = \frac{1}{2}\sum_{u=1}^{U}(R_n^{\max 2}\tau^2 + A_n^{\max 2})$, $R_n^{\max}$ and $A_n^{\max}$ are the maximum achievable computing rate and the maximum arrival workload, respectively.

Thus, $\mathbf{P}_{4.1}$ is converted to a series of per-time-slot deterministic optimization problem $\mathbf{P}_{4.2}$, which needs to be solved at each time slot and is given as in **Algorithm 4**.

In $\mathbf{P}_{4.2}$, $\mathbf{f}(t)$ and $\mathbf{p}(t)$ can be decoupled with each other in both the objective function and the constraints. Thus, the problem $\mathbf{P}_{4.2}$ can be decomposed into two sub-problems,

namely the optimal CPU-cycle frequency scheduling sub-problem and the optimal transmit power allocation sub-problem, which can be solved alternately in the following.

Optimal CPU-Cycle Frequencies Scheduling: The optimal CPU-cycle frequencies $\mathbf{f}(t)$ can be obtained by

$$\mathbf{P}_{4.21} \max_{0\leq f_n(t)\leq f_{\max}} \sum_{n=1}^{N}(Q_n(t)+V)(R_n^{off}(t)+f_n(t)/C_n) - V\eta^*(t)(\kappa_n f_n^3(t)+p_r+\zeta p_n(t))$$
$$s.t.\ \ \kappa_n f_n^3(t) \leq P_{\max} - P_n^{off}. \tag{4.10}$$

Since the objective function of $\mathbf{P}_{4.21}$ and the constraints are convex with respect to $f_n(t)$, the optimal $f_n(t)$ can be given as

$$f_n^* = \left[\sqrt{\frac{(V+Q_n(t))}{3V\eta\kappa_n C_n}}\right]_0^{\overline{f}_{\max}}, \tag{4.11}$$

where $\overline{f}_{\max} = \min\{f_{\max}, \sqrt[3]{(P_n^{\max}-\zeta p_n - p_r)/\kappa_n}\}$ is the upper bound of the frequency.

Optimal Transmit Power Allocation: For the transmission power allocation optimization, the problem $\mathbf{P}_{4.22}$ is transformed into

$$\mathbf{P}_{4.22}\ \max_{p_n(t)} \sum_{n=1}^{N} B\ln 2(Q_n(t)+V)[\ln(\sum_{i=1}^{n} p_i(t)h_{b,i}^2+\sigma_{b,n}^2) - \ln(\sum_{i=1}^{n-1} p_i(t)h_{b,i}^2+\sigma_{b,n}^2)$$
$$-\ln(\sum_{i=1}^{n} p_i h_{e,i}^2+\sigma_{e,n}^2) + \ln(\sum_{i=1}^{n-1} p_i h_{e,i}^2+\sigma_{e,n}^2) + \frac{f_n}{B\ln 2 C_n}] - V\eta^*(t)(\zeta p_n + p_r + \kappa_n f_n^3)$$
$$s.t.\ \ 0 \leq p_n(t) \leq (P_{\max} - p_r - \kappa_n f_n^3)/\zeta. \tag{4.12}$$

The minus logarithmic terms make the objective function not convex, which is addressed by Lemma 1 introduced in the following.

Lemma 1: By introducing the function $\phi(y) = -yx + \ln y + 1$, $\forall x > 0$, one has

$$-\ln x = \max_{y>0} \phi(y). \tag{4.13}$$

The optimal solution can be achieved at $y = 1/x$. The upper bound can be given by using

Lemma 1 as $\phi(y)$ [68]. By setting $x_{b,n} = \sum_{i=1}^{n-1} p_i(t)h_{b,i}^2 + \sigma_{b,n}^2$ and $x_{e,n} = \sum_{i=1}^{n} p_i(t)h_{e,i}^2 + \sigma_{e,n}^2$, one has

$$\begin{aligned}\mathbf{P}_{4.23} \quad & \max_{p_n(t), y_{b,n}, y_{e,n}} \sum_{n=1}^{N} B\ln 2(Q_n(t)+V)[\ln(\sum_{i=1}^{n} p_i(t)h_{b,i}^2 \\ & + \sigma_{b,n}^2) + \phi_{b,n}(y_{b,n}) + \phi_{e,n}(y_{e,n}) + \ln(\sum_{i=1}^{n-1} p_i(t)h_{e,i}^2 + \sigma_{e,n}^2) \\ & + \frac{f_n}{B\ln 2 C_n}] - V\eta^*(t)(\zeta p_n(t) + p_r + \kappa_n f_n^3) - Q_n(t)A_n(t) \\ & s.t. \;\; 0 \le p_n(t) \le (P_{\max} - p_r - \kappa_n f_n^3)/\zeta, \end{aligned} \tag{4.14}$$

where

$$\phi_{b,n}(y_{b,n}) = -y_{b,n}(\sum_{i=1}^{n-1} p_i(t)h_{b,i}^2 + \sigma_{b,n}^2) + \ln y_{b,n} + 1, \tag{4.15}$$

and

$$\phi_{e,n}(y_{e,n}) = -y_{e,n}(\sum_{i=1}^{n} p_i(t)h_{e,i}^2 + \sigma_{e,n}^2) + \ln y_{e,n} + 1. \tag{4.16}$$

The problem $\mathbf{P}_{4.23}$ is a convex problem with respect to both $p_n(t)$ and $y_{b,n}, y_{e,n}$. It can be solved by using a standard convex optimization tool. After we obtain $p_n^*(t)$, the values of $y_{b,n}^*$ and $y_{e,n}^*$ can be respectively given by $y_{b,n}^* = (\sum_{i=1}^{n-1} p_i^*(t)h_{b,i}^2 + \sigma_{b,n}^2)^{-1}$ and $y_{e,n}^* = (\sum_{i=1}^{n} p_i^*(t)h_{e,i}^2 + \sigma_{e,n}^2)^{-1}$. By alternately updating $p_n(t)$ and $y_{b,n}, y_{e,n}$, the optimal solutions of $\mathbf{P}_{4.23}$ can be achieved at convergence.

Remark 4.1: To obtain fundamental and insightful understanding of the offloading power allocation for a multi-user NOMA assisted secure MEC system, we consider a special

case with two UEs [69]. The problem with respect to p_n is given as

$$\begin{aligned}\mathbf{P}_{4.24} \max_{p_1(t),p_2(t)} & B\ln 2(V+Q_2(t))[\ln(p_2(t)h_{b,2}^2+p_1(t)h_{b,1}^2+\sigma_{b,2}^2)\\ & -\ln(p_1(t)h_{b,1}^2+\sigma_{b,2}^2)-y_{b2}(p_1(t)h_{b,1}^2+\sigma_{b,2}^2)\\ & +\ln y_{b2}+1+\ln(p_1(t)h_{e,1}^2+\sigma_{e,2}^2)+\frac{f_2}{C_2B\ln 2}]\\ & +B\ln 2(V+Q_1(t))[\ln(\sigma_{b,1}^2+p_1(t)h_{b,1}^2)-\ln\sigma_{b,1}^2\\ & -y_{e1}(\sigma_{e,1}^2+p_1(t)h_{e,1}^2)+\ln y_{e1}+1+\ln\sigma_{e,1}^2\\ & +\frac{f_1}{C_1B\ln 2}]-V\eta(\zeta(p_2(t)+p_1(t))+2p_r+\kappa_n({f_n}^3)\\ s.t.\ \ & 0\le p_n(t)\le(P_{max}-p_r-\kappa_n f_n^3)/\zeta. \end{aligned} \tag{4.17}$$

$\mathbf{P}_{4.24}$ is a convex problem with respect to $p_1(t)$ and $p_2(t)$, and the optimal solutions are given as

$$p_1^*(t)=\frac{-b_1\pm\sqrt{b_1^2-4b_2}}{2}, \tag{4.18}$$

and

$$p_2^*(t)=\frac{1}{(\frac{V\eta\zeta}{B\ln 2(V+Q_2(t))}+y_{e2}h_{e,2}^2)}-\frac{p_1h_{b,1}^2}{h_{b,2}^2}-\frac{\sigma_{b,2}^2}{h_{b,2}^2}, \tag{4.19}$$

where

$$\begin{aligned}a_1=&\frac{V\eta\zeta}{B\ln 2}+(V+Q_2(t))(y_{b2}h_{b,1}^2+y_{e2}h_{e,1}^2)+(V+Q_1(t))y_{e1}h_{e,1}^2\\ &-\frac{(V+Q_2(t))h_{b,1}^2(\frac{V\eta\zeta}{B\ln 2(V+Q_2(t))}+y_{e2}h_{e,2}^2)}{h_{b,2}^2},\end{aligned} \tag{4.20}$$

$$b_1=(\sigma_{b,1}^2/h_{b,1}^2+\sigma_{e,2}^2/h_{e,1}^2-\frac{(V+Q_1(t))}{a1}-\frac{(V+Q_2(t))}{a1}), \tag{4.21}$$

and

$$b_2=\frac{\sigma_{e,2}^2\sigma_{b,1}^2}{h_{e,1}^2h_{b,1}^2}-\frac{(V+Q_2(t))}{a1}\sigma_{b,1}^2/h_{b,1}^2-\frac{(V+Q_1(t))}{a1}\sigma_{e,2}^2/h_{e,1}^2. \tag{4.22}$$

4.4 Simulation Results

In this section, simulation results are provided to evaluate the proposed algorithm. The

simulation settings are based on the works in [66], [69]. We consider the configuration with 2 UEs, which can be readily extended to a more general case. The system bandwidth for computation offloading is set as $B = 1$ MHz, the time slot duration is $\tau = 1$ sec, path-loss exponent is $\theta = 4$, the noise variance is $\sigma_{i,j} = -60$ dBm, where $i \in \{b, e\}, j \in \{1, 2\}$. The size of the arrival workload $A_n(t)$ is uniformly distributed within $[1, 2] \times 10^6$ bits [70]. Other parameter settings include the reference distance $d_0 = 1$ m, $g_0 = -40$ dB, $d_{b,1} = 80$ m, $d_{b,2} = 40$ m, $d_{e,1} = 120$ m, $d_{e,2} = 80$ m. $\kappa_n = 10^{-28}$, $P_{\max} = 2$ W, $f_{\max} = 2.15$ GHz, $C_n = 737.5$ cycles/bit, the amplifier coefficient $\zeta = 1$, and the control parameter $V = 10^7$. The numerical results are obtained by averaging over 1000 random channel realizations. We consider two more cases as the benchmark schemes to compare with our proposed algorithm. In the first benchmark scheme, marked as "Full offloading", all the tasks are offloaded to the MEC server and there is no local computation at all. The second benchmark [69] is marked as "Eve fully decode", in which the Eve can correctly decode other users' information. This provides a worst-case scenario for comparison.

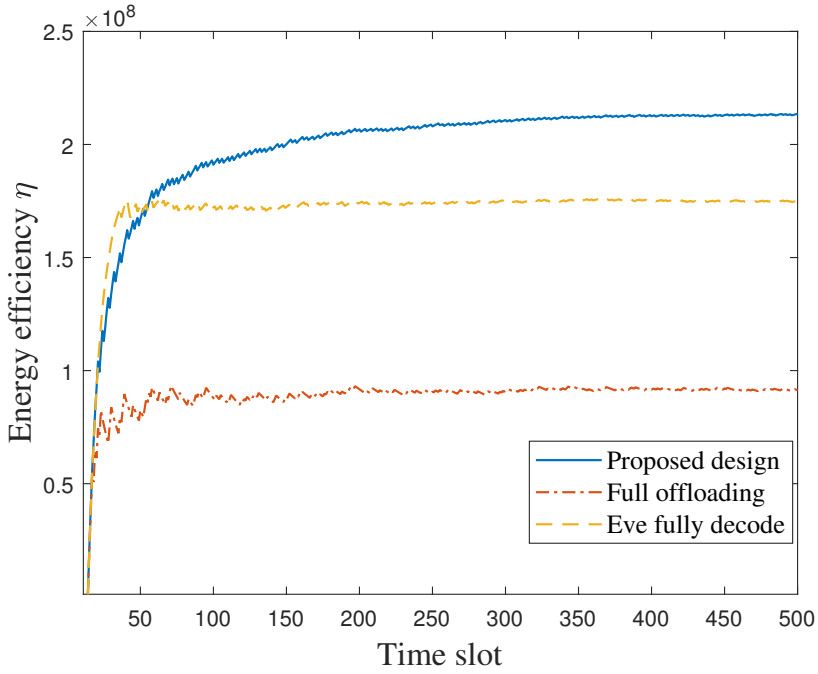

Fig. 4.2: System energy efficiency.

The performance of the system EE vs time is presented in Fig. 4.2. We can see that the proposed method can achieve the highest system EE compared with the other two benchmark schemes. Furthermore, owing to the flexibility of having both offloading and local computing in the proposed scheme and in the "Eve fully decode" scheme, the system can decide not to offload if the eavesdropper has a better channel on the offloading link while it can decide to offload if the link is secure enough. Therefore, these two schemes have a higher EE performance than the "Full offloading" scheme, which has to offload even when the links are insecure. The system EE stabilizes for all the three schemes after 200 time slots.

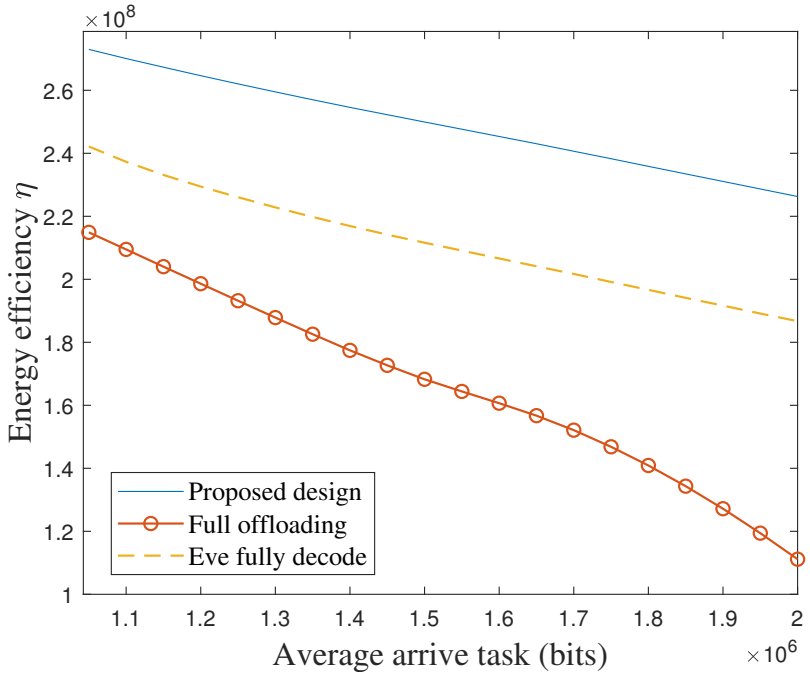

Fig. 4.3: System energy efficiency v.s. Average arrival task length.

The system EE versus the average arrival task length is presented in Fig. 4.3. The proposed method achieves the highest EE. For all the three schemes, EE decrease with the increase of the arrival task length because a higher workload forces the system to increase the computing rate to maintain the low queue level. This in turn decreases the system EE. Furthermore, we notice that the performance gap between the "Full offloading" scheme and

other two schemes goes up with the increase of the task length. This demonstrates that local computing is more energy efficient and secure for processing the computation tasks when the task size goes up.

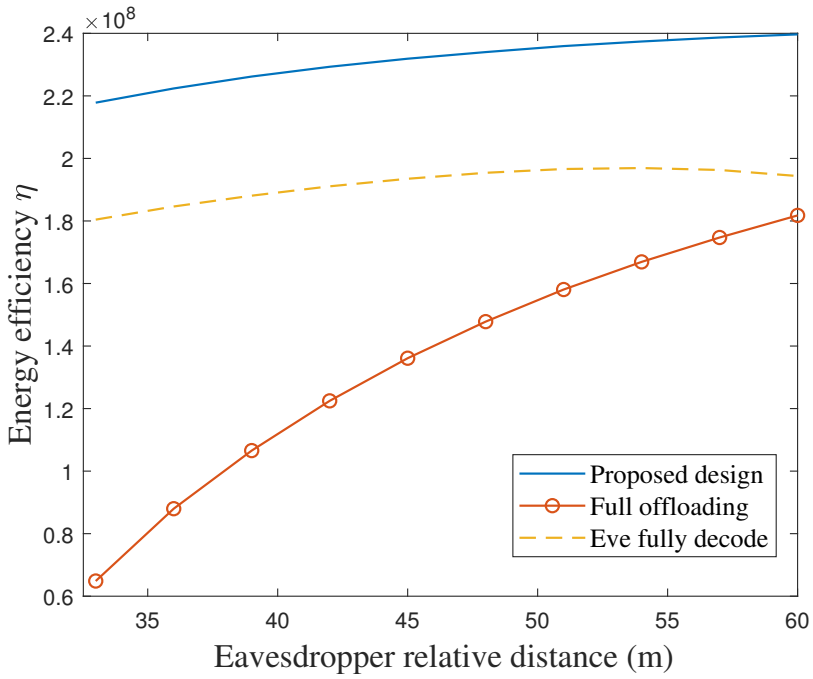

Fig. 4.4: System energy efficiency v.s. eavesdropper relative distance.

Fig. 4.4 shows the system EE versus the eavesdropper location. Here the eavesdropper relative distance is defined as the distance between the eavesdropper and the UE. The proposed design achieves the best performance among all the schemes. The system EE of all the schemes goes up as the eavesdropper relative distance increases since a larger distance leads to a worse intercepting channel at the eavesdropper. Furthermore, the performance gap between the "Full offloading" scheme and the other two schemes decreases quickly with the increase of the relative distance. This is because the secure offloading rate increases quickly when the eavesdropper moves away.

The relationship between EE and the maximum available power is illustrated in Fig. 4.5. It is observed that EE increases with available power and gradually converges to a constant value. This is because that when the available power is limited, the higher

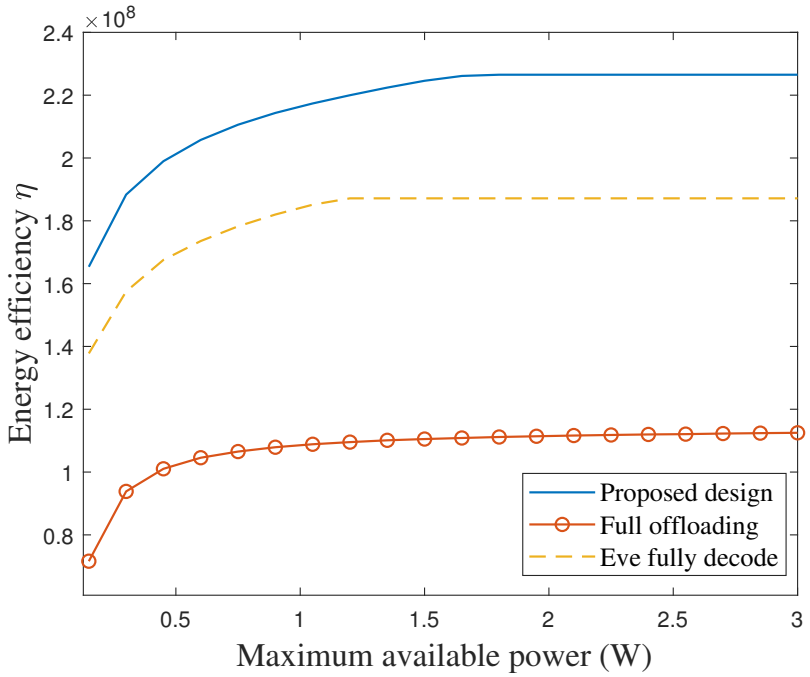

Fig. 4.5: System energy efficiency v.s. maximum available power $P_{\max}$.

computing rate and corresponding optimal EE cannot be achieved. With the power increase, EE of all the schemes keeps increasing and only stops when it achieves the highest level. After the optimal tradeoff has been reached, even there is more power available in the system, all the schemes maintain at the highest level without consuming any more power.

4.5 Chapter Conclusion

This chapter aims to design a secure and energy efficient computation offloading scheme in a NOMA enabled MEC network with the presence of a malicious eavesdropper. In order to achieve a long term performance gain by considering dynamic task arrivals and fading channels, we proposed a secure task offloading and computation resource allocation scheme that aims to maximize the long-term average EE and used Lyapunov optimization framework to solve the problem. Numerical results validated the advantages of the proposed design via comparisons with two other benchmark schemes.

CHAPTER 5

Energy Efficient Robust Beamforming and Cooperative Jamming Design for IRS-Assisted MISO Networks

5.1 Introduction

In previous chapters, the NOMA-assisted MEC networks have been thoroughly investigated with respect to EE and PLS. However, new spectrum bands, such as mmWave communications in 5G networks, bring many potential benefits to IoT networks. For example, the multiple-input multiple-output (MIMO) techniques enable the transceivers to explore the space diversity and seek a higher throughput and EE. However, severe pathloss and the tendency of LoS propagation links to block impair the higher frequency channels. To exploit the spatial diversity in a controllable way and protect against path fading and block, many researchers are looking to IRS to potentially improve EE and achieve secure communications at one blow. It has also been shown that secure performance can be improved significantly by cooperative jamming [71]. Thus, it is envisioned that IRS-assisted cooperative jamming is promising to further improve the secrecy rate of the legitimate users.

Most of the existing works that focus on IRS-assisted secure communication networks assumed that the channel state information of the link from the IRS to the eavesdropper can be perfectly obtained [72], [73]. However, in practice, it is extremely difficult to obtain perfect CSI of the link from the IRS to eavesdropper. The reasons are as follows. On one hand, the existence of channel estimation errors and quantization errors can result in imperfect CSI estimation [50]. On the other hand, since the locations of the eavesdroppers are unknown and there is no cooperation between the legitimate user and the eavesdroppers, perfect CSI is almost impossible to obtain [74]. Imperfect CSI can significantly deteriorate the beamforming and IRS performance. Thus, it is of crucial importance to design robust secure beamforming and phase shift matrix for IRS assisted cooperative jamming (CJ)

communication networks.

Motivated by the above-mentioned facts, in this paper, robust secure beamforming and phase shift matrix are designed for an IRS assisted MISO network with an independent cooperative jamming user. The EE maximization framework is formulated. To the author's best knowledge, this is the first work that considers robust beamforming and cooperative jamming design in IRS-assisted MISO networks with CJ and that studies the EE maximization problems in this type of network.

5.1.1 Related Work and Motivation

Optimal beamforming design plays an important role in the improvement of secure performance in wireless communication networks by using physical layer security. The related works can be classified into three categories, namely, secure beamforming design in conventional MISO networks with CJ under perfect CSI [75]- [76], robust secure beamforming design in conventional MISO networks with CJ under imperfect CSI [77]- [78], and secure beamforming design in IRS-assisted secure wireless networks [22]- [28].

For the conventional MISO network secure communication with perfect CSI, the beamforming and jamming design were jointly optimized to achieve different objectives, e.g., the secrecy rate maximization of users [75]- [79], the minimization of energy consumption [80], and the system efficiency maximization [81], [76]. Specifically, in [75], the authors exploited the CJ for multiple users via broadcast channels to enhance the secure performance with the help of a friendly jammer. The optimal CJ was designed to keep the achieved SINR at the eavesdroppers below the threshold to guarantee that the transmission from the base station to the legitimate users is confidential. To achieve a higher secrecy rate performance, in [82], Park *et al.* investigated a single relay assisted secure communication network. By using CJ to prevent the eavesdropper from intercepting the source message, they proposed three jamming power allocation strategies to minimize the outage probability of the secrecy rate. Different from the single relay system, a wireless network with multiple relays was considered in [83]. A two-slot cooperative relaying scheme was proposed to maximize the secrecy rate. The access method is another key element for increasing the system secrecy

rate. The authors in [79] studied the secrecy rate maximization problem in an orthogonal frequency division multiple (OFDM) system with a potential eavesdropper. With the assistance of a cooperative jammer, the approaches they proposed can significantly improve the secrecy rate by jointly optimizing the transmit power and time allocation. While the works in [75]- [79] aim to achieve a higher secrecy rate, they only consider one performance metric therefore may not be able to achieve a good tradeoff between conflicting performance goals such as high rate and low energy consumption. Recently, the authors in [80] considered secure resource allocations for OFDM networks under scenarios with and without CJ. The joint optimization problem of subcarrier assignments and power allocations subject to a limited power budget at the relay was solved to maximize the secrecy sum-rate and save energy. Different from the works in [75]- [80], energy-efficient secure communication was considered in [81]. By using two jamming strategies, namely, beamforming and cooperative diversity, they demonstrated that a cooperative diversity strategy is desirable. Significant EE can be achieved by selectively switching between the two strategies. Besides the strategy selection, the mode switch can also improve the EE. In [76], the authors proposed an intermittent jamming strategy where a jammer alternates between jamming and non-jamming modes during the legitimate transmission. By jointly measuring security requirements and energy costs, they formulated and solved an optimization problem with respect to the jamming duration proportion and jamming power.

In practice, the perfect CSI is not always available at the transmitter. The secure network designs presented above are not suitable for imperfect CSI cases. Thus, to achieve robust design of the secure communication network, the beamforming design problems with channel estimation error have been considered [77]- [78]. The authors in [77] studied robust transmission schemes with a single eavesdropper for MISO networks. Both the cases of direct transmission and CJ were investigated with imperfect CSI for the eavesdropper links. Robust transmission covariance matrices were obtained by solving the worst-case secrecy rate maximization. For the MISO system with multiple eavesdroppers, Ma *et al.* in [84] investigated a robust quality-of-service (QoS)-based and secrecy rate-based secure

transmission design. By jointly optimizing the transmit beamforming vector and the covariance matrix of jamming signals under individual power constraints, they proposed an algorithm for each problem through semidefinite relaxation (SDR). In [85], the authors aimed to minimize the total transmit power by jointly designing the beamforming vector at the transmitter and AN covariance at jammer under the reliability and secrecy constraints for all the possible distributions of CSI errors. Su *et al.* in [86] proposed a novel robust beamforming strategy for the direct transmission NOMA and cooperative jamming NOMA to minimize the worst-case sum power subject to secrecy rate constraint. In [87], Feng *et al.* investigated cooperative secure beamforming for simultaneous wireless information and power transfer (SWIPT) in AF relay networks with imperfect CSI. They proposed a joint cooperative beamforming (CB) and energy signal (ES) scheme to maximize the secrecy rate under both the power constraints and the wireless power transfer constraint. In [88], Chu *et al.* studied a MISO secrecy network with CJ and SWIPT to maximize the minimum harvested energy subject to the total power constraints while guaranteeing the minimum secrecy rate. By incorporating the norm-bound channel uncertainty model, they proposed a joint design of the robust secure transmission scheme which outperforms the separate AN-aided or CJ-aided schemes. By considering the secrecy rate and consumed energy of the robust secure communication network simultaneously, the tradeoff between them can be investigated to achieve the maximum EE. In [74], a MISO cognitive radio downlink network with SWIPT was studied. The tradeoff was elucidated between the secrecy rate and the harvesting energy under the max-min fairness criterion. The joint design of the beamforming vector and the artificial noise covariance matrix were investigated in [78] for the MISO multiple-eavesdropper SWIPT systems. The secrecy EE maximization problem was formulated and two suboptimal solutions were proposed based on the heuristic beamforming techniques.

Recently, the IRS-assisted MISO secure network has attracted increasingly elevated attention. The beamforming and phase shift matrix design schemes for different objectives were proposed in [89]- [28]. For the multi-user network, in [89], the authors investigated the

symbol-level precoding in IRS-assisted multiuser MISO systems to minimize the transmit power while guaranteeing the information transmissions. In [90], the authors considered the downlink multigroup multicast communication systems assisted by an IRS. By optimizing the precoding matrix and the reflection coefficients, the sum rate of all the multicasting groups was maximized. For the multi-IRS deployment problems, in [91], the deploying strategies for IRS were investigated for a single-cell multiuser system aided by multiple IRSs. It was shown that the IRS-aided system outperforms the full-duplex relay-aided counterpart system and that the deployment strategies and the elements of IRS have significant influence on the achievable spatial throughput. In [92], the authors analyzed the impact of the deployment of IRS on the downlink throughput and showed that IRS density can significantly enhance the signal power at the expense of only a marginally increasing interference.

Although beamforming design problems in CJ assisted secure MISO networks [75]- [76] and robust beamforming design problems in CJ assisted secure MISO networks under the imperfect CSI [77]- [78] have been investigated, few studies have been conducted for beamforming, friendly jamming and phase shift matrix design in IRS assisted wireless MISO networks. Moreover, EE optimization based on perfect CSI proposed in the above-mentioned works are not applicable to the imperfect CSI since the channel estimated errors can have a big impact on the performance of both base station and friendly jammer. Furthermore, with imperfect CSI, the application of IRS into the MISO network with friendly jamming can face more challenges that have not been considered in the works mentioned above. Thus, in order to improve EE performance and achieve robustness against the uncertainty introduced by the imperfect CSI, it is of crucial importance to study robust beamforming, friendly jamming, and phase shift matrix design problems in IRS-aided MISO networks. These problems are normally challenging to tackle due to two reasons. There exists dependence among different variables that makes the problems non-convex. The imperfect CSI model further increases the complexity of the problems by introducing the uncertainty constraints to the optimization problems.

5.1.2 Contribution and Organization

Motivated by the above-mentioned observations, in this chapter, the EE maximization problems are studied in an IRS-assisted MISO network with cooperative jamming under both perfect and imperfect CSI models. The corresponding robust design to address channel uncertainty is also provided. The major contributions of this paper are summarized as follows.

1) We investigate the joint design of information beamforming, cooperative jamming, and phase shift matrix to maximize the EE in an IRS-assisted secure network with eavesdroppers under the perfect CSI model. The problem is challenging to solve due to its non-convexity and coupling of the beamforming vector with the IRS phase shift matrix. An alternating optimization algorithm is proposed to solve the non-convex fractional problem by using SDR.

2) For the IRS aided MISO network under imperfect CSI model, the estimated channel error results in the uncertainty to the system and brings more difficulties for beamforming and phase shift matrix design compared with the perfect CSI case. To deal with this uncertainty, the bounded channel error model is considered and the $\mathcal{S}$-procedure method is applied for optimizing the robust beamforming and IRS phase shift matrix to maximize the EE.

3) The simulation results show that the proposed method with the perfect CSI can achieve the highest EE among all the benchmark methods. Moreover, it is found that there is a tradeoff between secrecy rate and the consumed energy. Furthermore, it is shown that the exploitation of IRS is beneficial for improving EE even under the imperfect CSI case.

Notation: $\mathbb{C}^{M\times N}$ denotes the $M\times N$ complex-valued matrices space. $\mathcal{CN}(\mu,\sigma^2)$ denotes the distribution of complex Gaussian random variable with mean μ and variance σ^2. For a square matrix $\mathbf{X}$, the trace of $\mathbf{X}$ is denoted as $\mathrm{Tr}(\mathbf{X})$ and $\mathrm{rank}(\mathbf{X})$ denotes the rank of matrix $\mathbf{X}$. $\angle(x)$ denotes the phase of complex number x. Matrices and vectors are denoted by boldface capital letters and boldface lower case letters. $[x]^+$ denotes the maximum

between 0 and x.

5.2 System Model

As shown in Fig. 5.1, an IRS assisted wireless communication system is considered. A multi-antenna base station transmits the confidential information to a single-antenna legitimate user. At the same time, K eavesdroppers (Eves) are trying to intercept the information from the base station. In order to improve the security, a friendly jammer intentionally issues the jamming signals. It is assumed that both the base station and the jammer are equipped with N antennas, and the IRS has M reflecting elements. Each Eve is equipped with a single antenna.

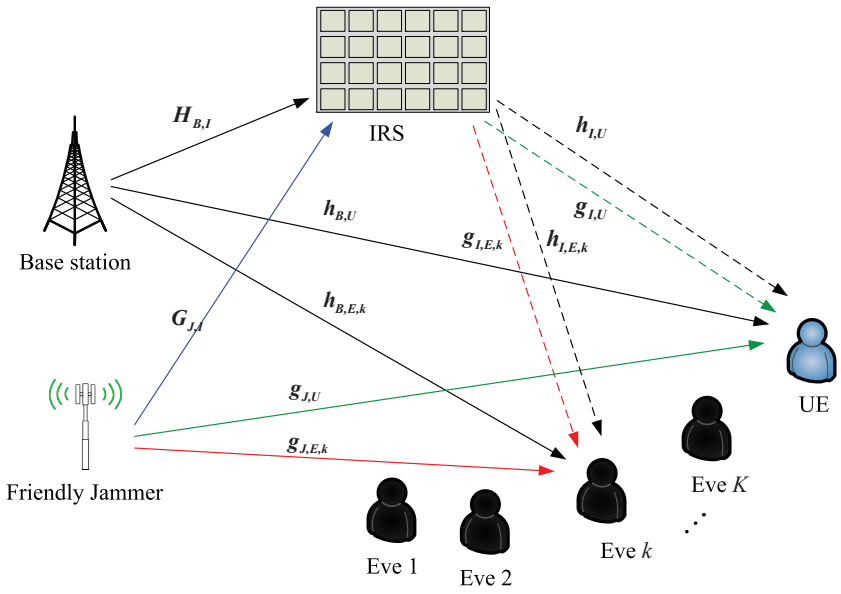

Fig. 5.1: An IRS-aided MISO wireless network with a friendly jammer.

The base band equivalent channel from the base station to the IRS, base station to the user, and base station to the kth Eve are denoted as $\mathbf{H}_{B,I} \in \mathbb{C}^{M\times N}$, $\mathbf{h}_{B,U} \in \mathbb{C}^{1\times N}$, and $\mathbf{h}_{B,E,k} \in \mathbb{C}^{1\times N}$, respectively. The baseband equivalent channel from the Jammer to the IRS, Jammer to user, and Jammer to Eve k are denoted as $\mathbf{G}_{J,I} \in \mathbb{C}^{M\times N}$, $\mathbf{g}_{J,U} \in \mathbb{C}^{1\times N}$, and $\mathbf{g}_{J,E,k} \in \mathbb{C}^{1\times N}$, respectively. The channel from the IRS to the user and Eve k are

denoted as $\mathbf{h}_{I,U}$, $\mathbf{h}_{I,E,k}$, $\mathbf{g}_{I,U}$, and $\mathbf{g}_{I,E,k}$, respectively, each of which is a $1 \times M$ complex vector. The performance achieved under the perfect CSI case can serve as an upper bound for the proposed secure communication design. According to the works in [93], [94], the Eves may be legitimate users in the past but cannot access the confidential information in the current communication process or the base station does not want to send confidential information to those users. To guarantee the communication security, the system has to treat those receivers as potential eavesdroppers. Therefore, the perfect CSI of those Eves can be acquired.

It is assumed that the channel information between the IRS and the user is available at both the base station and the jammer. However, since eavesdroppers normally try to hide their existence from the base station, it is difficult to obtain the perfect CSI between Eves and base station. In practice, the CSI knowledge of the links from the IRS to Eves is not accurate. This can also be caused by channel estimation and quantization errors. In order to develop a robust scheme under the imperfect CSI case, the worst case channel uncertainty model is considered. The bounded CSI error models for the channel vector $\mathbf{h}_{I,E,k}$, and $\mathbf{g}_{I,E,k}$ are given as

$$\begin{aligned}\mathbf{h}_{I,E,k} =& \overline{\mathbf{h}}_{I,E,k} + \Delta\mathbf{h}_{I,E,k},\ \mathcal{H}_{I,E,k} \qquad (5.1a)\\ =& \{\Delta\mathbf{h}_{I,E,k} \in \mathbb{C}^{M\times 1} : \Delta\mathbf{h}_{I,E,k}^{H}\Delta\mathbf{h}_{I,E,k} \leq \xi_{I,E,k}^{2}\},\\ \mathbf{g}_{I,E,k} =& \overline{\mathbf{g}}_{I,E,k} + \Delta\mathbf{g}_{I,E,k},\ \mathcal{G}_{I,E,k} \qquad (5.1b)\\ =& \{\Delta\mathbf{g}_{I,E,k} \in \mathbb{C}^{M\times 1} : \Delta\mathbf{g}_{I,E,k}^{H}\Delta\mathbf{g}_{I,E,k} \leq \xi_{J,E,k}^{2}\},\end{aligned}$$

where $\overline{\mathbf{h}}_{I,E,k}$, and $\overline{\mathbf{g}}_{I,E,k}$ are the estimated values of the channel vectors $\mathbf{h}_{I,E,k}$, and $\mathbf{g}_{I,E,k}$, respectively. $\mathcal{H}_{I,E,k}$, and $\mathcal{G}_{I,E,k}$ denote the uncertainty regions of $\mathbf{h}_{I,E,k}$, and $\mathbf{g}_{I,E,k}$, respectively. $\Delta\mathbf{h}_{I,E,k}$, and $\Delta\mathbf{g}_{I,E,k}$ represent the channel estimation errors. $\xi_{I,E,k}$, and $\xi_{J,E,k}$ are the radius of the uncertainty region $\mathcal{H}_{I,E,k}$, and $\mathcal{G}_{I,E,k}$, respectively [74].

In this paper, the IRS adjusts its elements to maximize the combined incident signal for the legitimate user. The diagonal phase-shift matrix can be denoted as $\mathbf{\Theta} =$

$\text{diag}(\exp(j\theta_1), \exp(j\theta_2), \cdots, \exp(j\theta_M))$, wherein its main diagonal, $\theta_m \in [0, 2\pi)$, denotes the phase shift on the combined incident signal by its mth element, $m = 1, 2, ..., M$ [95].

The transmitted signal from the base station to the user is given as $\mathbf{x}_B = \mathbf{f}_1 s_1$ and the jamming signal from the jammer is given as $\mathbf{x}_J = \mathbf{f}_2 s_2$, where $s_1 \sim \mathcal{CN}(0, 1)$ and $s_2 \sim \mathcal{CN}(0, 1)$ denote the independent information and jamming signal, respectively. $\mathbf{f}_1 \in \mathbb{C}^{N\times 1}$ and $\mathbf{f}_2 \in \mathbb{C}^{N\times 1}$ denote the beamforming and jamming precode vectors, respectively. Let $P_{1,max}$ and $P_{2,max}$ denote the maximum transmit power available at base station and jammer. We have $(\mathbf{f}_1^H \mathbf{f}_1) \leq P_{1,max}$ and $(\mathbf{f}_2^H \mathbf{f}_2) \leq P_{2,max}$. The signal received at legitimate user and Eve k can be respectively given as

$$\mathbf{y}_U = (\mathbf{h}_{B,U}^H + \mathbf{h}_{I,U}^H \boldsymbol{\Theta} \mathbf{H}_{B,I}) \mathbf{f}_1 s_1 + (\mathbf{g}_{J,U}^H + \mathbf{g}_{I,U}^H \boldsymbol{\Theta} \mathbf{G}_{J,I}) \mathbf{f}_2 s_2 + n_U, \tag{5.2}$$

and

$$\mathbf{y}_{E,k} = (\mathbf{h}_{B,E,k}^H + \mathbf{h}_{I,E,k}^H \boldsymbol{\Theta} \mathbf{H}_{B,I}) \mathbf{f}_1 s_1 + (\mathbf{g}_{J,E,k}^H + \mathbf{g}_{I,E,k}^H \boldsymbol{\Theta} \mathbf{G}_{J,I}) \mathbf{f}_2 s_2 + n_{E,k}, \tag{5.3}$$

where n_U and $n_{E,k} \sim \mathcal{CN}(0, \sigma^2)$ are the complex additive white Gaussian noise (AWGN). Thus, the signal of interference plus noise ratio (SINR) of the legitimate user and Eve k can be given as

$$\gamma_U = \frac{|(\mathbf{h}_{B,U}^H + \mathbf{h}_{I,U}^H \boldsymbol{\Theta} \mathbf{H}_{B,I}) \mathbf{f}_1|^2}{|(\mathbf{g}_{J,U}^H + \mathbf{g}_{I,U}^H \boldsymbol{\Theta} \mathbf{G}_{J,I}) \mathbf{f}_2|^2 + \sigma^2}, \tag{5.4}$$

and

$$\gamma_{E,k} = \frac{|(\mathbf{h}_{B,E,k}^H + \mathbf{h}_{I,E,k}^H \boldsymbol{\Theta} \mathbf{H}_{B,I}) \mathbf{f}_1|^2}{|(\mathbf{g}_{J,E,k}^H + \mathbf{g}_{I,E,k}^H \boldsymbol{\Theta} \mathbf{G}_{J,I}) \mathbf{f}_2|^2 + \sigma^2}. \tag{5.5}$$

The achievable secrecy rate is defined as

$$R_S = [R_U - R_E]^+ = [B \log_2(1 + \gamma_U) - \max_{k\in K} B \log_2(1 + \gamma_{E,k})]^+. \tag{5.6}$$

The energy consumed by the base station and the jammer consists of the transmit power and the circuit power consumption P_{BS} and P_G. The power consumed by the IRS is

denoted as P_{IRS}. Thus, the total power consumed in the system can be given as

$$P_{tot} = \zeta(\mathbf{f}_1^H \mathbf{f}_1 + \mathbf{f}_2^H \mathbf{f}_2) + P_{BS} + P_G + P_{IRS}, \tag{5.7}$$

where ζ is the amplifier coefficient.

According to [95], the energy efficiency (bps/ Watt or bits/Joule) is defined as

$$\eta = \frac{[B \log_2(1+\gamma_U) - \max_{k \in K} B \log_2(1+\gamma_{E,k})]^+}{\zeta(\mathbf{f}_1^H \mathbf{f}_1 + \mathbf{f}_2^H \mathbf{f}_2) + P_{BS} + P_G + P_{IRS}}. \tag{5.8}$$

In order to maximize the energy efficiency, the beamforming and jamming vectors and the phase shift matrix are jointly optimized. Since the energy efficiency maximization problem is extremely challenging under the imperfect CSI case, the problem is firstly studied under the perfect CSI case in order to provide some meaningful insights in Section 5.3. Based on the results obtained in Section 5.3, the energy efficiency maximization problem is further studied under the imperfect CSI in Section 5.4.

5.3 System Design With Perfect CSI

In this section, the energy efficiency maximization problem with perfect CSI is studied by jointly optimizing the beamforming vector, jamming vector, and phase shift matrix. An alternating algorithm is proposed to tackle the challenging non-convex problem.

5.3.1 Problem Formulation

Under the perfect CSI model, the energy efficiency maximization problem is formulated as

$$\begin{aligned}
\mathbf{P}_5: \ & \max_{\mathbf{f}_1,\mathbf{f}_2,\boldsymbol{\Theta}} \ \eta \\
s.t. \ & \mathbf{f}_1^H\mathbf{f}_1 \le P_{1,\max}, && (5.9a)\\
& \mathbf{f}_2^H\mathbf{f}_2 \le P_{2,\max}, && (5.9b)\\
& R_s \ge R_{th}, && (5.9c)\\
& |\exp(j\theta_m)| = 1, && (5.9d)
\end{aligned}$$

where R_{th} is the minimum required secure rate threshold. It is evident that problem $\mathbf{P}_5$ is non-convex due to the fractional structure of the objective function and the non-convex constraints. In order to tackle it, an alternating algorithm is proposed to solve this problem.

The problem $\mathbf{P}_5$ is non-convex due to the coupling of the beamforming vector, jamming vector and IRS phase shift matrix. By introducing $\mathbf{w}^H = [w_1, w_2, \cdots, w_M]$, one has $\mathbf{h}_{I,j}^H\boldsymbol{\Theta}\mathbf{H}_{B,I} = \mathbf{w}^H\mathbf{H}_{I,j}$, where $w_m = \exp(j\theta_m)$, $\mathbf{H}_{I,j} = \text{diag}(\mathbf{h}_{I,j}^H)\mathbf{H}_{B,I},\ j \in \{U,(E,k)\}$. The interference from the jammer can be denoted as $\mathbf{g}_{I,j}^H\boldsymbol{\Theta}\mathbf{G}_{J,I} = \mathbf{w}^H\mathbf{G}_{I,j}$, where $\mathbf{G}_{I,j} = \text{diag}(\mathbf{g}_{I,j}^H)\mathbf{G}_{J,I},\ j \in \{U,(E,k)\}$. Thus, the SINRs of user and Eve k are given as $\gamma_j = \frac{a_0|\overline{\mathbf{w}}^H\mathbf{H}_j\mathbf{f}_1|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{G}_j\mathbf{f}_2|^2+1},\ \ j \in \{U,(E,k)\}$, where $a_0 = 1/\sigma^2$, $\mathbf{H}_j = \begin{bmatrix} \mathbf{H}_{I,j} \\ \mathbf{h}_{B,j} \end{bmatrix}$, $\mathbf{G}_j = \begin{bmatrix} \mathbf{G}_{I,j} \\ \mathbf{g}_{J,j} \end{bmatrix}$, $\overline{\mathbf{w}}^H = \exp(j\overline{w})[\mathbf{w}^H, 1]$ and $\overline{w}$ is an arbitrary phase rotation. The problem can be transformed into

$$\begin{aligned}
\mathbf{P}_{5.1}: \ & \max_{\mathbf{f}_1,\mathbf{f}_2,\overline{\mathbf{w}}} \ \frac{1}{P_{tot}}\{\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{w}}^H\mathbf{H}_U\mathbf{f}_1|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{G}_U\mathbf{f}_2|^2+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{w}}^H\mathbf{H}_{E,k}\mathbf{f}_1|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{G}_{E,k}\mathbf{f}_2|^2+1})\} \\
s.t. \ & (5.9a), (5.9c), \\
& \frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{w}}^H\mathbf{H}_U\mathbf{f}_1|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{G}_U\mathbf{f}_2|^2+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{w}}^H\mathbf{H}_{E,k}\mathbf{f}_1|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{G}_{E,k}\mathbf{f}_2|^2+1}) \ge R_{th}. \qquad (5.10a)
\end{aligned}$$

The problem $\mathbf{P}_{5.1}$ is yet still non-convex. In order to tackle it, the beamforming and jamming vectors can be optimized for a given $\overline{\mathbf{w}}$, and then $\overline{\mathbf{w}}$ can be optimized for the obtained optimal $\mathbf{f}_1$ and $\mathbf{f}_2$. This process iteratively continues till convergence.

5.3.2 Optimizing the Beamforming for a Given $\overline{\mathbf{w}}$

In this section, we solve the problem $\mathbf{P}_{5.1}$ to achieve the optimal secure transmit beamformer $\mathbf{f}_1$ and jammer vector $\mathbf{f}_2$ for a given $\overline{\mathbf{w}}$. Let $\overline{\mathbf{h}}_U^H = \overline{\mathbf{w}}^H\mathbf{H}_U$, $\overline{\mathbf{g}}_U^H = \overline{\mathbf{w}}^H\mathbf{G}_U$, $\overline{\mathbf{h}}_{E,k}^H = \overline{\mathbf{w}}^H\mathbf{H}_{E,k}$, and $\overline{\mathbf{g}}_{E,k}^H = \overline{\mathbf{w}}^H\mathbf{G}_{E,k}$. The problem $\mathbf{P}_{5.1}$ can be transformed into

$$\mathbf{P}_{5.2}: \max_{\mathbf{f}_1,\mathbf{f}_2} \frac{1}{P_{tot}}\{\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{h}}_U^H\mathbf{f}_1|^2}{a_0|\overline{\mathbf{g}}_U^H\mathbf{f}_2|^2+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{h}}_{E,k}^H\mathbf{f}_1|^2}{a_0|\overline{\mathbf{g}}_{E,k}^H\mathbf{f}_2|^2+1})\}$$

$$s.t. \quad \frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{h}}_U^H\mathbf{f}_1|^2}{a_0|\overline{\mathbf{g}}_U^H\mathbf{f}_2|^2+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0|\overline{\mathbf{h}}_{E,k}^H\mathbf{f}_1|^2}{a_0|\overline{\mathbf{g}}_{E,k}^H\mathbf{f}_2|^2+1}) \geq R_{th}, \tag{5.11a}$$

$$\mathbf{f}_1^H\mathbf{f}_1 \leq P_{1,max}, \mathbf{f}_2^H\mathbf{f}_2 \leq P_{2,max}. \tag{5.11b}$$

Let $|\overline{\mathbf{h}}_j^H\mathbf{f}_1|^2 = \mathrm{Tr}(\overline{\mathbf{H}}_\mathrm{j}\mathbf{f}_1\mathbf{f}_1^\mathrm{H})$ and $|\overline{\mathbf{g}}_j^H\mathbf{f}_2|^2 = \mathrm{Tr}(\overline{\mathbf{G}}_\mathrm{j}\mathbf{f}_2\mathbf{f}_2^\mathrm{H})$. By defining $\overline{\mathbf{H}}_j = \overline{\mathbf{h}}_j\overline{\mathbf{h}}_j^H$, $\overline{\mathbf{G}}_j = \overline{\mathbf{g}}_j\overline{\mathbf{g}}_j^H$, $j \in \{U,(E,k)\}$, $\mathbf{F}_1 = \mathbf{f}_1\mathbf{f}_1^H$ and $\mathbf{F}_2 = \mathbf{f}_2\mathbf{f}_2^H$, one has $\mathbf{F}_1 \succeq 0$, $\mathbf{F}_2 \succeq 0$ and $\mathrm{rank}(\mathbf{F}_1) = \mathrm{rank}(\mathbf{F}_2) = 1$. The rank-1 constraint makes the problem difficult to solve. Thus, we apply the SDR method to relax the constraints. The problem $\mathbf{P}_{5.2}$ is thus expressed as

$$\mathbf{P}_{5.3}: \max_{\mathbf{F}_1,\mathbf{F}_2} \frac{1}{P_{tot}}(\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\overline{\mathbf{H}}_\mathrm{U}\mathbf{F}_1)}{a_0\mathrm{Tr}(\overline{\mathbf{G}}_\mathrm{U}\mathbf{F}_2)+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\overline{\mathbf{H}}_\mathrm{E,k}\mathbf{F}_1)}{a_0\mathrm{Tr}(\overline{\mathbf{G}}_\mathrm{E,k}\mathbf{F}_2)+1}))$$

$$s.t. \quad (\mathbf{F}_1,\mathbf{F_2}) \in \mathcal{F}, \tag{5.12a}$$

$$\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\overline{\mathbf{H}}_\mathrm{U}\mathbf{F}_1)}{a_0\mathrm{Tr}(\overline{\mathbf{G}}_\mathrm{U}\mathbf{F}_2)+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\overline{\mathbf{H}}_\mathrm{E,k}\mathbf{F}_1)}{a_0\mathrm{Tr}(\overline{\mathbf{G}}_\mathrm{E,k}\mathbf{F}_2)+1}) \geq R_{th}, \tag{5.12b}$$

where $\mathcal{F} = \{(\mathbf{F}_1,\mathbf{F}_2)|\mathrm{Tr}(\mathbf{F}_1) \leq P_{1,\mathrm{max}},\ \mathrm{Tr}(\mathbf{F}_2) \leq P_{2,\mathrm{max}},\ \mathbf{F}_1 \succeq 0,\ \mathbf{F}_2 \succeq 0\}$. However, the problem $\mathbf{P}_{5.3}$ is still a non-convex problem due to the objective function and the non-convex second constraint with respect to $\mathbf{F}_1$ and $\mathbf{F}_2$. To solve this, the following lemma is applied [68].

Lemma 5.1: By introducing the function $\phi(t) = -tx + \ln t + 1$ for any $x > 0$, one has

$$-\ln x = \max_{t>0} \phi(t). \tag{5.13}$$

The optimal solution can be achieved at $t = 1/x$. The upper bound can be given by using Lemma 1 as $\phi(t)$. By setting $x = a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + 1$, and $t = t_U$, one has

$$R_U\frac{\ln 2}{B} = [\ln(a_0\mathrm{Tr}(\overline{\mathbf{H}}_U\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + 1) - \ln(a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + 1)] = \max_{t_U>0} \phi_u(\mathbf{F}_1, \mathbf{F}_2, t_U), \tag{5.14}$$

where $\phi_U(\mathbf{F}_1, \mathbf{F}_2, t_U) = \ln(a_0\mathrm{Tr}(\overline{\mathbf{H}}_U\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + 1) - t_U(a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2)+1)+\ln t_U+1$.

In the same way, let $x = a_0\mathrm{Tr}(\overline{\mathbf{H}}_{E,k}\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1$ and $t = t_{E,k}$, one has

$$\begin{aligned} R_{E,k}\frac{\ln 2}{B} &= [\ln(a_0\mathrm{Tr}(\overline{\mathbf{H}}_{E,k}\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1) - \ln(a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1)] \\ &= \min_{t_{E,k}>0} \phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2, t_{E,k}), \end{aligned} \tag{5.15}$$

where $\phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2, t_{E,k}) = t_{E,k}(a_0\mathrm{Tr}(\overline{\mathbf{H}}_{E,k}\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1) - \ln(a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1) - \ln t_{E,k} - 1$. By using Sion's minimax theorem [96], the problem given by eq. (5.12) can be transformed into

$$\begin{aligned} \mathbf{P}_{5.4} \quad & \max_{\mathbf{F}_1, \mathbf{F}_2, t_U, t_{E,k}} \frac{\phi_U(\mathbf{F}_1, \mathbf{F}_2, t_U) - \max_k \phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2 t_{E,k})}{\frac{\ln 2}{B}(\mathrm{Tr}(\mathbf{F}_1 + \mathbf{F}_2) + P_{BS} + P_G + P_{IRS})} \\ & \text{s.t.} \quad (\mathbf{F}_1, \mathbf{F}_2) \in \mathcal{F}, && \text{(5.16a)} \\ & \quad \phi_U(\mathbf{F}_1, \mathbf{F}_2, t_U) - \max_k \phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2, t_{E,k}) \geq R_{th}\frac{\ln 2}{B}, && \text{(5.16b)} \\ & \quad t_U, t_{E,k} \geq 0. && \text{(5.16c)} \end{aligned}$$

According to Lemma 1, the optimal values of t_U and $t_{E,k}$ can be achieved when $t_U^* = (a_0\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + 1)^{-1}$ and $t_{E,k}^* = (a_0\mathrm{Tr}(\overline{\mathbf{H}}_{E,k}\mathbf{F}_1) + a_0\mathrm{Tr}(\overline{\mathbf{G}}_{E,k}\mathbf{F}_2) + 1)^{-1}$. Here, a slack variable $l \geq \max_{k\in K} \phi_{E,k}$ is introduced. Thus, the optimization problem $\mathbf{P}_{5.4}$ for $\mathbf{F}_1$ and

$\mathbf{F}_2$ based on t_U^* and $t_{E,k}^*$ can be given as

$$\mathbf{P}_{5.5} \quad \max_{\mathbf{F}_1,\mathbf{F}_2} \frac{\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U^*) - l}{\frac{\ln 2}{B}(\zeta \mathrm{Tr}(\mathbf{F}_1+\mathbf{F}_2) + P_{BS} + P_G + P_{IRS})}$$

$$\text{s.t.} \quad (\mathbf{F}_1,\mathbf{F}_2) \in \mathcal{F}, \tag{5.17a}$$

$$\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U^*) - l \geq R_{th}\frac{\ln 2}{B}, \tag{5.17b}$$

$$\phi_{E,k}(\mathbf{F}_1,\mathbf{F}_2,t_{E,k}^*) \leq l. \tag{5.17c}$$

The objective function of $\mathbf{P}_{5.5}$ is now a concave function over a convex function, and the constraints are all convex, since $\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U^*)$ is concave and $\phi_{E,k}(\mathbf{F}_1,\mathbf{F}_2,t_{E,k}^*)$ is convex. It is a single ratio maximization problem and can be solved with the Dinkelbach's method [97] [98]. Using the Dinkelbach's method [99], $\mathbf{P}_{5.5}$ can be solved by iteratively solving the following problem, given as

$$\mathbf{P}_{5.6} \quad \max_{\mathbf{F}_1,\mathbf{F}_2} \phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U^*) - l - \frac{\ln 2}{B}\eta_1^*(\zeta \mathrm{Tr}(\mathbf{F}_1+\mathbf{F}_2) + P_{BS} + P_G + P_{IRS})$$

$$\text{s.t.} \quad (5.17\text{a}), (5.17\text{b}), (5.17\text{c}),$$

where η_1^* is a non-negative parameter. $\mathbf{P}_{5.6}$ is convex and can be solved by using a standard convex optimization tool [93].

After the $\mathbf{F_1}$ and $\mathbf{F_2}$ are obtained, if $\mathrm{rank}(\mathbf{F_1}) = \mathrm{rank}(\mathbf{F_2}) = 1$, $\mathbf{f_1}$ and $\mathbf{f_2}$ can be obtained from $\mathbf{F_1} = \mathbf{f_1}\mathbf{f_1^H}$ and $\mathbf{F_2} = \mathbf{f_2}\mathbf{f_2^H}$ by performing the eigenvalue decomposition. Otherwise, the Gaussian randomization can be used for recovering the approximate $\mathbf{f_1}$ and $\mathbf{f_2}$ [93]. Thus, the problem $\mathbf{P}_{5.2}$ can be solved by alternately updating $(t_U, t_{E,k})$ and $(\mathbf{f_1}, \mathbf{f_2})$, which is summarized in Algorithm 5.1.

5.3.3 Optimizing w with $(\mathbf{f}_1, \mathbf{f}_2)$

After obtaining the beamforming vectors $\mathbf{f}_1$ and $\mathbf{f}_2$, by setting $\mathbf{h}_{W,U} = \mathbf{H}_U\mathbf{f_1}$, $\mathbf{g}_{W,U} = \mathbf{G}_U\mathbf{f_2}$, $\mathbf{h}_{W,E,k} = \mathbf{H}_{E,k}\mathbf{f_1}$, and $\mathbf{g}_{W,E,k} = \mathbf{G}_{E,k}\mathbf{f_2}$, the SINR of user and eavesdroppers can be denoted as $\gamma_j = \frac{a_0|\overline{\mathbf{w}}^H\mathbf{h}_{W,j}|^2}{a_0|\overline{\mathbf{w}}^H\mathbf{g}_{W,j}|^2+1}$, $j \in \{U, (E,k)\}$. Similar to the previous section, let $\mathbf{W} =$

$\overline{\mathbf{w}\mathbf{w}}^H$, $\mathbf{H}_{W,j} = \mathbf{h}_{W,j}\mathbf{h}_{W,j}^H$ and $\mathbf{G}_{W,j} = \mathbf{g}_{W,j}\mathbf{g}_{W,j}^H$. The problem of $\mathbf{P}_{5.1}$ can be transformed into

$$\mathbf{P}_{5.7}: \max_{\mathbf{W}} \frac{1}{P_{tot}}\{\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\mathbf{H}_{W,U}\mathbf{W})}{a_0\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W})+1}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{a_0\mathrm{Tr}(\mathbf{H}_{W,E,k}\mathbf{W})}{a_0\mathrm{Tr}(\mathbf{G}_{W,E,k}\mathbf{W})+1})\}$$
$$s.t.\ \ (5.9b),\ (5.9c).$$

By applying Lemma 1 with SDR and introducing the variable $l_W \geq \max_{k\in K}\phi_{W,E,k}$, the problem $\mathbf{P}_{5.7}$ can be transformed into

$$\mathbf{P}_{5.8}: \max_{\mathbf{W},t_{W,U},t_{W,E,k}} \frac{1}{\frac{\ln 2}{B}P_{tot}}[\phi_{W,U}(\mathbf{W},t_{W,U}) - l_W]$$
$$s.t.\ \ \phi_{W,E,k}(\mathbf{W},t_{W,E,k}) \leq l_W, \tag{5.20a}$$
$$\phi_{W,U}(\mathbf{W},t_{W,U}) - l_W \geq \frac{\ln 2}{B}R_{th}, \tag{5.20b}$$
$$\mathbf{W}\succeq 0, \mathbf{W}_{mm}=1,\ m=1,2,...,M, \tag{5.20c}$$

where

$$\phi_{W,U} = \ln(1+a_0\mathrm{Tr}(\mathbf{G}_{W,U}+\mathbf{H}_{W,U})\mathbf{W}) - t_{W,U}(a_0\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W})+1) + \ln t_{W,U} + 1, \tag{5.21}$$

and

$$\phi_{W,E,k} = t_{W,E,k}(1+a_0\mathrm{Tr}(\mathbf{G}_{W,E,k}+\mathbf{H}_{W,E,k})\mathbf{W}) - \ln(a_0\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W})+1) - \ln t_{W,E,k} - 1. \tag{5.22}$$

Since the objective function is a concave-convex fractional function, By using the Dinkelbach's method [99], $\mathbf{P}_{5.8}$ can be solved by iteratively solving the following problem, given

as

$$\mathbf{P}_{5.9}: \max_{\mathbf{W},t_{W,U},t_{W,E,k}} \phi_{W,U}(\mathbf{W},t_{W,U}) - l_W - \frac{\ln 2}{B}\eta_2^* P_{tot}$$

$$s.t. \quad \phi_{W,E,k}(\mathbf{W},t_{W,E,k}) \leq l_W, \tag{5.23a}$$

$$\phi_{W,U}(\mathbf{W},t_{W,U}) - l_W \geq \frac{\ln 2}{B} R_{th}, \tag{5.23b}$$

$$\mathbf{W} \succeq 0, \mathbf{W}_{mm} = 1, \ m = 1,2,...,M, \tag{5.23c}$$

The problem $\mathbf{P}_{5.9}$ is a convex problem and can be solved by using the standard convex optimization tool. After the optimal $\mathbf{W}$ is obtained, $t_{W,U}$ and $t_{W,E,k}$ can be given as $t_{W,U}^* = (a_0\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W})+1)^{-1}$ and $t_{W,E,k}^* = (1+a_0\mathrm{Tr}(\mathbf{G}_{W,E,k}+\mathbf{H}_{W,E,k})\mathbf{W})^{-1}$. After obtaining $\mathbf{W}$, the $\overline{\mathbf{w}}$ can be given by eigenvalue decomposition if $\mathrm{rank}(\mathbf{W}) = 1$, otherwise, the Gaussian randomization can be used for recovering the approximate $\overline{\mathbf{w}}$ [93]. The reflection coefficients can be given by $w_m = \angle(\frac{\overline{w}_m}{\overline{w}_{M+1}})$, $m = 1,2,..,M$. The overall optimization algorithm for solving $\mathbf{P}_5$ is summarized in Algorithm 5.1, where δ is the threshold and T is the maximum number of iterations.

5.3.4 Convergence Analysis

For the convergence of the proposed algorithm, similar to [21], the proof is given as follows. Let $(\mathbf{W}^k, \mathbf{F}_1^k, \mathbf{F}_2^k)$ denote the feasible solution in the kth iteration, and let $J_{5.4}$ denote the objective function of $\mathbf{P}_{5.4}$. It can be seen that for the given $\mathbf{W}^{k+1}$ and $\mathbf{W}^k$ from two iterations, one has $J_{5.4}(\mathbf{W}^{k+1}, \mathbf{F}_1^{k+1}, \mathbf{F}_2^{k+1}) \overset{(a)}{\geq} J_{5.4}(\mathbf{W}^{k+1}, \mathbf{F}_1^k, \mathbf{F}_2^k) \overset{(b)}{\geq} J_{5.4}(\mathbf{W}^k, \mathbf{F}_1^k, \mathbf{F}_2^k)$, where (a) holds because for the given $\mathbf{W}^{k+1}$ in Algorithm 5.1, $(\mathbf{F}_1^{k+1}, \mathbf{F}_2^{k+1})$ are the optimal solutions of problem $\mathbf{P}_{5.4}$, and (b) holds because from the objective function of $\mathbf{P}_{5.7}$, we have $\ln[1+a_0\mathrm{Tr}(\mathbf{H}_u\mathbf{F}_1\mathbf{H}_u^H\mathbf{W}^{k+1})+\mathrm{Tr}(\mathbf{G}_u\mathbf{F}_2\mathbf{G}_u^H\mathbf{W}^{k+1})] - t_{W,U}[a_0\mathrm{Tr}(\mathbf{G}_u\mathbf{F}_2\mathbf{G}_u^H\mathbf{W}^{k+1})+1] + \ln t_{W,U}+1 = \ln[1+a_0\mathrm{Tr}(\mathbf{H}_u^H\mathbf{W}^{k+1}\mathbf{H}_u\mathbf{F}_1)+\mathrm{Tr}(\mathbf{G}_u^H\mathbf{W}^{k+1}\mathbf{G}_u\mathbf{F}_2)] - \ln(a_0\mathrm{Tr}(\mathbf{G}_u^H\mathbf{W}^{k+1}\mathbf{G}_u\mathbf{F}_2)+1) \geq \ln[1 + a_0\mathrm{Tr}(\mathbf{H}_u^H\mathbf{W}^k\mathbf{H}_u\mathbf{F}_1) + \mathrm{Tr}(\mathbf{G}_u^H\mathbf{W}^k\mathbf{G}_u\mathbf{F}_2)] - \ln(a_0\mathrm{Tr}(\mathbf{G}_u^H\mathbf{W}^k\mathbf{G}_u\mathbf{F}_2) + 1)$. Similarly, one has $R_{E,k}(\mathbf{W}^{k+1}, \mathbf{F}_1^k, \mathbf{F}_2^k) \leq R_{E,k}(\mathbf{W}^k, \mathbf{F}_1^k, \mathbf{F}_2^k)$. Therefore, the objective function of problem $\mathbf{P}_{5.4}$ is non-decreasing over the iterations in the proposed algorithm, and the ob-

jective value of $\mathbf{P}_{5.4}$ is finite due to the limited resource in the system. Thus, the proposed method is able to converge to a stationary point. A similar proof can be obtained for $\mathbf{P}_{5.7}$.

When the obtained solutions $\mathbf{F}_1$, $\mathbf{F}_2$, and $\mathbf{W}$ are not rank-one matrices, based on Gaussian randomization, a set of $\zeta_1 \sim \mathcal{CN}(0, \mathbf{F}_1)$, $\zeta_2 \sim \mathcal{CN}(0, \mathbf{F}_2)$, and $\zeta_3 \sim \mathcal{CN}(0, \mathbf{W})$ are generated. Then, the feasibility of $\mathbf{P}_5$ is checked with ζ_i, $i \in \{1, 2, 3\}$, and the monotonicity is also checked by comparing the current results with the results from the previous iteration. Via independently generating enough feasible ζ_i, $i \in \{1, 2, 3\}$, the optimal value of problem $\mathbf{P}_5$ can be approximated by the best ζ_i among all random vectors with an arbitrary small bias $\epsilon > 0$ [100].

The proposed method can provide a sub-optimal solution when the Gaussian randomization is applied. In Section V, we will compare the proposed method with the existing maximum ratio transmission (MRT) beamforming method to verify the superiority of our proposed scheme in terms of energy efficiency.

5.4 System Design With Imperfect CSI

In this section, based on the results obtained in Section 5.3, the energy efficiency maximization problem is extended into the case that the CSIs of the links from the IRS to the Eves are imperfect. The beamforming and jamming vectors and the phase shift matrix are jointly optimized to maximize the energy efficiency.

5.4.1 Problem Formulation

By considering the imperfect CSI model between the Eves and IRS, the energy efficiency maximization problem can be formulated as

$$\mathbf{P}_{5.2.1}: \max_{\mathbf{f}_1, \mathbf{f}_2, \boldsymbol{\Theta}} \ \eta$$

$$s.t. \ \ \mathbf{f}_1^H \mathbf{f}_1 \leq P_{1,\max}, \ \mathbf{f}_2^H \mathbf{f}_2 \leq P_{2,\max}, \tag{5.24a}$$

$$R_s \geq R_{th}, \ \Delta\mathbf{h}_{I,E,k} \in \mathcal{H}_{I,E,k}, \Delta\mathbf{g}_{I,E,k} \in \mathcal{G}_{I,E,k}, \tag{5.24b}$$

$$|\exp(j\theta_m)| = 1. \tag{5.24c}$$

Motivated by the method used for solving $\mathbf{P}_{5.1}$, the problem $\mathbf{P}_{5.2.1}$ can be also solved by using an alternating optimization method.

5.4.2 Optimizing the Beamforming with a Given $\mathbf{\Theta}$

In this section, we solve the problem $\mathbf{P}_{5.2.1}$ to achieve the optimal secure transmit beamforming vector $\mathbf{f}_1$ and jamming vector $\mathbf{f}_2$ for a given $\mathbf{\Theta}$.

Let $\mathbf{H}_{B,W} = \mathbf{\Theta}\mathbf{H}_{B,I}$ and $\mathbf{G}_{J,W} = \mathbf{\Theta}\mathbf{G}_{J,I}$ to simplify the formulas. By setting $\mathbf{H}_{B,E,k,X} = \begin{bmatrix} \mathbf{H}_{B,W} \\ \mathbf{h}_{B,E,k}^H \end{bmatrix}$, $\mathbf{G}_{J,E,k,X} = \begin{bmatrix} \mathbf{G}_{J,W} \\ \mathbf{g}_{J,E,k}^H \end{bmatrix}$, and introducing $\mathbf{h}_{I,E,k,X}=\overline{\mathbf{h}}_{I,E,k,X}+\Delta\mathbf{h}_{I,E,k,X}$, $\mathbf{g}_{I,E,k,X} = \overline{\mathbf{g}}_{I,E,k,X} + \Delta\mathbf{g}_{I,E,k,X}$, where $\overline{\mathbf{h}}_{I,E,k,X} = \begin{bmatrix} \overline{\mathbf{h}}_{I,E,k} \\ 1 \end{bmatrix}$, $\Delta\mathbf{h}_{I,E,k,X} = \begin{bmatrix} \Delta\mathbf{h}_{I,E,k} \\ 0 \end{bmatrix}$, $\overline{\mathbf{g}}_{I,E,k,X} = \begin{bmatrix} \overline{\mathbf{g}}_{I,E,k} \\ 1 \end{bmatrix}$, and $\Delta\mathbf{g}_{I,E,k,X} = \begin{bmatrix} \Delta\mathbf{g}_{I,E,k} \\ 0 \end{bmatrix}$, respectively, the SINR of Eve k can be reformulated as $\gamma_{E,k} = \frac{|(\mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X})\mathbf{f}_1|^2}{|(\mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X})\mathbf{f}_2|^2+\sigma^2}$. The problem $\mathbf{P}_{5.2.1}$ can be transformed into

$$\mathbf{P}_{5.2.2}\text{: } \max_{\mathbf{f}_1,\mathbf{f}_2} \frac{}{P_{tot}}\{\frac{B}{\ln 2}\ln(1+\frac{|\overline{\mathbf{w}}^H\mathbf{H}_U\mathbf{f}_1|^2}{|\overline{\mathbf{w}}^H\mathbf{G}_U\mathbf{f}_2|^2+\sigma^2}) - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{|(\mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X})\mathbf{f}_1|^2}{|(\mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X})\mathbf{f}_2|^2+\sigma^2})\}$$

$$s.t.\ \ \mathbf{f}_1^H\mathbf{f}_1 \le P_{1,max},\ \mathbf{f}_2^H\mathbf{f}_2 \le P_{2,max}, \tag{5.25a}$$

$$\frac{B}{\ln 2}\ln(1+\frac{|\overline{\mathbf{w}}^H\mathbf{H}_U\mathbf{f}_1|^2}{|\overline{\mathbf{w}}^H\mathbf{G}_U\mathbf{f}_2|^2+\sigma^2} - \max_{k\in K}\frac{B}{\ln 2}\ln(1+\frac{|(\mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X})\mathbf{f}_1|^2}{|(\mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X})\mathbf{f}_2|^2+\sigma^2})$$

$$\ge R_{th},\ \Delta\mathbf{h}_{I,E,k,X} \in \mathcal{H}_{I,E,k}, \Delta\mathbf{g}_{I,E,k,X} \in \mathcal{G}_{I,E,k}. \tag{5.25b}$$

Similar to the method used in Section 5.3, by defining $\mathbf{F}_1 = \mathbf{f}_1\mathbf{f}_1^H$ and $\mathbf{F}_2 = \mathbf{f}_2\mathbf{f}_2^H$, one has $\mathbf{F}_1 \succeq 0$, $\mathbf{F}_2 \succeq 0$ and $\text{rank}(\mathbf{F}_1) = \text{rank}(\mathbf{F}_2) = 1$. The rank-1 constraint makes problem hard to be solved. By applying the SDR method to relax the rank-1 constraints [101], the

problem $\mathbf{P}_{5.2.2}$ can be transformed into

$$\mathbf{P}_{5.2.3}: \max_{\mathbf{F}_1,\mathbf{F}_2} \frac{1}{P_{tot}}\Bigg(\frac{B}{\ln 2}[\ln(\mathrm{Tr}(\overline{\mathbf{H}}_U\mathbf{F}_1)+\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2)+\sigma^2)-\ln(\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2)+\sigma^2)]$$
$$-\max_{k\in K}\frac{B}{\ln 2}[\ln(\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}^H_{B,E,k,X}\mathbf{h}_{I,E,k,X}+\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)$$
$$-\ln(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)]\Bigg)$$
$$s.t.\ \ (\mathbf{F}_1,\mathbf{F}_2)\in\mathcal{F}, \tag{5.26a}$$
$$\frac{B}{\ln 2}[\ln(\mathrm{Tr}(\overline{\mathbf{H}}_U\mathbf{F}_1)+\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2)+\sigma^2)-\ln(\mathrm{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2)+\sigma^2)]$$
$$-\max_{k\in K}\frac{B}{\ln 2}[\ln(\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}^H_{B,E,k,X}\mathbf{h}_{I,E,k,X}+\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)$$
$$-\ln(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)]\geq R_{th}, \Delta\mathbf{h}_{I,E,k,X}\in\mathcal{H}_{I,E,k},\ \Delta\mathbf{g}_{I,E,k,X}\in\mathcal{G}_{J,E,k}, \tag{5.26b}$$

where $\mathcal{F}=\{(\mathbf{F}_1,\mathbf{F}_2)|\mathrm{Tr}(\mathbf{F}_1)\leq P_{1,max},\ \mathrm{Tr}(\mathbf{F}_2)\leq P_{2,max},\ \mathbf{F}_1\succeq 0,\ \mathbf{F}_2\succeq 0)\}$. Lemma 1 can be applied to solve the non-convexity caused by the second term in objective function and constraint (26b).

Let $x_{E,k}=\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}^H_{B,E,k,X}\mathbf{h}_{I,E,k,X}+\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2$ and $t=t_{E,k}$, the transmit rate of Eve k can be denoted as

$$\begin{aligned}R_{E,k}\frac{\ln 2}{B}&=\ln(\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}^H_{B,E,k,X}\mathbf{h}_{I,E,k,X}+\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)\\&\quad-\ln(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H{}_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)\\&=\min_{t_{E,k}\geq 0}\ \phi_{E,k}(\mathbf{F}_1,\mathbf{F}_2,t_{E,k}),\end{aligned} \tag{5.27}$$

where $\phi(t_{E,k})=t_{E,k}(\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}^H_{B,E,k,X}\mathbf{h}_{I,E,k,X}+\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)+\ln(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}^H_{J,E,k,X}\mathbf{g}_{I,E,k,X}+\sigma^2)-\ln(t_{E,k})-1$. Therefore, the problem

Table 5.1: Alternating Algorithm for Solving $\mathbf{P}_{5.1}$

Algorithm 5.1: Alternating Algorithm for Solving $\mathbf{P}_{5.1}$

1) **Input settings:**
   δ, $R_{th}, P_{1,max}, P_{2,max} > 0$, and T.
2) **Initialization:**
   $t_U(0)$, $t_{E,k}(0)$, $t_{W,U}(0)$, $t_{W,E,K}(0)$, $\mathbf{w}(0)$, $\eta(0)$;
3) **Optimization:**
   ⊵ **for** τ_1**=1:T**
      solve $\mathbf{P}_{5.6}$ with $(\mathbf{w}^*(\tau_1 - 1))$;
      obtain the solution $\mathbf{f}_1^*(\tau_1)$, $\mathbf{f}_2^*(\tau_1)$;
      solve $\mathbf{P}_{5.9}$ with $(\mathbf{f}_1^*(\tau_1), \mathbf{f}_2^*(\tau_1))$;
      obtain the solution $\mathbf{w}^*(\tau_1)$;
      calculate energy efficiency $\eta(\tau_1)$;
      if $\|\eta(\tau_1) - \eta(\tau_1 - 1)\| \leq \delta$;
         the optimal energy efficiency η^* is obtained;
      end
   ⊵ **end**
4) **Output:**
   $\{\mathbf{f}_1^*, \mathbf{f}_2^*, \mathbf{w}^*\}$ and energy efficiency η^*.

$\mathbf{P}_{5.2.3}$ can be transformed into

$$\mathbf{P}_{5.2.4}: \max_{\mathbf{F}_1,\mathbf{F}_2} \max_{t_U} \frac{1}{\frac{\ln 2}{B} P_{tot}} [\phi_U(\mathbf{F}_1, \mathbf{F}_2, t_U) - \min_{t_{E,k}} \phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2, t_{E,k})]$$

$$s.t. \quad (\mathbf{F}_1, \mathbf{F}_2) \in \mathcal{F}, \tag{5.28a}$$

$$\begin{aligned} &\max_{t_U} \phi_U(\mathbf{F}_1, \mathbf{F}_2, t_U) - \min_{t_{E,k}} \phi_{E,k}(\mathbf{F}_1, \mathbf{F}_2, t_{E,k}) \geq R_{th}, \\ &\Delta\mathbf{h}_{I,E,k,X} \in \mathcal{H}_{I,E,k}, \Delta\mathbf{g}_{I,E,k,X} \in \mathcal{G}_{J,E,k}. \end{aligned} \tag{5.28b}$$

By using Sion's minimax theorem [96], and introducing the slack variable $l \geq \max_{k\in K} \phi_{E,k}$, the problem $\mathbf{P}_{5.2.4}$ can be further transformed into

$$\mathbf{P}_{5.2.5} \quad \max_{\mathbf{F}_1,\mathbf{F}_2,t_U,t_{E,k}} \frac{\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U)-l}{\frac{\ln 2}{B}(\mathrm{Tr}(\mathbf{F}_1+\mathbf{F}_2)+P_{BS}+P_G+P_{IRS})}$$

$$\text{s.t.} \quad (\mathbf{F}_1,\mathbf{F}_2)\in\mathcal{F}, t_U, t_{E,k}\geq 0, \tag{5.29a}$$

$$\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U)-l \geq R_{th}\frac{\ln 2}{B}, \tag{5.29b}$$

$$\phi_{E,k}(\mathbf{F}_1,\mathbf{F}_2,t_{E,k})\leq l, \tag{5.29c}$$

$$\Delta\mathbf{h}_{I,E,k,X}\in\mathcal{H}_{I,E,k},\ \Delta\mathbf{g}_{I,E,k,X}\in\mathcal{G}_{I,E,k}. \tag{5.29d}$$

However, the problem is still difficult to be solved due to the uncertainty of the CSI from the IRS to the Eves. We introduce the slack variable $\psi_{B,E,k}$, and $\psi_{J,E,k}$ to deal with this uncertainty.

$$\mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H\mathbf{h}_{I,E,k,X}\leq\psi_{B,E,k}, \tag{5.30a}$$

$$\mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H\mathbf{g}_{I,E,k,X}\geq\psi_{J,E,k}. \tag{5.30b}$$

Then we have $\phi_{E,k}\leq t_{E,k}(\psi_{B,E,k}+\psi_{J,E,k}+\sigma^2)-\ln(\psi_{J,E,k}+\sigma^2)-\ln(t_{E,k})-1$. The problem $\mathbf{P}_{5.2.5}$ can be transformed into

$$\mathbf{P}_{5.2.6} \quad \max_{\mathbf{F}_1,\mathbf{F}_2,t_U,t_{E,k},\psi_{B,E,k},\psi_{J,E,k}} \frac{1}{\frac{\ln 2}{B}P_{tot}}[\phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U)-l]$$

$$\text{s.t.} \quad (5.29\text{a}),(5.29\text{b}),(5.30\text{a}),(5.30\text{b}),$$

$$t_{E,k}(\psi_{B,E,k}+\psi_{J,E,k}+\sigma^2)-\ln(\psi_{J,E,k}+\sigma^2) \tag{5.31a}$$

$$-\ln(t_{E,k})-1\leq l,\forall k. \tag{5.31b}$$

$\mathbf{P}_{5.2.6}$ can be solved by alternately solving $(t_U,t_{E,k})$ and $(\mathbf{F}_1,\mathbf{F}_1)$. First, with the given $(t_U^*,t_{E,k}^*)$, to solve the problem $\mathbf{P}_{5.2.6}$ for $(\mathbf{F}_1,\mathbf{F}_1)$, the $\mathcal{S}$-Procedure is applied.

Lemma 5.2: Let $f_i(\mathbf{z})=\mathbf{z}^H\mathbf{A}_i\mathbf{z}+2\Re(\mathbf{b}_i^H\mathbf{z})+c_i, i\in\{1,2\}$, where $\mathbf{z}\in\mathbb{C}^{M\times 1}$, $\mathbf{A}_i\in\mathbb{C}^{M\times M}$, $\mathbf{b_i}\in\mathbb{C}^{M\times 1}$, and $c_i\in\mathbb{R}$. Then, the expression $f_1(\mathbf{z})\leq 0\Rightarrow f_2(\mathbf{z})\leq 0$ holds if and

only if there exists a $\lambda \geq 0$ such that

$$\lambda \begin{bmatrix} \mathbf{A}_1 & \mathbf{b}_1 \\ \mathbf{b}_1^H & c_1 \end{bmatrix} - \begin{bmatrix} \mathbf{A}_2 & \mathbf{b}_2 \\ \mathbf{b}_2^H & c_2 \end{bmatrix} \succeq 0, \tag{5.32}$$

which assumes that there exists a vector $\overline{\mathbf{z}}$ such that $f(\overline{\mathbf{z}}) < 0$ [74]. By applying Lemma 5.2, let $\overline{\mathbf{h}}_{E,k,X} = \mathbf{H}_{B,E,k,X}^H \overline{\mathbf{h}}_{I,E,k,X}$ and $\overline{\mathbf{g}}_{E,k,X} = \mathbf{G}_{J,E,k,X}^H \overline{\mathbf{g}}_{I,E,k,X}$, the constraint (5.30a)-(5.30b) can be transformed into (5.33) and (5.34).

$$\begin{bmatrix} \lambda_{B,E,k}\mathbf{I} - \mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H & -\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H\overline{\mathbf{h}}_{I,E,k,X} \\ -\overline{\mathbf{h}}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H & \psi_{B,E,k} - \lambda_{B,E,k}\xi_{I,E,k}^2 - \overline{\mathbf{h}}_{E,k,X}^H\mathbf{F}_1\overline{\mathbf{h}}_{E,k,X} \end{bmatrix} \succeq 0, \tag{5.33}$$

$$\begin{bmatrix} \lambda_{J,E,k}\mathbf{I} + \mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H & \mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H\overline{\mathbf{g}}_{I,E,k,X} \\ \overline{\mathbf{g}}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H & -\lambda_{J,E,k}\xi_{J,E,k}^2 - \psi_{J,E,k} + \overline{\mathbf{g}}_{E,k,X}^H\mathbf{F}_2\overline{\mathbf{g}}_{E,k,X} \end{bmatrix} \succeq 0. \tag{5.34}$$

Then, similar to the previous section, by introducing the variable η_3^*, the optimization problem $\mathbf{P}_{5.2.6}$ for $\mathbf{F}_1$ and $\mathbf{F}_2$ based on t_U and $t_{E,k}$ can be given as

$$\mathbf{P}_{5.2.7} \quad \max_{\mathbf{F}_1,\mathbf{F}_2,\psi_{B,E,k},\psi_{J,E,k},\lambda_{B,E,k},\lambda_{J,E,k}} \phi_U(\mathbf{F}_1,\mathbf{F}_2,t_U) - l - \frac{\ln 2}{B}\eta^* P_{tot}$$

$$\text{s.t.} \quad (5.29\text{a}), (5.29\text{b}), (5.33), (5.34).$$

The problem $\mathbf{P}_{5.2.7}$ is a convex problem since the objective function and the constraints are all convex. It can be solved by using a standard convex optimization tool.

After $\mathbf{F_1}$ and $\mathbf{F_2}$ are obtained, if $\text{rank}(\mathbf{F_1}) = \text{rank}(\mathbf{F_2}) = 1$, $\mathbf{f_1}$ and $\mathbf{f_2}$ can be obtained from $\mathbf{F_1} = \mathbf{f_1}\mathbf{f_1^H}$ and $\mathbf{F_2} = \mathbf{f_2}\mathbf{f_2^H}$ by applying the eigenvalue decomposition. Otherwise, the Gaussian randomization can be used for recovering the approximate $\mathbf{f_1}$ and $\mathbf{f_2}$. After $\mathbf{f_1}$ and $\mathbf{f_2}$ are obtained, according to Lemma 1, the optimal value of t_U can be achieved when

$$t_U^* = (\text{Tr}(\overline{\mathbf{G}}_U\mathbf{F}_2) + \sigma^2)^{-1}. \tag{5.36}$$

To optimize $t_{E,k}$, the following problem should be solved.

$$\begin{aligned}\max_{t_{E,k}} -t_{E,k}(&\max_{\Delta\mathbf{h}_{I,E,k,X}} \mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H\mathbf{h}_{I,E,k,X} \\ &+ \min_{\Delta\mathbf{g}_{J,E,k,X}} \mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H\mathbf{g}_{I,E,k,X} + \sigma^2) + \ln(t_{E,k}) + 1.\end{aligned} \tag{5.37}$$

This needs to first solve the following problems, given as,

$$\Gamma_{1,k} = \max_{\Delta\mathbf{h}_{I,E,k,X}} \mathbf{h}_{I,E,k,X}^H\mathbf{H}_{B,E,k,X}\mathbf{F}_1\mathbf{H}_{B,E,k,X}^H\mathbf{h}_{I,E,k,X} \tag{5.38a}$$

$$\text{s.t. } \Delta\mathbf{h}_{I,E,k,X}^H\Delta\mathbf{h}_{I,E,k,X} \le \xi_{I,E,k}^2, \tag{5.38b}$$

and

$$\Gamma_{2,k} = \min_{\Delta\mathbf{g}_{J,E,k,X}} \mathbf{g}_{I,E,k,X}^H\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H\mathbf{g}_{I,E,k,X} + \sigma^2 \tag{5.39a}$$

$$\text{s.t. } \Delta\mathbf{g}_{I,E,k,X}^H\Delta\mathbf{g}_{I,E,k,X} \le \xi_{J,E,k}^2. \tag{5.39b}$$

For notational simplification, we denote $\mathbf{H}_{I,E,k,X}\mathbf{F}_1\mathbf{H}_{I,E,k,X}^H = \mathbf{F}_{1,k,X}$. Then, the Lagrangian function of problem (5.38) can be given as

$$L_{1,k} = (\overline{\mathbf{h}}_{I,E,k,X}^H + \Delta\mathbf{h}_{I,E,k,X}^H)\mathbf{F}_{1,k,X}(\overline{\mathbf{h}}_{I,E,k,X} + \Delta\mathbf{h}_{I,E,k,X}) + \mu_{1,k}(\xi_{I,E,k}^2 - \Delta\mathbf{h}_{I,E,k,X}^H\Delta\mathbf{h}_{I,E,k,X}), \tag{5.40}$$

where $\mu_{1,k}$ is the Lagrange multiplier. $L_{1,k}$ is convex respect to $\Delta\mathbf{h}_{I,E,k,X}$. The Karush-Kuhn-Tucker (KKT) condition can be applied to solve this problem. Thus, one has

$$\Gamma_{1,k} = \mathrm{tr}[\mathbf{F}_{1,k,X}(\overline{\mathbf{h}}_{I,E,k,X}\overline{\mathbf{h}}_{I,E,k,X}^H + \xi_{I,E,k}^2\mathbf{I} + 2\xi_{I,E,k}\sqrt{\frac{\overline{\mathbf{h}}_{I,E,k,X}^H\mathbf{F}_{1,k,X}\mathbf{h}_{I,E,k,X}}{\mathrm{tr}(\mathbf{F}_{1,k,X})}}\mathbf{I})]. \tag{5.41}$$

Similarly, letting $\mathbf{G}_{J,E,k,X}\mathbf{F}_2\mathbf{G}_{J,E,k,X}^H = \mathbf{F}_{2,k,X}$, one has

$$\Gamma_{2,k} = \mathrm{tr}[\mathbf{F}_{2,k,X}(\overline{\mathbf{g}}_{I,E,k,X}\overline{\mathbf{g}}_{I,E,k,X}^H + \xi_{J,E,k}^2\mathbf{I} - 2\xi_{J,E,k}\sqrt{\frac{\overline{\mathbf{g}}_{I,E,k,X}^H\mathbf{F}_{2,k,X}\mathbf{g}_{I,E,k,X}}{\mathrm{tr}(\mathbf{F}_{2,k,X})}}\mathbf{I})]. \tag{5.42}$$

The closed-form expression for solution for $t_{E,k}$ can be given as

$$t^*_{E,k} = (\Gamma_{1,k} + \Gamma_{2,k} + \sigma^2)^{-1}. \tag{5.43}$$

Thus, the problem $\mathbf{P}_{5.2.2}$ can be solved by alternately updating $(t_U, t_{E,k})$ and $(\mathbf{f_1}, \mathbf{f_2})$, which is summarized at Algorithm 5.2.

5.4.3 Optimizing w with Given $(\mathbf{f}_1, \mathbf{f}_2)$

After obtaining $\mathbf{f}_1$ and $\mathbf{f}_2$, by setting $\mathbf{H}_{B,E,k,F} = \begin{bmatrix} \mathrm{diag}(\mathbf{H}_{B,I}\mathbf{f}_1)^H & \\ & \mathbf{h}^H_{B,E,k}\mathbf{f}_1 \end{bmatrix}$, and $\mathbf{G}_{J,E,k,F} = \begin{bmatrix} \mathrm{diag}(\mathbf{G}_{J,I}\mathbf{f}_2)^H & \\ & \mathbf{g}^H_{J,E,k}\mathbf{f}_2 \end{bmatrix}$, the SINR of Eve k can be given as

$$\gamma_{E,k} = \frac{\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X}}{\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X} + \sigma^2}, \quad \forall k \in K, \tag{5.44}$$

where $\mathbf{W} = \overline{\mathbf{w}\mathbf{w}}^H$, $\mathbf{W} \succeq 0$, and $\mathrm{Rank}(\mathbf{W}) = 1$. The problem of $\mathbf{P}_{5.2.1}$ can be transformed into

$$\begin{aligned} \mathrm{P}_{5.2.8}: \max_{\mathbf{W}} \ & \frac{1}{P_{tot}}\Bigg[\frac{B}{\ln 2}\ln(1 + \frac{\mathrm{Tr}(\mathbf{H}_{W,U}\mathbf{W})}{\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W}) + \sigma^2}) \\ & - \max_{k\in K}\frac{B}{\ln 2}\ln(1 + \frac{\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X}}{\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X} + \sigma^2})\Bigg] \\ s.t. \ \ & \mathbf{W} \succeq 0, \mathrm{Rank}(\mathbf{W}) = 1, \mathbf{W}_{m,m} = 1, \ m = 1, 2, ..., M, \qquad (5.45a) \\ & R_s \geq R_{th}, \ \Delta\mathbf{h}_{I,E,k} \in \mathcal{H}_{I,E,k}, \Delta\mathbf{g}_{I,E,k} \in \mathcal{G}_{I,E,k}. \qquad (5.45b) \end{aligned}$$

Similar to the previous section, by applying Lemma 1 with SDR and introducing the variable $t_{W,U}$, $t_{W,E,K}$, and $l_W \geq \max_{k\in K}\phi_{W,E,k}$, the problem $\mathrm{P}_{5.2.8}$ can be transformed into

$$\mathrm{P}_{5.2.9}: \max_{\mathbf{W},t_{W,U},t_{W,E,k}} \frac{1}{\frac{\ln 2}{B}P_{tot}}[\phi_{W,U}(\mathbf{W},t_{W,U})-l_W]$$

$$s.t.\ \ \phi_{W,U}(\mathbf{W},t_{W,U})-l_W \geq \frac{\ln 2}{B}R_{th}, \tag{5.46a}$$

$$\phi_{W,E,k}(\mathbf{W},t_{W,E,k}) \leq l_W, \tag{5.46b}$$

$$\Delta\mathbf{h}_{I,E,k} \in \mathcal{H}_{I,E,k}, \Delta\mathbf{g}_{I,E,k} \in \mathcal{G}_{I,E,k}, \tag{5.46c}$$

$$\mathbf{W} \succeq 0, \mathbf{W}_{mm}=1,\ m=1,2,...,M, \tag{5.46d}$$

$$t_{W,U}>0, t_{W,E,k}>0,\ k=1,...,K, \tag{5.46e}$$

where

$$\begin{aligned}\phi_{W,E,k} =& t_{W,E,k}(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X}+\sigma^2+\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X})\\ &-\ln(\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X}+\sigma^2)-\ln t_{W,E,k}-1.\end{aligned} \tag{5.47}$$

To solve the uncertainty channel constraints, we introduce the variables $\psi^W_{B,E,k}$ and $\psi^W_{J,E,k}$, which are given as

$$\begin{aligned}&\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X} \leq \psi^W_{B,E,k},\\ &\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X} \geq \psi^W_{J,E,k}.\end{aligned} \tag{5.48}$$

Thus, problem $\mathbf{P}_{5.2.9}$ can be transformed into

$$\mathbf{P}_{5.2.10}: \max_{\mathbf{W},t_{W,U},t_{W,E,k},\psi^W_{B,E,k},\psi^W_{J,E,k}} \frac{[\phi_{W,U}(\mathbf{W},t_{W,U})-l_W]}{\frac{\ln 2}{B}P_{tot}}$$

$$s.t.\ \ (5.46a), (5.46c), (5.46d),$$

$$t_{W,E,k}(\psi^W_{J,E,k}+\sigma^2+\psi^W_{B,E,k})-\ln(\psi^W_{J,E,k}+\sigma^2)-\ln(t_{W,E,k})-1 \leq l_W, \tag{5.49a}$$

$$\mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X} \leq \psi^W_{B,E,k}, \Delta\mathbf{h}_{I,E,k} \in \mathcal{H}_{I,E,k}, \tag{5.49b}$$

$$\mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X} \geq \psi^W_{J,E,k}, \Delta\mathbf{g}_{I,E,k} \in \mathcal{G}_{I,E,k}. \tag{5.49c}$$

$$\mathbf{P}_{5.2.11}: \max_{\mathbf{W},t_{W,U},t_{W,E,k},\psi^W_{B,E,k},\psi^W_{J,E,k},\lambda^W_{B,E,k},\lambda^W_{J,E,k}} \phi_{W,U}(\mathbf{W},t_{W,U}) - l_W - \frac{\ln 2}{B}\eta_4^* P_{tot}$$

$$s.t.\ \ (5.46a),(46c),(46d),(49a),$$

$$\begin{bmatrix} \lambda^W_{B,E,k}\mathbf{I} - \mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F} & -\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\overline{\mathbf{h}}_{I,E,k,X} \\ -\overline{\mathbf{h}}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F} & -\lambda^W_{B,E,k}\xi^2_{I,E,k} + \psi^W_{B,E,k} - \overline{\mathbf{h}}^H_{E,k,F}\mathbf{W}\overline{\mathbf{h}}_{E,k,F} \end{bmatrix} \succeq 0, \tag{5.50a}$$

$$\begin{bmatrix} \lambda^W_{J,E,k}\mathbf{I} + \mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F} & \mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\overline{\mathbf{g}}_{I,E,k,X} \\ \overline{\mathbf{g}}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F} & -\lambda^W_{J,E,k}\xi^2_{J,E,k} - \psi^W_{J,E,k} + \overline{\mathbf{g}}^H_{E,k,F}\mathbf{W}\overline{\mathbf{g}}_{E,k,F} \end{bmatrix} \succeq 0. \tag{5.50b}$$

By using Lemma 2, letting $\overline{\mathbf{h}}_{E,k,F} = \mathbf{H}^H_{B,E,k,F}\overline{\mathbf{h}}_{I,E,k,X}$ $\overline{\mathbf{g}}_{E,k,F} = \mathbf{G}^H_{J,E,k,F}\overline{\mathbf{g}}_{J,E,k,X}$, and introducing the variable η_4^*, the problem $\mathbf{P}_{5.2.10}$ can be transformed as $\mathbf{P}_{5.2.11}$. The problem $\mathbf{P}_{5.2.11}$ is a convex problem with respect to $\mathbf{W}$ or $(t_{W,U}, t_{W,E,k})$ when other variables are fixed and can be solved by using a standard convex optimization tool. After obtaining $\mathbf{W}$, $\overline{\mathbf{w}}$ can be given by eigenvalue decomposition if $\mathrm{rank}(\mathbf{W}) = 1$; otherwise, the Gaussian randomization can be used for recovering the approximate $\mathbf{w}$. With the optimal $\mathbf{W}$, one has

$$t^*_{W,U} = (\mathrm{Tr}(\mathbf{G}_{W,U}\mathbf{W}) + \sigma^2)^{-1}. \tag{5.51}$$

And $t_{W,E,k}$ can obtained by solving the following problems.

$$\begin{aligned} \max_{t_{W,E,k}} & -t_{W,E,k}\Big(\min_{\Delta\mathbf{g}_{J,E,k,X}} \mathbf{g}^H_{I,E,k,X}\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F}\mathbf{g}_{I,E,k,X} \\ & + \max_{\Delta\mathbf{h}_{I,E,k,X}} \mathbf{h}^H_{I,E,k,X}\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F}\mathbf{h}_{I,E,k,X} + \sigma^2\Big) + \ln t_{W,E,k} + 1. \end{aligned} \tag{5.52}$$

Let $\mathbf{H}_{B,E,k,F}\mathbf{W}\mathbf{H}^H_{B,E,k,F} = \mathbf{W}_{B,k,X}$, and $\mathbf{G}_{J,E,k,F}\mathbf{W}\mathbf{G}^H_{J,E,k,F} = \mathbf{W}_{J,k,X}$. The solution for $t_{W,E,k}$ can be given as

$$t^*_{W,E,k} = (\Gamma_{W,1,k} + \Gamma_{W,2,k} + \sigma^2)^{-1}, \tag{5.53}$$

Table 5.2: Alternating Algorithm for Solving $\mathrm{P}_{5.2.1}$

Algorithm 5.2: Alternating Algorithm for Solving $\mathbf{P}_{5.2.1}$

1) **Input settings:**
   δ, $R_{th}, P_{1,\max}, P_{2,\max} > 0$, and T.
2) **Initialization:**
   $t_U(0)$, $t_{E,k}(0)$, $t_{W,U}(0)$, $t_{W,E,K}(0)$, $\mathbf{w}(0), \eta(0)$;
3) **Optimization:**
   for τ_2**=1:T**
      solve $\mathbf{P}_{5.2.2}$ with $(\mathbf{w}^*(\tau_2 - 1))$;
      obtain the solution $\mathbf{f}_1^*(\tau_2)$, $\mathbf{f}_2^*(\tau_2)$;
      solve $\mathbf{P}_{5.2.8}$ with $(\mathbf{f}_1^*(\tau_2), \mathbf{f}_2^*(\tau_2))$;
      obtain the solution $\mathbf{w}^*(\tau_2)$
      calculate energy efficiency $\eta(\tau_2)$;
      if $\|\eta(\tau_2) - \eta(\tau_2 - 1)\| \leq \delta$;
         the optimal energy efficiency η^* is obtained;
      end
   end
4) **Output:**
   $\{\mathbf{f}_1^*, \mathbf{f}_2^*, \mathbf{w}^*\}$ and energy efficiency η^*.

$\Gamma_{W,1,k}$ and $\Gamma_{W,2,k}$ are respectively given as

$$\Gamma_{W,1,k} = \mathrm{tr}[\mathbf{W}_{B,k,X}(\overline{\mathbf{h}}_{I,E,k,X}\overline{\mathbf{h}}_{I,E,k,X}^H + \xi_{I,E,k}^2\mathbf{I} + 2\xi_{I,E,k}\sqrt{\frac{\overline{\mathbf{h}}_{I,E,k,X}^H\mathbf{W}_{B,k,X}\mathbf{h}_{I,E,k,X}}{\mathrm{tr}(\mathbf{W}_{B,k,X})}}\mathbf{I})], \tag{5.54}$$

and

$$\Gamma_{W,2,k} = \mathrm{tr}[\mathbf{W}_{J,k,X}(\overline{\mathbf{g}}_{I,E,k,X}\overline{\mathbf{g}}_{I,E,k,X}^H + \xi_{J,E,k}^2\mathbf{I} - 2\xi_{J,E,k}\sqrt{\frac{\overline{\mathbf{g}}_{I,E,k,X}^H\mathbf{W}_{J,k,X}\mathbf{g}_{I,E,k,X}}{\mathrm{tr}(\mathbf{W}_{J,k,X})}}\mathbf{I})]. \tag{5.55}$$

The overall optimization algorithm for solving $\mathbf{P}_{5.2.1}$ is summarized in Algorithm 5.2, where δ is the threshold and T is the maximum number of iterations.

5.5 Simulation Results

In this section, simulation results are provided to verify the proposed algorithms. We consider a three-dimensional Cartesian coordinate system. The simulation settings are based on the work in [18], [93]. The locations of the base station, the Jammer, the IRS, and the legitimate user are respectively set as $(5, 0, 20)$, $(5, 0, 15)$, $(0, 100, 2)$, $(3, 100, 0)$ and the locations of 5 different Eves are set as $(2, 105, 0)$, $(2, 102.5, 0)$, $(2, 100, 0)$, $(2, 97.5, 0)$, $(2, 95, 0)$, respectively [93]. The channels are generated by the model $h_{i,j} = \sqrt{G_0 d_{i,j}^{-c_{i,j}}} g_{i,j}$, where $G_0 = -30$ dB is the path loss at the reference point. $d_{i,j}$, $c_{i,j}$ and $g_{i,j}$ denote the distance, path loss exponent, and fading between i and j, respectively, where $i \in \{B, J, I\}$ and $j \in \{U, (E, k)\}$ [102]. The path loss exponents are set as $c_{B,U} = c_{B,E,k} = c_{J,U} = c_{J,E,k} = 5$, $c_{B,J} = c_{G,J} = 3.5$, $c_{J,U} = 2$, and $c_{J,E,k} = 3$. We consider that the vertical location of the IRS is higher than those of the user and Eves. In this case, a less scattered environment is expected and one has $c_{B,I} \leq c_{B,i}$, $c_{J,I} \leq c_{J,i}$, $i \in \{B, (E, k)\}$. For the path loss exponents between IRS and the receivers, since IRS is deployed to support the legitimate user, it is assumed that the path loss between IRS and user is smaller than that of Eves, one has $c_{I,U} \leq c_{I,E,k}$. The bandwidth B is normalized to 1. The other parameters are set as $\xi_{I,E,k} = \xi_{J,E,k} = 10^{-4}$, $P_{1,max} = P_{2,max} = P_{max}$, $\sigma^2 = -105$ dBm, $\zeta = 1$, $P_{BS} = P_G = 23$ dBm, $P_{IRS} = 20$ dBm, and $\delta = 10^{-7}$.

Our proposed scheme for the perfect CSI model is marked as 'Efficiency-IRS'. The proposed scheme for the imperfect CSI model is marked as 'Robust-IRS'. We consider five cases as benchmarks to compare with our proposed method. The first benchmark optimizes the transmit rate, which is marked as 'Rate-IRS'. The second benchmark minimizes the transmit power, which is marked as 'Power-IRS'. The third benchmark without IRS is marked as 'Efficiency-NoIRS'. The fourth benchmark is the method that has IRS but no phase adjustment, which is marked as 'Efficiency-NoAngle'. The fifth benchmark is the method that is based on the maximum ratio transmission (MRT) method for beamforming design under perfect CSI case [103] and is marked 'MRT-IRS'.

Fig. 5.2 shows the energy efficiency versus the maximum transmit power achieved by different designs. The minimum secrecy rate threshold is set as $R_{th} = 0.5$ Bits/Hz/s. It

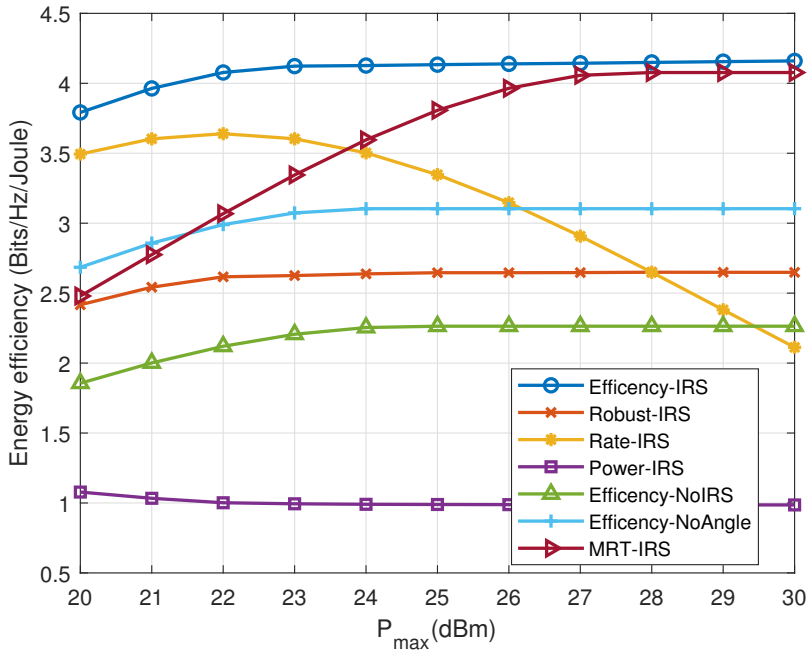

Fig. 5.2: Energy efficiency versus the maximum transmit power.

is observed that the energy efficiency achieved by the proposed method with the perfect CSI is the best among all the schemes. This indicates that our proposed IRS assisted cooperative jamming scheme is efficient in improving energy efficiency and achieving secure communications.

The system energy efficiency of the proposed method under the imperfect CSI condition is smaller than those achieved with the 'Efficiency-IRS' method, 'MRT-IRS' method, 'Efficiency-Noangle' method, and 'Rate-IRS' method at the beginning. This is because even without the phase optimization, IRS can help to increase the energy efficiency with the perfect CSI. Under the imperfect CSI, the energy efficiency degrades compared to that achieved under the perfect CSI case due to the CSI uncertainty. However, compared with the method without IRS, the 'Robust-IRS' method can still achieve a higher energy efficiency. This further indicates that the application of IRS is effective to improve energy efficiency even under the imperfect CSI.

It is worth noting that the system efficiencies obtained by the proposed method, the benchmark 'Efficiency-NoIRS', and 'Efficiency-NoAngle' all increase first with $P_{\max}$ and

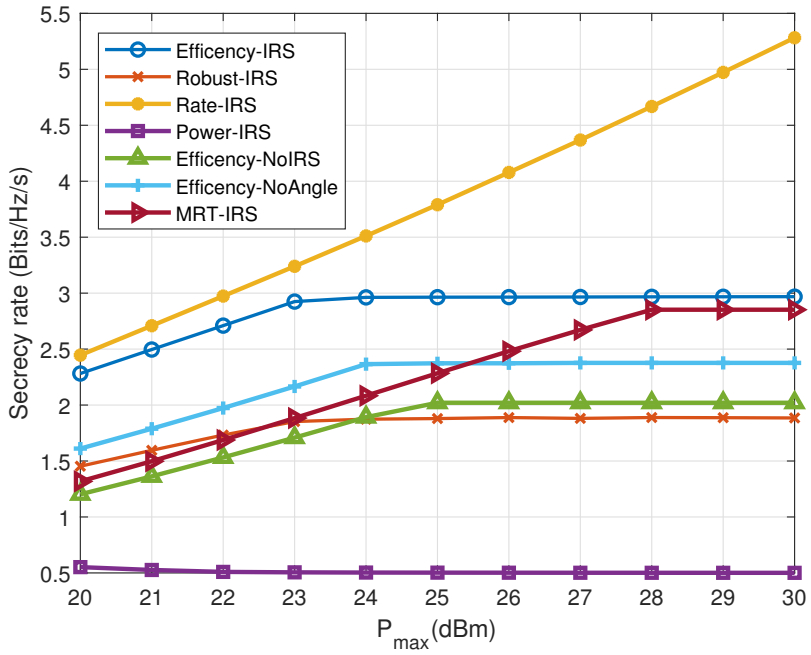

Fig. 5.3: Secrecy rate versus the maximum transmit power.

finally converge. For these methods, when the available power is limited, the increase of the secrecy rate is beneficial for the system to obtain a higher energy efficiency with only a slightly more power consumption. However, when the power availability is sufficient, e.g., $P_{\max}$ is larger than 23 dBm in this setting, further increase of the secrecy rate causes repaid elevation of the energy consumption, which leads to a decrease in energy efficiency. Similarly, the energy efficiency of the 'Rate-IRS' method first increases with the transmit power and then gradually decreases. The reason is that this method aims to maximize the secrecy rate without the constraint on the power consumption. Thus the study shows that there is a tradeoff between the energy efficiency and the secrecy rate. The energy efficiency of the 'Power-IRS' method first slightly decreases and then keeps at a low level. The reason is that this method aims to minimize power consumption, and thus it achieves the minimum secrecy rate R_{th} to save energy. In this case, both the energy efficiency and secrecy rate are relatively low. The energy efficiency of the 'MRT-IRS' method keeps increasing with $P_{\max}$ until reaching the highest efficiency, which is lower than that obtained with 'Efficiency-IRS' method. This validates the superiority of the proposed design.

Fig. 5.3 shows the achievable secrecy rate versus the maximum available transmits power $P_{\max}$. The secrecy rate obtained by the proposed scheme is comparable with the 'Rate-IRS' scheme when $P_{\max}$ is smaller than 23 dBm. When $P_{\max}$ is larger than 23 dBm, the 'Rate-IRS' method continues to use all the available energy to increase the achievable secrecy rate. The 'MRT-IRS' method shows a similar trend with the 'Efficiency-IRS' method but achieves a lower rate at the optimal level, which validates the observation in Fig. 5.2. However, the proposed scheme maintains the secrecy rate at a stable level in order to achieve the maximum energy efficiency. Similar trends can also be observed from the 'Robust-IRS' method, 'Efficiency-NoIRS' method, and 'Efficiency-NoAngle' method. The secrecy rate achieved by the 'Power-IRS' method first decreases and then stabilizes at the lowest level in order to save energy. The achievable secrecy rate of 'Robust-IRS' stabilizes at a lower level than other methods because based on the estimated channel quality, this algorithm needs to decrease the transmission rate to achieve the optimal energy efficiency under this setting. The curves in Fig. 5.3 indicate that with the aided IRS, our proposed method with the perfect CSI can achieve a higher secrecy rate and obtain the maximum energy efficiency.

Fig. 5.4 presents the power consumption for different methods versus $P_{\max}$. The results of all the methods in Fig. 5.4 are consistent with what have been shown in Fig. 5.2 and Fig. 5.3. It is worth noting that the power consumption by the proposed method with the perfect CSI and imperfect CSI are almost the same and both are quite low. This indicates that even with channel estimation errors, the 'Robust-IRS' method can still use less energy to achieve a higher rate, which demonstrates the advantage of the exploitation of IRS in improving energy efficiency.

Fig. 5.5 shows the energy efficiency versus the minimum secrecy rate threshold R_{th}. The maximum available transmit power is set to $P_{\max} = 36$ dBm. The energy efficiency achieved by the proposed method is the best among all the schemes. This indicates that the IRS assisted cooperative jamming can help guarantee the secrecy rate requirement and achieve the maximum energy efficiency. The energy efficiency of the proposed method, and the 'Efficiency-NoAngle' method initially maintain at a stable level and then decreases

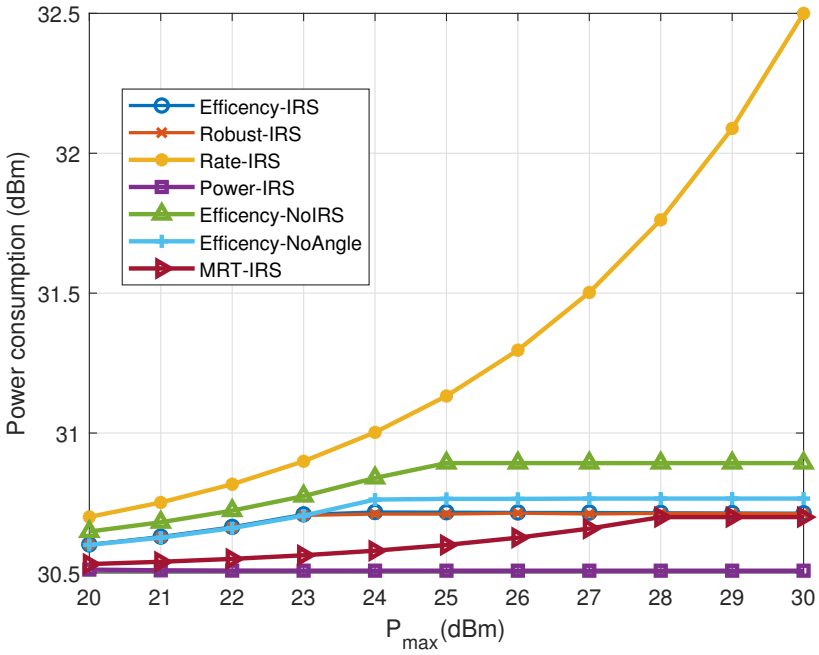

Fig. 5.4: Power consumption versus the maximum transmit power.

with the increase of R_{th}. When the minimum secrecy requirement is low, a higher rate can help the system to obtain a higher energy efficiency. However, when R_{th} is larger than the optimal rate, the system has to consume excessive energy to increase the secrecy rate in order to meet the minimum secrecy rate constraint, which causes the decrease of the energy efficiency.

In Fig. 5.5, the curves of the 'Efficiency-NoIRS' method, the 'Robust-IRS' method, the 'MRT-IRS' method, and the 'Efficiency-NoAngle' method vanish when R_{th} is larger than 1.5 Bits/Hz/s, 2 Bits/Hz/s, 2.5 Bits/Hz/s, and 3 Bits/Hz/s, respectively. The reason is that there is no feasible solution that can meet a higher R_{th} in those regions even with the maximum available transmit power. Moreover, the energy efficiency of the 'Power-IRS' method first increases and the curve starts to decrease when R_{th} is larger than 3 Bits/Hz/s, When the secrecy rate is smaller than 3 Bits/Hz/s, the increase of the secrecy rate can bring more performance gains (say rate gain) than the energy consumption. Thus, it results in the increase of the system energy efficiency. However, when the secrecy rate becomes larger and larger, the power cost for increasing the secrecy rate goes higher than the benefits that

it brings to the system, which causes a lower energy efficiency. This also indicates that there is a tradeoff between energy efficiency and the secrecy rate. The energy efficiency of 'Rate-IRS' stays at a constant level. This can be explained by the fact that the system uses all the available power to maximize the secrecy rate without considering the achievable energy efficiency. Thus, the curve does not change with the increase of R_{th}.

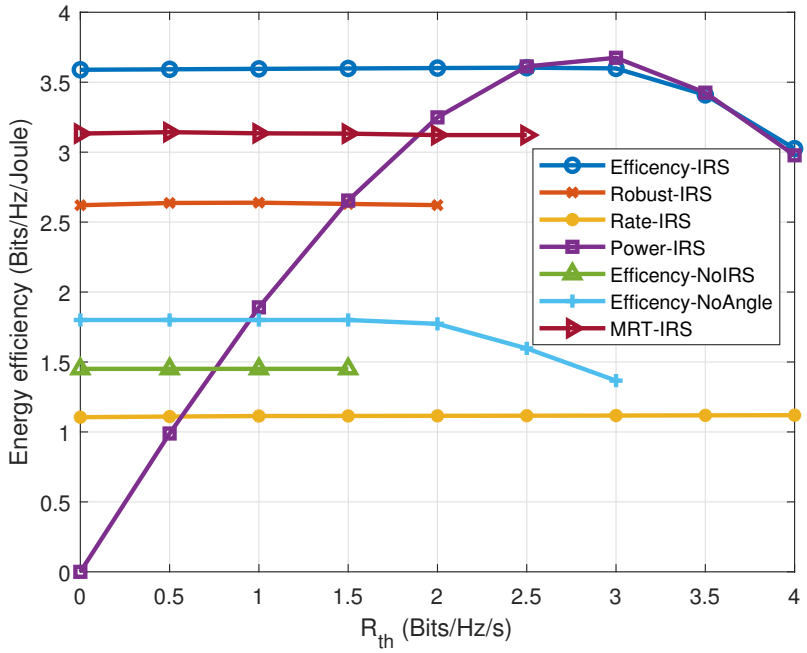

Fig. 5.5: Energy efficiency versus the secrecy rate threshold.

A comparison of the achievable secrecy rate versus the rate threshold R_{th} is presented in Fig. 5.6. The secrecy rates obtained by the proposed method, the 'Efficiency-NoIRS' method, and the 'Efficiency-NoAngle' method are first maintained at the stable level to guarantee the maximum energy efficiency. After R_{th} is larger than the optimal rate, the secrecy rate constraint enforces a linear increase of the rate with R_{th}. Similar to the reason for Fig. 5.5, the missing points are caused by lack of feasible solutions for the two benchmark schemes in certain R_{th} regions. With the assistance of the IRS, the system can use a smaller transmit power to achieve a higher secrecy rate. Additionally, the secrecy rate of the 'Power-IRS' method increases with the R_{th} linearly, which also verifies the observation

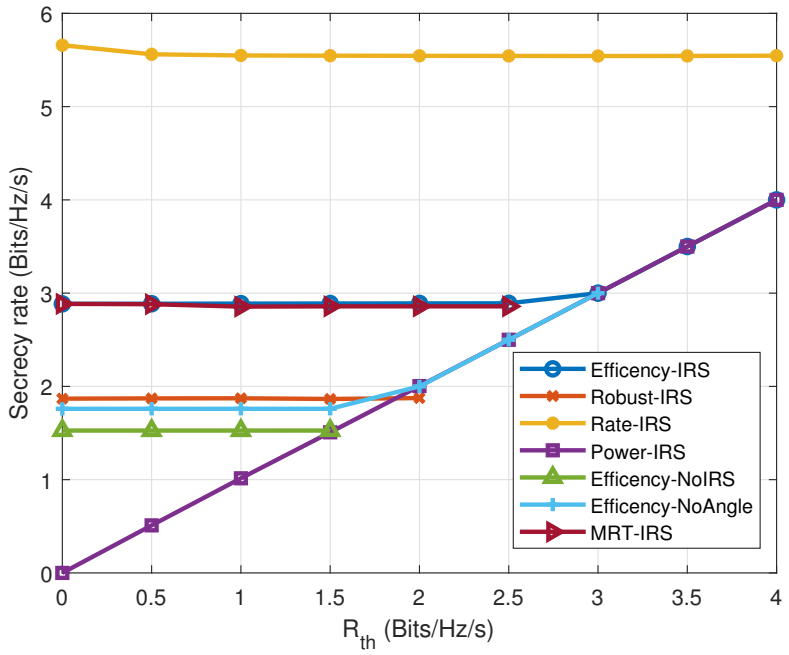

Fig. 5.6: Achievable secrecy rate versus the secrecy rate threshold.

in Fig. 5.5. For the 'Robust-IRS method, the system efficiency and secrecy rate are both higher than those of the 'Efficiency-NoIRS' method and the 'Efficiency-NoAngle' method under this setting, which indicates that even with imperfect CSI, the proposed method can still achieve a better performance than the method without IRS under perfect CSI. The secrecy rate achieved by the 'Rate-IRS' method is the largest among all the methods and remains constant.

Fig. 5.7 shows the energy efficiency versus the relative distance between the user and IRS. The curves for all the methods with IRS decrease with the increase of the distance. This is because the increase of the distance results in the increase of the path loss and the reduction of the power gain from the reflecting path through the IRS. Therefore, the achievable secure rate and energy efficiency both are decreased. It is also seen that the 'Efficiency-IRS' method still has the highest performance among all the methods, which validates the superiority of our proposed design.

Fig. 5.8 shows the achievable secure rate versus the relative distance. The trend is consistent with that shown in Fig. 5.7. It is worth noting that although the secrecy rate of

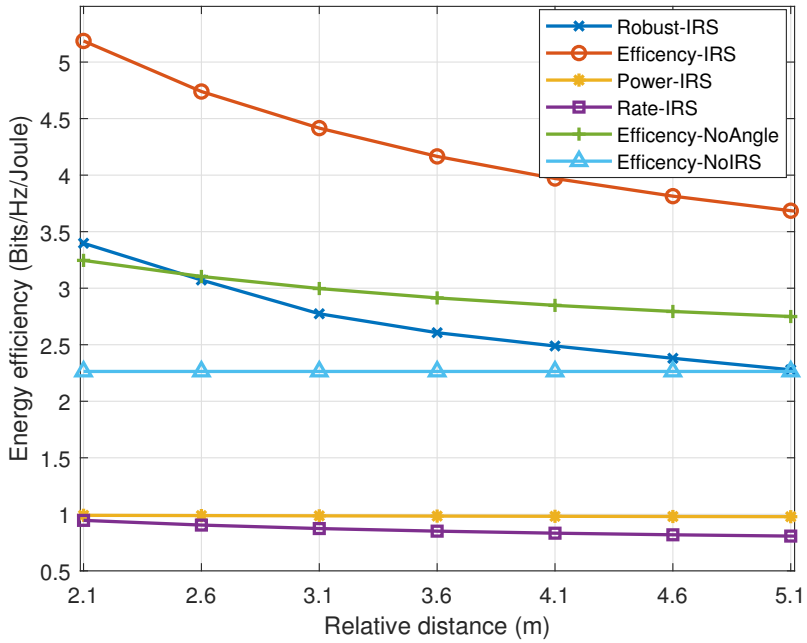

Fig. 5.7: Energy efficiency versus the relative distance of UE-IRS.

the 'Robust-IRS' method is lower than that obtained with 'Efficiency-NoIRS' method due to the uncertainty under the imperfect CSI, the energy efficiency of the 'Robust-IRS' is still larger than that achieved with the 'Efficiency-NoIRS' method. This further demonstrates the efficiency of the proposed robust design.

Fig. 5.9 shows energy efficiency versus the number of reflecting elements on the IRS. It is seen that the higher the number of the reflecting elements on the IRS, the better the energy efficiency obtained in the IRS- assisted network. The reason is that a better performance can be achieved by employing a higher number of reflecting elements to enhance the desired signals for the legitimate user. The increasing gain of the proposed method is higher than those obtained with other benchmarks, which demonstrates that the proposed scheme can effectively exploit the IRS to assist the secure transmission.

Fig. 5.10 shows the impact of the number of IRS elements on the achievable secure rate. It can be observed that both the 'Rate-IRS' method and the 'Efficiency-IRS' method achieve evident improvement on the secure rate. Compared with Fig. 5.9, it is worth noting that energy efficiency of the 'Rate-IRS' method is also increased. The reason is that the

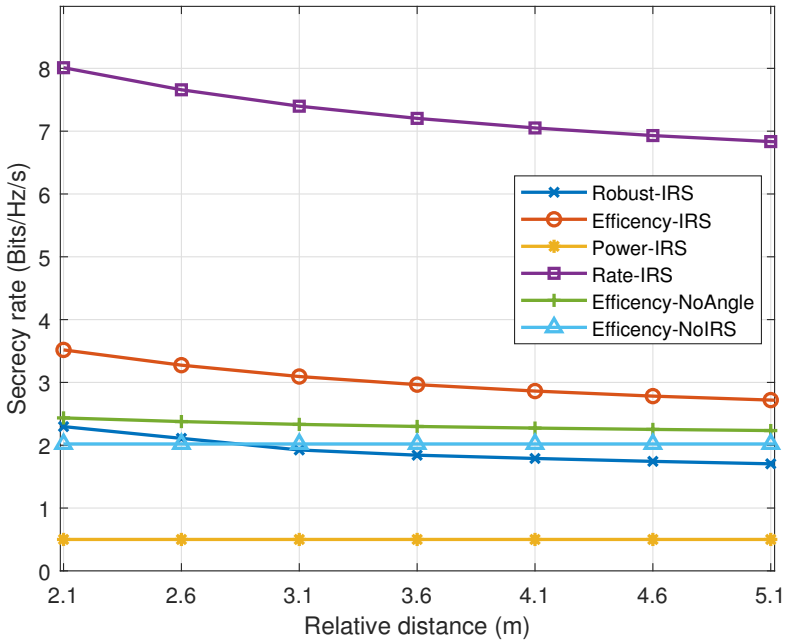

Fig. 5.8: Secrecy rate versus the relative distance of UE-IRS.

diversity gain can be achieved by increasing the number of the reflecting elements.

5.6 Chapter Conclusions

In this chapter, an IRS-assisted MISO wireless communication network was considered with the independently cooperative jamming in order to achieve secure communications. The energy efficiency was maximized by jointly optimizing the beamforming, jamming precode vectors, and IRS phase shift matrix under both perfect and imperfect CSI conditions. Two alternating algorithms were proposed to solve the challenging non-convex fractional optimization problems. It was shown that our proposed method outperforms other schemes in terms of energy efficiency. Although there is a tradeoff between the secrecy rate and energy efficiency, the application of IRS can effectively improve the energy efficiency even under the imperfect CSI case. The proposed alternating algorithm can be extended to the multi-IRS multi-UE MIMO communication network and the research will be done in our future works.

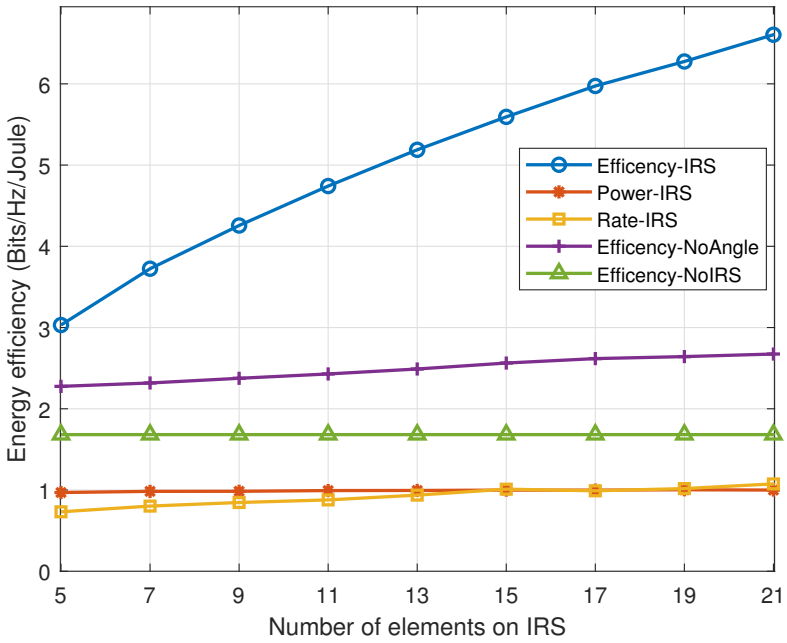

Fig. 5.9: Energy efficiency versus the number of elements on IRS.

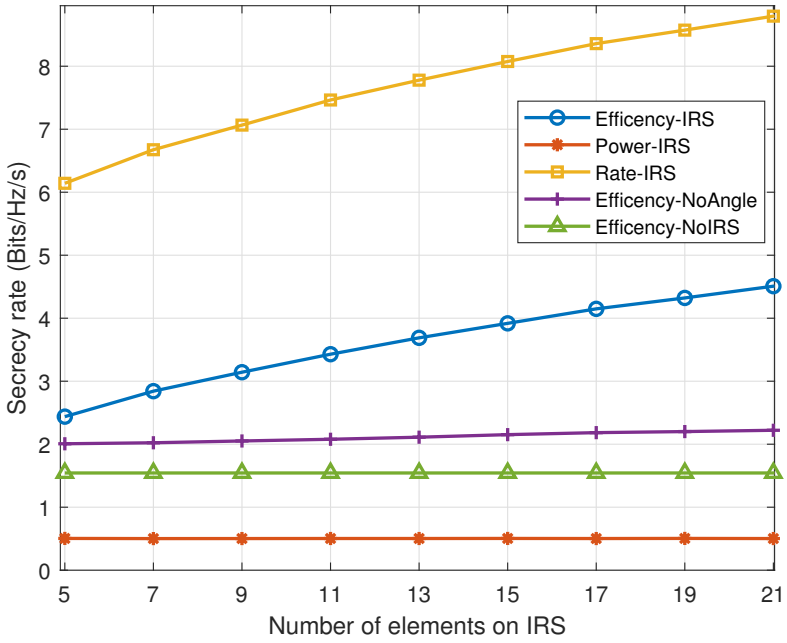

Fig. 5.10: Secrecy rate versus the number of elements on IRS.

CHAPTER 6

Energy-Efficient Design for IRS-Assisted MEC Networks with NOMA

6.1 Introduction

In the last chapter, we look at how the IRS can help the wireless network achieve a higher EE. Considering the NOMA-assisted MEC networks in previous chapters, combining IRS into this framework is imperative. However, there are new challenges that need to be overcome before this method can be successfully implemented.

The following studies have tested the efficacy of using IRS with NOMA. In [104], an IRS-assisted uplink NOMA system was considered to maximize the sum rate of all the users under the individual power constraint. The considered problem requires a joint power control at the users and beamforming design at the IRS, and an SDR-based solution has been developed. In [105], the problem of joint user association, subchannel assignment, power allocation, phase shifts design, and decoding order determination was formulated for maximizing the sum-rate for an IRS-assisted NOMA network. In [106], an EE algorithm was proposed to yield a tradeoff between the rate maximization and power minimization for an IRS-assisted NOMA network. The authors aimed to maximize the system EE by jointly optimizing the transmit beamforming and the reflecting beamforming. It was shown that NOMA can improve EE compared to OMA.

Furthermore, recently application of IRS into NOMA-based MEC networks has been studied. In [107], the authors investigated an IRS-aided MEC system with NOMA. By jointly optimizing the passive phase shifters, the size of transmission data, transmission rate, power, and time, as well as the decoding order, they aimed to minimize the sum energy consumption. A block coordinate descent method was developed to alternately optimize two separated subproblems. In [108], an IRS-aided MEC system was considered and a flexible time-sharing NOMA scheme was proposed to allow users to divide their data into

two parts that are transmitted via NOMA and TDMA respectively. By designing the IRS passive reflection and users' computation-offloading scheduling, the delay was minimized.

However, neither [107] nor [108] considered the EE performance of IRS-assisted MEC networks with NOMA, which is very important for system design to obtain the optimal trade-off between achievable rate and consumed power. Motivated by the above-mentioned observations, in this paper, the EE maximization problem is studied in an IRS-assisted MEC network with NOMA. To the authors' best knowledge, this is the first work that focuses on EE performance for applying both NOMA and IRS in the MEC network. The major contributions of this paper are summarized as follows.

We investigate the joint design of the receiver beamforming, offloading power, phase shift matrix, and local computing frequency to maximize the EE in an IRS-assisted MEC network with NOMA. The problem is challenging to solve due to its non-convexity fractional objective function and coupling of the beamforming vector with the IRS phase shift matrix. An alternating optimization algorithm is proposed to solve the non-convex fractional problem by using semidefinite programming relaxation (SDR). The simulation results show that the proposed method can achieve the highest EE among all the benchmarks.

6.2 System Model

As shown in Fig. 6.1, an IRS-assisted MEC system is considered. There are K single-antenna user equipments (UEs) in the system, which can do both local computing and data offloading. The access point (AP) with an MEC server is equipped with N antennas and the IRS has M reflecting elements.

6.2.1 Offloading Model

The baseband equivalent channel from UE k to IRS, IRS to AP, and UE k to AP are denoted as $\mathbf{h}_{I,U,k} \in \mathbb{C}^{1\times M}$, $\mathbf{H}_{B,I,k} \in \mathbb{C}^{M\times N}$, and $\mathbf{h}_{B,U,k} \in \mathbb{C}^{1\times N}$, respectively. In this paper, IRS adjusts its elements to maximize the combined incident signal from each UE to the AP. The diagonal phase-shift matrix can be denoted as $\mathbf{\Theta} = \text{diag}(\exp(j\theta_1), \exp(j\theta_2),$

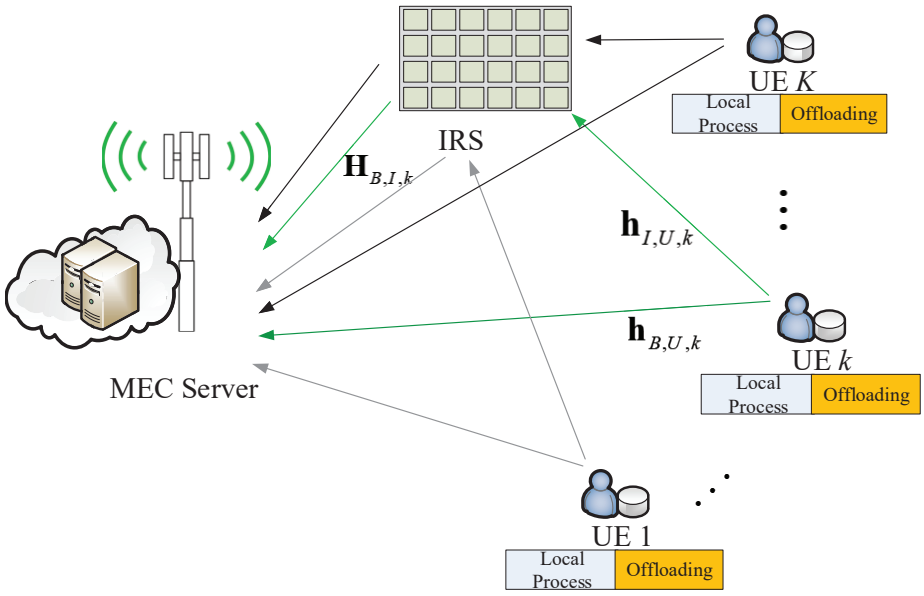

Fig. 6.1: An IRS-aided MEC system with NOMA.

$\cdots, \exp(j\theta_M))$, wherein its main diagonal, $\theta_m \in [0, 2\pi)$, denotes the phase shift on the combined incident signal by its mth element, $m = 1, 2, ..., M$ [95].

The transmitted signal from UE k is given as $\sqrt{p_k}s_k$, where $\sqrt{p_k}$ denotes the transmit power and s_k denotes the independent information. $\mathbf{m}_{B,k} \in \mathbb{C}^{N\times 1}$ denotes the receive beam vectors with unit norm, i.e., $\|\mathbf{m}_{B,k}\|^2 = 1$ [104]. Therefore, the signal received at AP can be given as

$$\mathbf{y}_{B,U} = \sum_{k=1}^{K}(\mathbf{h}_{B,U,k}^H + \mathbf{h}_{I,U,k}^H \Theta \mathbf{H}_{B,I,k})\mathbf{m}_{B,k}\sqrt{p_k}s_k + n_{B,U,k}, \tag{6.1}$$

where $n_{B,U,k} \sim \mathcal{CN}(0, \sigma^2)$ is the complex additive white Gaussian noise (AWGN) [106], [109].

NOMA is used to improve SE and mitigate the interference between different UEs. By exploiting the SIC techniques, the received signal at AP is sequentially decoded and the UE with the best channel conditions is firstly decoded. The channel of each UE includes a direct link and a reflect link. Since the reflect link depends on the unknown parameters $\mathbf{\Theta}$, the effective channels cannot be used to order the users at the receiver side. Similar to [104], we simply remove unknown reflect matrix by considering it as an identity matrix $\mathbf{I}$. UEs are then sorted based on this channel gain $|(\mathbf{h}_{B,U,k}^H + \mathbf{h}_{I,U,k}^H \mathbf{I}\mathbf{H}_{B,I,k})|$. Without loss of generality,

we assume that UEs are sorted in an increasing order, i.e., $|(\mathbf{h}_{B,U,1}^H + \mathbf{h}_{I,U,1}^H \mathbf{I}\mathbf{H}_{B,I,1})| \leq |(\mathbf{h}_{B,U,2}^H + \mathbf{h}_{I,U,2}^H \mathbf{I}\mathbf{H}_{B,I,2})| \leq \cdots \leq |(\mathbf{h}_{B,U,K}^H + \mathbf{h}_{I,U,K}^H \mathbf{I}\mathbf{H}_{B,I,K})|$. When decoding the signal for UE k, the signals from $i = 1, 2, \cdots, k-1$ are treated as interference. Thus, the signal to interference plus noise ratio (SINR) for UE k is expressed as

$$\gamma_{B,k} = \frac{p_k |(\mathbf{h}_{B,U,k}^H + \mathbf{h}_{I,U,k}^H \mathbf{\Theta} \mathbf{H}_{B,I,k}) \mathbf{m}_{B,k}|^2}{\sum_{i=1}^{k-1} p_i |(\mathbf{h}_{B,U,i}^H + \mathbf{h}_{I,U,i}^H \mathbf{\Theta} \mathbf{H}_{B,I,i}) \mathbf{m}_{B,i}|^2 + \sigma^2}. \tag{6.2}$$

The achievable offloading rate is

$$R_k^{off} = B \log_2(1 + \gamma_{B,k}). \tag{6.3}$$

6.2.2 Local Processing Model

Let C_k be the number of computation cycles required to process one bit of data for UE k locally. UE can compute and transmit simultaneously. Let f_k denote the computing frequency of the processor (cycles/second) [29]. Therefore, the local computing rate can be given as

$$R_k^{loc} = \frac{f_k}{C_k}. \tag{6.4}$$

The power consumption of local computing is modeled as a function of processor speed f_k. It can be given as $p_k^{loc} = \epsilon f_k^3$, where ϵ is effective capacitance coefficient of processor chip.

6.2.3 Energy Efficiency

The energy consumed by each UE consists of transmit power, local computing power, and circuit power consumption. Thus, the total power consumed by each UE is given as

$$P_k^{tot} = p_k + \epsilon f_k^3 + P_k^{cn}, \tag{6.5}$$

where P_k^{cn} denotes the constant circuit power consumed for signal processing and it is assumed to be the same for all UEs. The total achievable rate for each UE is

$$R_k^{tot} = R_k^{off} + R_k^{loc}. \tag{6.6}$$

According to [95], EE is defined as

$$\eta = \frac{\sum_{k=1}^{K} R_k^{tot}}{\sum_{k=1}^{K} P_k^{tot}}. \tag{6.7}$$

In order to maximize the EE, the local CPU frequency, offloading power, decoding vectors, and the phase shift matrix need to be jointly optimized.

6.3 Resource Optimization

In this section, the EE maximization problem is studied by jointly optimizing the local CPU frequency, offloading power, decoding vectors, and phase shift matrix. An alternating algorithm is further proposed to tackle the formulated problem.

6.3.1 Problem Formulation

The EE maximization problem is formulated as

$$\begin{aligned}
\mathbf{P}_{6.1}: \quad & \max_{p_k, f_k, \mathbf{m}_{B,k}, \boldsymbol{\Theta}} \ \eta \\
s.t. \quad & P_k^{tot} \leq P_k^{th}, && \text{(6.8a)} \\
& R_k^{tot} \geq R_{th}, && \text{(6.8b)} \\
& |\exp(j\theta_m)| = 1, && \text{(6.8c)} \\
& \|\mathbf{m}_{B,k}\|^2 = 1, && \text{(6.8d)}
\end{aligned}$$

where R_{th} is the minimum required rate threshold. P_k^{th} is the maximum available power of each UE. It is evident that problem $\mathbf{P}_{6.1}$ is non-convex due to the fractional structure of the objective function and the non-convex constraints. In order to tackle it, an alternating

algorithm is proposed.

By introducing $\mathbf{w}^H = [w_1, w_2, \cdots, w_M]$, one has $\mathbf{h}_{I,U,k}^H \mathbf{\Theta} \mathbf{H}_{B,I,k} = \mathbf{w}^H \mathbf{H}_{B,k}$, where $w_m = \exp(j\theta_m)$, $\mathbf{H}_{B,k} = \text{diag}(\mathbf{h}_{I,U,k}^H)\mathbf{H}_{B,I,k}$. Thus, the SINR of UE k is given as $\gamma_{B,k} = \frac{a_0 p_k |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,k} \mathbf{m}_{B,k}|^2}{a_0 \sum_{i=1}^{k-1} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1}$, where $a_0 = 1/\sigma^2$, $\overline{\mathbf{H}}_{B,k} = \begin{bmatrix} \mathbf{H}_{B,k} \\ \mathbf{h}_{B,U,k} \end{bmatrix}$, $\overline{\mathbf{w}}^H = \exp(j\overline{w})[\mathbf{w}^H, 1]$, and $\overline{w}$ is an arbitrary phase rotation. The objective of the optimization problem can be transformed into

$$\frac{\frac{B}{\ln 2}(\ln(1 + \sum_{k=1}^{K} a_0 p_k |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,k} \mathbf{m}_{B,k}|^2)) + \sum_{k=1}^{K} R_k^{loc}}{\sum_{k=1}^{K} P_k^{tot}}. \tag{6.9}$$

To tackle the complexity introduced by the logarithmic function of R_k^{off} in (6.8b), Lemma 1 is introduced. First, we have

$$R_k^{off} = \frac{B}{\ln 2}[\ln(a_0 \sum_{i=1}^{k} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1) - \ln(a_0 \sum_{i=1}^{k-1} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1)]. \tag{6.10}$$

Lemma 6.1: By introducing the function $\phi(t) = -tx + \ln t + 1$ for any $x > 0$, one has

$$-\ln x = \max_{t>0} \phi(t). \tag{6.11}$$

The optimal solution can be achieved at $t = 1/x$. By setting $x = a_0 \sum_{i=1}^{k-1} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1$, and $t = t_{B,k}$, one has

$$\begin{aligned} R_k^{off} &= \frac{B}{\ln 2} \max_{t_{B,k}>0} \phi_{B,k}(p_k, f_k, \mathbf{m}_{B,k}, \overline{\mathbf{w}}, t_{B,k}) \\ &= \frac{B}{\ln 2}[\ln(a_0 \sum_{i=1}^{k} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1) + \ln(t_{B,k}) + 1 - t_{B,k}(a_0 \sum_{i=1}^{k-1} p_i |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,i} \mathbf{m}_{B,i}|^2 + 1)]. \end{aligned} \tag{6.12}$$

By further introducing a variable η_1 to deal the fractional structure of (6.9), $\mathbf{P}_1$ can be transformed into

$$\mathbf{P}_{6.2}: \max_{p_k, f_k, \mathbf{m}_{B,k}, t_{B,k}, \overline{\mathbf{w}}} [\sum_{k=1}^{K} R_k^{loc} + \frac{B}{\ln 2}(\ln(1 + \sum_{k=1}^{K} a_0 p_k |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,k} \mathbf{m}_{B,k}|^2))] - \eta_1 \sum_{k=1}^{K} P_k^{tot}$$

$$s.t. \quad (6.8a), (6.8c), (6.8d),$$

$$\frac{B}{\ln 2}\phi_{B,k}(p_k, f_k, \mathbf{m}_{B,k}, \overline{\mathbf{w}}, t_{B,k}) + R_k^{loc} \geq R_{th}. \tag{6.13a}$$

$\mathbf{P}_{6.2}$ is still non-convex due to the coupling of variables. An alternating algorithm is proposed. To be specific, p_k and f_k are first optimized with a given $\mathbf{m}_{B,k}$, and $\overline{\mathbf{w}}$. $\mathbf{m}_{B,k}$ can then be optimized with the obtained p_k, f_k, and $\overline{\mathbf{w}}$. Further $\overline{\mathbf{w}}$ can be optimized with the obtained p_k, f_k, and $\mathbf{m}_{B,k}$. This process iteratively continues until convergence.

6.3.2 CPU Frequency and Offloading Power Optimization

With the given $\mathbf{m}_{B,k}$ and $\overline{\mathbf{w}}$, let $A_{B,k} = a_0 |\overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,k} \mathbf{m}_{B,k}|^2$, the problem can be transformed into

$$\mathbf{P}_{6.3}: \max_{p_k, f_k} \frac{B}{\ln 2}(\ln(\sum_{k=1}^{K} p_k A_{B,k} + 1)) + \sum_{k=1}^{K} \frac{f_k}{C_k} - \eta_1 \sum_{k=1}^{K} (\zeta p_k + \epsilon f_k^3 + P_k^{cn})$$

$$s.t. \quad p_k + \epsilon f_k^3 + P_k^{cn} \leq P_k^{th}, \tag{6.14a}$$

$$\frac{B}{\ln 2}\phi_{B,k}(p_k, f_k) + \frac{f_k}{C_k} \geq R_{th}. \tag{6.14b}$$

Problem $\mathbf{P}_{6.3}$ is convex with respect to f_k and p_k, therefore, it can be solved by using a standard convex optimization tool.

6.3.3 Optimizing the Receiving Beamforming

In this section, we solve the problem $\mathbf{P}_{6.2}$ to achieve the receive beamforming vector $\mathbf{m}_{B,k}$ for a given $\overline{\mathbf{w}}$, p_k, and f_k. Let $\overline{\mathbf{h}}_{B,k}^H = \overline{\mathbf{w}}^H \overline{\mathbf{H}}_{B,k}$, $\mathbf{P}_{6.3}$ can be transformed into

$$\begin{aligned}
\mathbf{P}_{6.4}: \ & \max_{\mathbf{m}_{B,k}} \ \frac{\ln 2}{B} \ln(a_0 \sum_{k=1}^{K} p_k |\overline{\mathbf{h}}_{B,k}^H \mathbf{m}_{B,k}|^2 + 1) + \sum_{k=1}^{K} R_k^{loc} - \eta_1 \sum_{k=1}^{K} P_k^{tot} \\
& s.t. \ \ \frac{B}{\ln 2} \ln(a_0 \sum_{i=1}^{k} p_i |\overline{\mathbf{h}}_{B,k}^H \mathbf{m}_{B,i}|^2 + 1) + \ln(t_{B,k}) + 1 \\
& \quad - t_{B,k}(a_0 \sum_{i=1}^{k-1} p_i |\overline{\mathbf{h}}_{B,i}^H \mathbf{m}_{B,i}|^2 + 1) + R_k^{loc} \geq R_{th}, \qquad (6.15a) \\
& \quad \|\mathbf{m}_{B,k}\|^2 = 1. \qquad (6.15b)
\end{aligned}$$

Let $|\overline{\mathbf{h}}_{B,k}^H \mathbf{m}_{B,k}|^2 = \mathrm{Tr}(\tilde{\mathbf{H}}_{\mathrm{B,k}} \mathbf{m}_{\mathrm{B,k}} \mathbf{m}_{\mathrm{B,k}}^{\mathrm{H}})$. By defining $\tilde{\mathbf{H}}_{B,k} = \overline{\mathbf{h}}_{B,k} \overline{\mathbf{h}}_{B,k}^H$, $\mathbf{M}_{B,k} = \mathbf{m}_{B,k} \mathbf{m}_{B,k}^H$, one has $\mathbf{M}_{B,k} \succeq 0$ and $\mathrm{rank}(\mathbf{M}_{B,k}) = 1$. The rank-1 constraint makes the problem difficult to solve. Thus, we apply the SDR method to relax the constraints [110]. $\mathbf{P}_{6.4}$ is then expressed as

$$\begin{aligned}
\mathbf{P}_{6.5}: \ & \max_{\mathbf{M}_{B,k}} \ [\frac{B}{\ln 2} \ln(a_0 \sum_{k=1}^{K} p_k \mathrm{Tr}(\tilde{\mathbf{H}}_{\mathrm{B,k}} \mathbf{M}_{\mathrm{B,k}}) + 1) + \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{R}_{\mathrm{k}}^{\mathrm{loc}}] - \eta_1 \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{P}_{\mathrm{k}}^{\mathrm{tot}} \\
& s.t. \ \ \frac{B}{\ln 2} \ln(a_0 \sum_{i=1}^{k} p_i \mathrm{Tr}(\tilde{\mathbf{H}}_{\mathrm{B,i}} \mathbf{M}_{\mathrm{B,i}}) + 1) + \ln(\mathrm{t}_{\mathrm{B,k}}) + 1 \\
& \quad - t_{B,k}(a_0 \sum_{i=1}^{k-1} p_i \mathrm{Tr}(\tilde{\mathbf{H}}_{\mathrm{B,i}} \mathbf{M}_{\mathrm{B,i}}) + 1) + \mathrm{R}_{\mathrm{k}}^{\mathrm{loc}} \geq \mathrm{R}_{\mathrm{th}}, \qquad (6.16a) \\
& \quad \mathrm{Tr}(\mathbf{M}_{\mathrm{B,k}}) = 1. \qquad (6.16b)
\end{aligned}$$

$\mathbf{P}_{6.5}$ is convex and can be solved by using a standard convex optimization tool [93]. After $\mathbf{M}_{B,k}$ is obtained, if $\mathrm{rank}(\mathbf{M}_{B,k}) = 1$, $\mathbf{m}_{B,k}$ can be obtained from $\mathbf{M}_{B,k} = \mathbf{m}_{B,k} \mathbf{m}_{B,k}^H$ by performing the eigenvalue decomposition. Otherwise, the Gaussian randomization can be used for recovering $\mathbf{m}_{B,k}$ [93].

6.3.4 Optimizing the IRS Reflecting Shifts $\overline{\mathbf{w}}$

After obtaining the beamforming vectors $\mathbf{m}_{B,k}$, by setting $\mathbf{h}_{W,B,k} = \overline{\mathbf{H}}_{B,k}\mathbf{m}_{B,k}$, problem $\mathbf{P}_{6.2}$ can be transformed into

$$\mathbf{P}_{6.6}: \max_{\overline{\mathbf{w}}} \ \frac{B}{\ln 2}\ln(a_0\sum_{k=1}^{K}p_k(\overline{\mathbf{w}}^H\mathbf{h}_{W,B,k})+1)+\sum_{k=1}^{K}\frac{f_k}{C_k}-\eta_1\sum_{k=1}^{K}P_k^{tot}$$

$$s.t. \ \ |w_m| = 1, \ m = 1, 2, ..., M, \tag{6.17a}$$

$$\ln(a_0\sum_{i=1}^{k}p_i(\overline{\mathbf{w}}^H\mathbf{h}_{W,B,i})+1)+\ln(t_{B,k})+1-t_{B,k}(a_0\sum_{i=1}^{k-1}p_i(\overline{\mathbf{w}}^H\mathbf{h}_{W,B,i})+1)+\frac{f_k}{C_k}\geq R_{th}. \tag{6.17b}$$

Similar to the previous section, let $\mathbf{W} = \overline{\mathbf{w}}\overline{\mathbf{w}}^H$, $\mathbf{H}_{W,B,k} = \mathbf{h}_{W,B,k}\mathbf{h}_{W,B,k}^H$. By applying the SDR method, we have

$$\mathbf{P}_{6.7}: \max_{\mathbf{W}} \ \frac{B}{\ln 2}\ln(a_0\sum_{k=1}^{K}p_k\mathrm{Tr}(\mathbf{H}_{W,B,k}\mathbf{W})+1)+\sum_{k=1}^{K}\frac{f_k}{C_k}-\eta_1\sum_{k=1}^{K}P_k^{tot}$$

$$s.t. \ \ \mathbf{W}\succeq 0, \mathbf{W}_{mm} = 1, \ m = 1, 2, ..., M, \tag{6.18a}$$

$$\begin{aligned}&\ln(a_0\sum_{i=1}^{k}p_i\mathrm{Tr}(\mathbf{H}_{W,B,k}\mathbf{W})+1)+\ln(t_{B,k})+1\\&-t_{B,k}(a_0\sum_{i=1}^{k-1}p_i\mathrm{Tr}(\mathbf{H}_{W,B,k}\mathbf{W})+1)+\frac{f_k}{C_k}\geq R_{th}.\end{aligned} \tag{6.18b}$$

The problem $\mathbf{P}_{6.7}$ is a convex problem and can be solved by using a standard convex optimization tool. After obtaining $\mathbf{W}$, $\overline{\mathbf{w}}$ can be given by eigenvalue decomposition if $\mathrm{rank}(\mathbf{W}) = 1$; otherwise, the Gaussian randomization can be used for recovering the approximate $\overline{\mathbf{w}}$ [93]. The reflection coefficients can be given by $w_m = \angle(\frac{\overline{w}_m}{\overline{w}_{M+1}})$, $m = 1, 2, .., M$. The overall optimization algorithm is summarized in Algorithm 6.1, where δ is the threshold and T is the maximum number of iterations.

6.4 Simulation Results

In this section, simulation results are provided to evaluate the performance of the

Table 6.1: Alternating Algorithm for Solving $\mathbf{P}_{6.1}$

Algorithm 6.1: Alternating Algorithm for Solving $\mathbf{P}_{6.1}$

1) **Input settings:**
   δ, $R_{th}, P_k^{th} > 0$, and T.
2) **Initialization:**
   $t_{B,k}(0)$, $\overline{\mathbf{w}}(0)$,$\mathbf{m}_{B,k}(0)$, and $\eta_1(0)$;
3) **Optimization:**
   ⊵ **for** τ_1**=1:T**
      solve $\mathbf{P}_{6.4}$ with $\overline{\mathbf{w}}^*(\tau_1 - 1)$,$\mathbf{m}^*_{B,k}(\tau_1 - 1)$;
      obtain the solution $p_k^*(\tau_1)$, $f_k^*(\tau_1)$;
      solve $\mathbf{P}_{6.5}$ with $p_k^*(\tau_1)$, $f_k^*(\tau_1)$, and $\overline{\mathbf{w}}^*(\tau_1 - 1)$;
      obtain the solution $\mathbf{m}^*_{B,k}(\tau_1)$;
      solve $\mathbf{P}_{6.7}$ with $p_k^*(\tau_1)$, $f_k^*(\tau_1)$, and $\mathbf{m}^*_{B,k}(\tau_1)$;
      obtain the solution $\overline{\mathbf{w}}^*(\tau_1)$;
      calculate EE $\eta(\tau_1)$ and update $t_{B,k}(\tau_1)$ and $\eta_1(\tau_1)$;
      if $|\frac{\eta(\tau_1)-\eta(\tau_1-1)}{\eta(\tau_1)}| \leq \delta$;
         the optimal EE η^* is obtained;
      end
   ⊵ **end**
4) **Output:**
   p_k^*, f_k^*, $\mathbf{m}^*_{B,k}$, and $\mathbf{w}^*$ and EE η^*.

proposed algorithms. We consider a three-dimensional Cartesian coordinate system. The simulation settings are based on those used in [18], [93]. We consider a 2-UE case and it can be readily extend to multiple UE cases. The locations of the MEC, the IRS, UE1, and UE2 are set as $(5, 0, 20)$, $(0, 50, 2)$, $(5, 75, 5)$ and $(5, 50, 10)$, respectively [93]. The channels are generated by $h_{i,j} = \sqrt{G_0 d_{i,j}^{-c_{i,j}}} g_{i,j}$, where $G_0 = -30$ dB is the path loss at the reference point. $d_{i,j}$, $c_{i,j}$ and $g_{i,j}$ denote the distance, path loss exponent, and fading between i and j, respectively, where $i \in \{B, I\}$ and $j \in \{U, k\}$. The path loss exponents are set as $c_{B,U,k} = 5$, $c_{B,I} = 3.5$, and $c_{I,U,k} = 2$. The bandwidth B is set to 1 Mhz. Other parameters are set as $\sigma^2 = -105$ dBm, $P_k^{th} = 31$ dBm, $P_k^{cn} = 23$ dBm, $C_k = 10^3$ cycles/bit, and $\epsilon = 10^{-28}$.

The proposed scheme is marked as 'Efficiency-IRS'. We consider three other cases as benchmarks to compare with the proposed method. The first benchmark, marked as 'OMA-IRS', uses FDMA with equally allocated bandwidth to all the users. The second benchmark, marked as 'OnlyOff-IRS', has no local computing and all the tasks are offloaded. The third benchmark, marked as 'Efficiency-NoIRS", aims to investigate the performance without IRS.

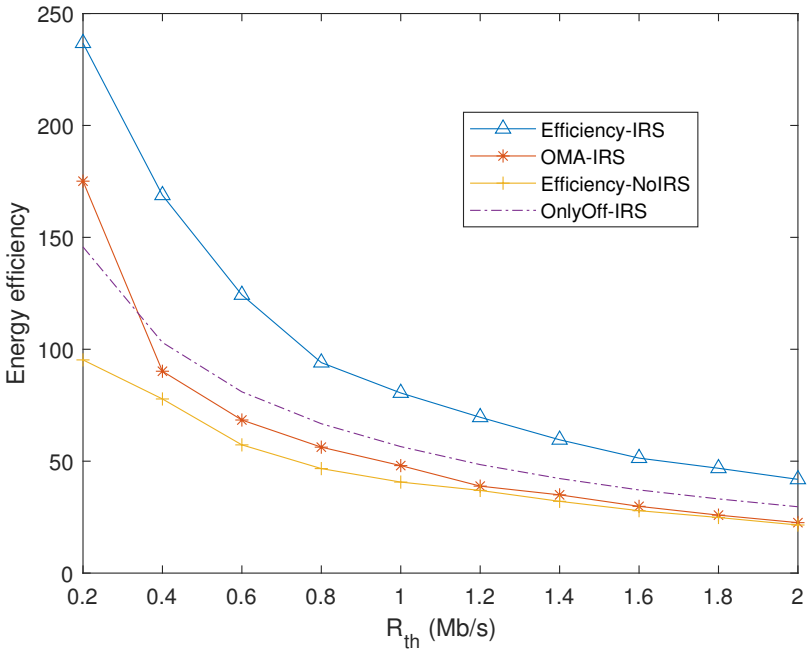

Fig. 6.2: Energy efficiency versus the minimum rate threshold.

Fig. 6.2 shows EE versus the minimum rate threshold R_{th}. The EE achieved by the proposed method is the best among all the schemes. This indicates that the IRS assisted MEC with NOMA can help improve the system rate and achieve high EE. With the increase of R_{th}, all the curves are decreasing. The system has to consume excessive energy to increase the rate in order to meet the minimum rate constraint, which decreases EE.

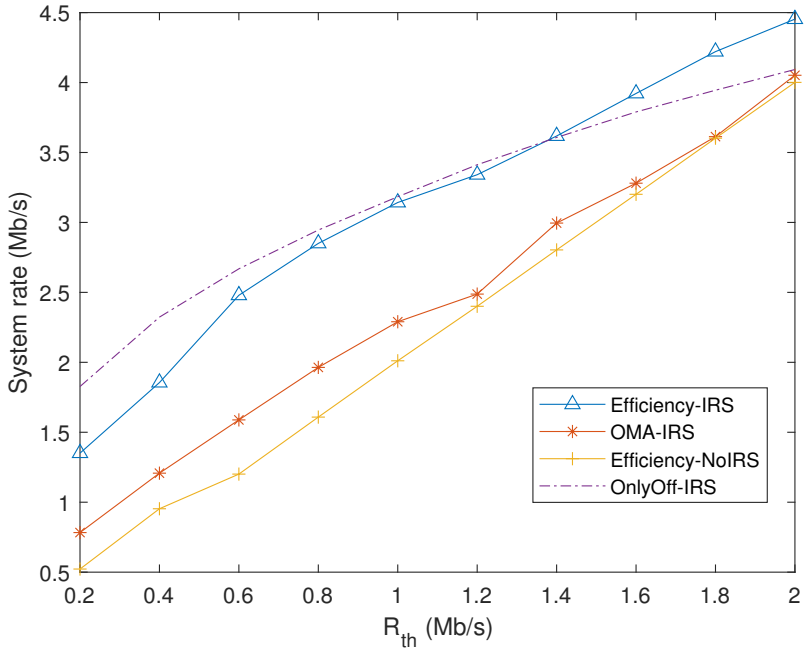

Fig. 6.3: Achievable rate versus the minimum rate threshold.

A comparison of the system rate versus the rate threshold R_{th} is presented in Fig. 6.3. All the curves increase with R_{th} in order to meet the service requirement, which verifies the observation in Fig. 6.2. The system rates obtained by the proposed method and the 'OnlyOff-IRS' method are higher than those of the other two methods, which indicates that combining IRS with NOMA can significantly help the system to achieve a higher rate. It is worth noting that even though the 'OnlyOff-IRS' method can achieve the highest rate when R_{th} is low, its efficiency is lower than the proposed method. This indicates that the overall efficiency performance degrades when there is no local computing.

Fig. 6.4 presents the power consumption versus R_{th} for different methods. The results

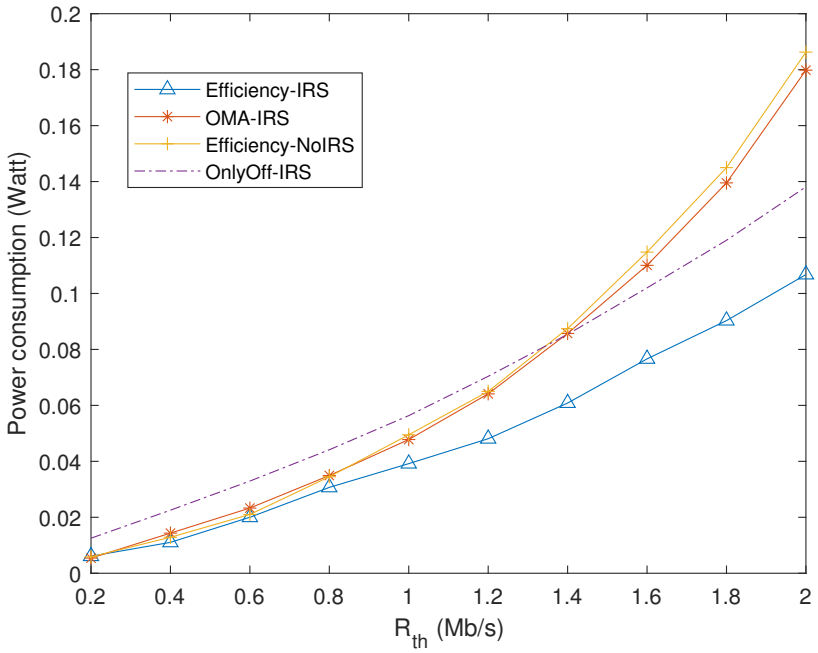

Fig. 6.4: Power consumption versus the minimum rate threshold.

of all the methods in Fig. 6.4 are consistent with what are shown in Fig. 6.2 and Fig. 6.3. It is worth noting that the power consumption by the proposed method is quite low. So UEs can use less energy to achieve a higher rate, which demonstrates the advantage of combining NOMA and IRS to MEC network in improving EE.

Fig. 6.5 shows EE versus the distance between UEs and IRS. The distance is the relative increased amount compared with UEs' original position. The curves for all the methods with IRS decrease with the increase of the distance except 'Efficiency-NoIRS'. This is because the increase of the distance results in the increase of the path loss and the reduction of the power gain from the reflecting path through the IRS. Therefore, the achievable rate and EE are both decreased. It can also be seen that the 'Efficiency-IRS' method still has the highest performance among all the methods, which validates the superiority of the proposed design.

In Fig. 6.6, the coverage of proposed methods based on different R_{th} setting are investigated. It can be observed from Fig. 6.6 that only several iterations are needed for the proposed algorithms to converge, showing the computation efficiency of the proposed

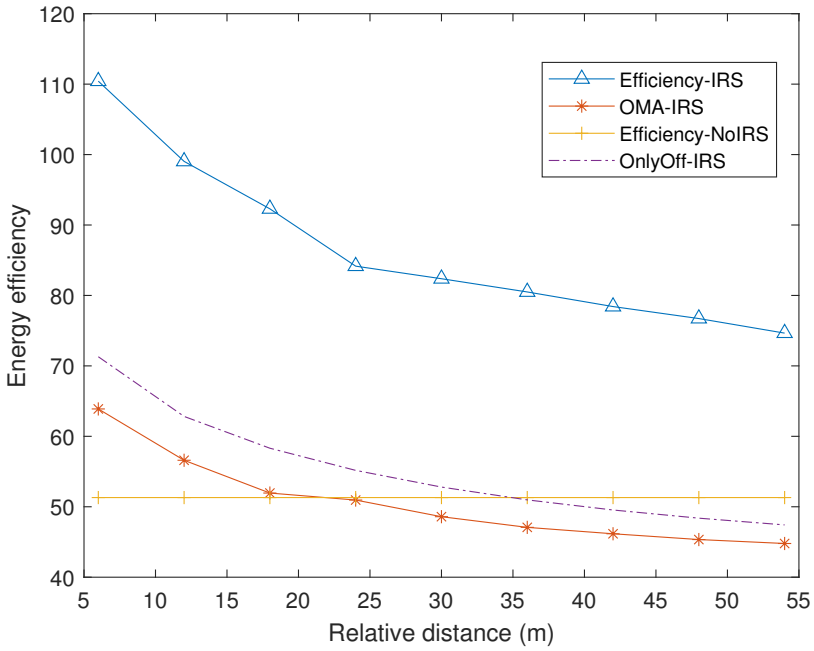

Fig. 6.5: Energy efficiency versus the relative distance of UE-IRS.

algorithm.

6.5 Chapter Conclusions

In this chapter, an IRS-assisted MEC network with NOMA was considered. EE was maximized by jointly optimizing the offloading power, local computing frequency, beamforming vectors, and IRS phase shift matrix. An alternating algorithm was proposed to solve the challenging non-convex fractional optimization problems. The numerical results showed that our proposed method outperforms other benchmark schemes in terms of EE. It was proved that NOMA and IRS could help the MEC network to achieve a higher rate with a lower power. The convergence of the proposed algorithm was also verified.

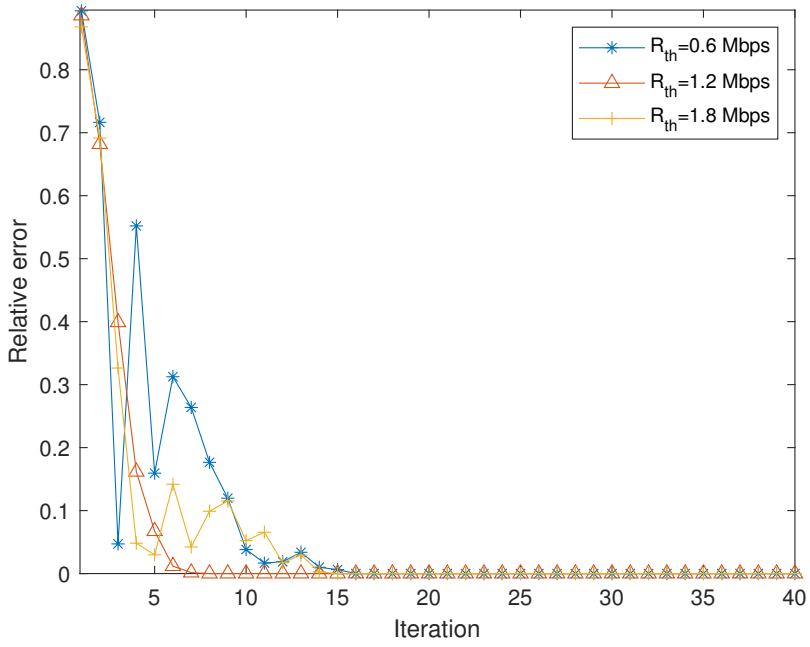

Fig. 6.6: Convergence with Iteration.

CHAPTER 7

Conclusions

7.1 Summary

In this thesis, we employed multiple techniques to improve the EE and PLS for IoT networks under the MEC framework. We started from a NOMA-based MEC network. By considering fairness between different users with respect to their overall rate, we developed several optimization algorithms to find the optimal trade-off between users' rate and the corresponding consumed power for three different fairness indicators.

We then expanded this model to a more complicated dynamic setting. Considering dynamic arrival traffic loads and environmental changes, we proposed a twin time scale resource allocation approach to improve the system EE. LSTM was adopted to predict the arrival tasks at the edge server-side, and a corresponding server coordination algorithm was developed to optimize the power consumption and guarantee the QoS for each user. A Lyapunov optimization theory-based algorithm was proposed in the short time slot to solve the resource allocation problem.

We then further investigated PLS together with EE for this dynamic model and verified that NOMA could help the system achieve a higher EE and PLS. Enlighted by the space diversity brought by multiple antennas, we begin to investigate a MISO network's performance with IRS. We then formed a secure EE maximization problem which included IRS, PLS, and FJ. Based on the channel state information, we developed two alternating approaches to optimize the beamform vector and phase shift matrix for the system. Finally, we applied IRS to our MEC networks with NOMA and verified that IRS could further improve the system EE.

7.2 Future Work

In our preliminary work, we have investigated the performance of the IRS-assisted MEC network for the EE. However, the improvement of the PLS of the MEC system by combining IRS with NOMA needs further research. Moreover, while considering multiple MEC servers and multiple IRS for IoT networks, the server association and coordination will introduce more challenges to the resource allocation optimization. Classical optimization theory may not be able to deal with such complexity. Fortunately, machine learning provides a powerful tool for researchers to dive deeply into those complex challenges. Therefore, in our future work, we will keep exploring new approaches that can further improve the system EE and PLS under more practical and complicated settings.

REFERENCES

[1] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, 2014.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[3] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA–MEC offloading," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 6207–6211, 2019.

[4] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Communications Letters*, vol. 23, no. 2, pp. 310–313, 2019.

[5] L. P. Qian, A. Feng, Y. Huang, Y. Wu, B. Ji, and Z. Shi, "Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IoT networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2806–2816, 2019.

[6] H. Sun, Q. Wang, S. Ahmed, and R. Q. Hu, "Non-orthogonal multiple access in a mmwave based IoT wireless system with SWIPT," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5.

[7] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Communications Letters*, vol. 23, no. 2, pp. 310–313, 2019.

[8] W. Wu, X. Wang, F. Zhou, K. Wong, C. Li, and B. Wang, "Resource allocation for enhancing offloading security in NOMA-enabled MEC networks," *IEEE Systems Journal*, pp. 1–4, 2020.

[9] X. Wang, W. Wu, B. Lyu, and H. Wang, "Delay minimization for secure NOMA mobile-edge computing," in *2019 IEEE 19th International Conference on Communication Technology (ICCT)*, 2019, pp. 1529–1534.

[10] H. Lin, Y. Cao, Y. Zhong, and P. Liu, "Secure computation efficiency maximization in NOMA-enabled mobile edge computing networks," *IEEE Access*, vol. 7, pp. 87 504–87 512, 2019.

[11] H. Sun, Q. Wang, X. Ma, Y. Xu, and R. Q. Hu, "Towards green mobile edge computing offloading systems with security enhancement," in *2020 Intermountain Engineering, Technology and Computing (IETC)*, 2020, pp. 1–6.

[12] D. Lymberopoulos, A. Bamis, and A. Savvides, "Extracting spatiotemporal human activity patterns in assisted living using a home sensor network," in *Proceedings of the 1st International Conference on PErvasive Technologies Related to Assistive Environments*, 2008, pp. 1–8.

[13] T. G. Rodrigues, K. Suto, H. Nishiyama, N. Kato, and K. Temma, "Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration," *IEEE Transactions on Computers*, vol. 67, no. 9, pp. 1287–1300, 2018.

[14] M. Hu, L. Zhuang, D. Wu, Y. Zhou, X. Chen, and L. Xiao, "Learning driven computation offloading for asymmetrically informed edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 30, no. 8, pp. 1802–1815, 2019.

[15] Y. Miao, G. Wu, M. Li, A. Ghoneim, M. Al-Rakhami, and M. S. Hossain, "Intelligent task prediction and computation offloading based on mobile-edge cloud computing," *Future Generation Computer Systems*, vol. 102, pp. 925–931, 2020.

[16] Q. Wang, Q. Xie, N. Yu, H. Huang, and X. Jia, "Dynamic server switching for energy efficient mobile edge networks," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[17] Z. Zhou, X. Chen, W. Wu, D. Wu, and J. Zhang, "Predictive online server provisioning for cost-efficient iot data streaming across collaborative edges," in *Proceedings of the Twentieth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, ser. Mobihoc '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 321–330. [Online]. Available: https://doi.org/10.1145/3323679.3326530

[18] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[19] C. Huang, G. C. Alexandropoulos, C. Yuen, and M. Debbah, "Indoor signal focusing with deep learning designed reconfigurable intelligent surfaces," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[20] Q. Wu and R. Zhang, "Beamforming optimization for wireless network aided by intelligent reflecting surface with discrete phase shifts," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1838–1851, 2020.

[21] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5394–5409, 2019.

[22] X. Yu, D. Xu, and R. Schober, "Enabling secure wireless communications via intelligent reflecting surfaces," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[23] D. Xu, X. Yu, Y. Sun, D. W. K. Ng, and R. Schober, "Resource allocation for secure IRS-assisted multiuser MISO systems," in *2019 IEEE Globecom Workshops (GC Wkshps)*, 2019, pp. 1–6.

[24] J. Chen, Y. C. Liang, Y. Pei, and H. Guo, "Intelligent reflecting surface: A programmable wireless environment for physical layer security," *IEEE Access*, vol. 7, pp. 82 599–82 612, 2019.

[25] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Communications Letters*, vol. 23, no. 9, pp. 1488–1492, 2019.

[26] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Communications Letters*, vol. 9, no. 1, pp. 108–112, 2020.

[27] B. Feng, Y. Wu, and M. Zheng, "Secure transmission strategy for intelligent reflecting surface enhanced wireless system," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.

[28] W. Shi, X. Zhou, L. Jia, Y. Wu, F. Shu, and J. Wang, "Enhanced secure wireless information and power transfer via intelligent reflecting surface," *IEEE Communications Letters*, pp. 1–1, 2020.

[29] H. Sun, F. Zhou, and R. Q. Hu, "Joint offloading and computation energy efficiency maximization in a mobile edge computing system," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3052–3056, 2019.

[30] M. Li, S. Yang, Z. Zhang, J. Ren, and G. Yu, "Joint subcarrier and power allocation for ofdma based mobile edge computing system," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017, pp. 1–6.

[31] S.-H. Kim, S. Park, M. Chen, and C.-H. Youn, "An optimal pricing scheme for the energy-efficient mobile edge computation offloading with ofdma," *IEEE Communications Letters*, vol. 22, no. 9, pp. 1922–1925, 2018.

[32] S. Wang, C. Pan, and C. Yin, "Joint heterogeneous tasks offloading and resource allocation in mobile edge computing systems," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2018, pp. 1–6.

[33] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1397–1411, 2017.

[34] Y. Wang, Y. Wu, F. Zhou, Z. Chu, Y. Wu, and F. Yuan, "Multi-objective resource allocation in a noma cognitive radio network with a practical non-linear energy harvesting model," *IEEE Access*, vol. 6, pp. 12 973–12 982, 2018.

[35] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna noma," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–7.

[36] Q.-V. Pham and W.-J. Hwang, "Fairness-aware spectral and energy efficiency in spectrum-sharing wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 207–10 219, 2017.

[37] F. Zhou, Y. Wu, R. Q. Hu, and Y. Qian, "Computation rate maximization in uav-enabled wireless-powered mobile-edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 1927–1941, 2018.

[38] Z. Zhang, H. Sun, and R. Q. Hu, "Downlink and uplink non-orthogonal multiple access in a dense wireless network," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2771–2784, 2017.

[39] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 90–96, 2014.

[40] Y. Wang, M. Sheng, X. Wang, L. Wang, and J. Li, "Mobile-edge computing: Partial computation offloading using dynamic voltage scaling," *IEEE Transactions on Communications*, vol. 64, no. 10, pp. 4268–4282, 2016.

[41] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking*, vol. 8, no. 5, pp. 556–567, 2000.

[42] Z.-Q. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 57–73, 2008.

[43] J. Papandriopoulos and J. S. Evans, "Scale: A low-complexity distributed protocol for spectrum balancing in multiuser dsl networks," *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3711–3724, 2009.

[44] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 5994–6009, 2017.

[45] X. Lyu, W. Ni, H. Tian, R. P. Liu, X. Wang, G. B. Giannakis, and A. Paulraj, "Optimal schedule of mobile edge computing for internet of things using partial information," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 11, pp. 2606–2615, 2017.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] Y. Mao, J. Zhang, and K. B. Letaief, "A lyapunov optimization approach for green cellular networks with hybrid energy supplies," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 12, pp. 2463–2477, 2015.

[48] S.-J. Huang and K.-R. Shih, "Short-term load forecasting via arma model identification including non-gaussian process considerations," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 673–679, 2003.

[49] L. T. Tan and L. B. Le, "Compressed sensing based data processing and mac protocol design for smartgrids," in *2015 IEEE Wireless Communications and Networking Conference (WCNC)*, 2015, pp. 2138–2143.

[50] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 190–10 203, 2018.

[51] L. T. Tan, R. Q. Hu, and L. Hanzo, "Twin-timescale artificial intelligence aided mobility-aware edge caching and computing in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 4, pp. 3086–3099, 2019.

[52] A. Zahin, L. T. Tan, and R. Q. Hu, "Sensor-based human activity recognition for smart healthcare: A semi-supervised machine learning," in *Artificial Intelligence for Communications and Networks*, S. Han, L. Ye, and W. Meng, Eds. Cham: Springer International Publishing, 2019, pp. 450–472.

[53] L. T. Tan, R. Q. Hu, and L. Hanzo, "Heterogeneous networks relying on full-duplex relays and mobility-aware probabilistic caching," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 5037–5052, 2019.

[54] A. Zahin, L. T. Tan, and R. Q. Hu, "A machine learning based framework for the smart healthcare system," in *2020 Intermountain Engineering, Technology and Computing (IETC)*, 2020, pp. 1–6.

[55] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[56] Y.-T. Lee, K.-T. Chen, Y.-M. Cheng, and C.-L. Lei, "World of warcraft avatar history dataset," in *Proceedings of the second annual ACM conference on Multimedia systems*, 2011, pp. 123–128.

[57] W. Xiao, W. Bao, X. Zhu, C. Wang, L. Chen, and L. T. Yang, "Dynamic request redirection and resource provisioning for cloud-based video services under heterogeneous environment," *IEEE Transactions on Parallel and Distributed Systems*, vol. 27, no. 7, pp. 1954–1967, 2016.

[58] S. Wang and S. Dey, "Adaptive mobile cloud computing to enable rich mobile multimedia applications," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 870–883, 2013.

[59] Y. Yu, J. Cao, and J. Zhu, "An lstm short-term solar irradiance forecasting under complicated weather conditions," *IEEE Access*, vol. 7, pp. 145 651–145 666, 2019.

[60] Q. Wang, L. T. Tan, R. Q. Hu, and G. Wu, "Hierarchical collaborative cloud and fog computing in IoT networks," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2018, pp. 1–7.

[61] K. Park, Y. Choi, W. J. Choi, H.-Y. Ryu, and H. Kim, "Lstm-based battery remaining useful life prediction with multi-channel charging profiles," *IEEE Access*, vol. 8, pp. 20 786–20 798, 2020.

[62] Q. Wang and F. Zhou, "Fair resource allocation in an mec-enabled ultra-dense iot network with noma," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[63] M. Zeng, N.-P. Nguyen, O. A. Dobre, and H. V. Poor, "Securing downlink massive mimo-noma networks with artificial noise," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 685–699, 2019.

[64] Z. Chu, M. Johnston, and S. Le Goff, "Robust beamforming techniques for miso secrecy communication with a cooperative jammer," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, 2015, pp. 1–5.

[65] N. Nouri, A. Entezari, J. Abouei, M. Jaseemuddin, and A. Anpalagan, "Dynamic power–latency tradeoff for mobile edge computation offloading in NOMA-based networks," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2763–2776, 2020.

[66] Q. Wang, L. T. Tan, R. Q. Hu, and Y. Qian, "Hierarchical energy-efficient mobile-edge computing in iot networks," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 626–11 639, 2020.

[67] S. Mao, S. Leng, S. Maharjan, and Y. Zhang, "Energy efficiency and delay tradeoff for wireless powered mobile-edge computing systems with multi-access schemes," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 1855–1867, 2020.

[68] Q. Li, M. Hong, H.-T. Wai, Y.-F. Liu, W.-K. Ma, and Z.-Q. Luo, "Transmit solutions for MIMO wiretap channels using alternating optimization," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1714–1727, 2013.

[69] W. Wu, F. Zhou, R. Q. Hu, and B. Wang, "Energy-efficient resource allocation for secure noma-enabled mobile edge computing networks," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 493–505, 2020.

[70] Y. Deng, Z. Chen, X. Yao, S. Hassan, and A. M. A. Ibrahim, "Parallel offloading in green and sustainable mobile edge computing for delay-constrained iot system," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 12, pp. 12 202–12 214, 2019.

[71] M. Pei, A. L. Swindlehurst, D. Ma, and J. Wei, "Adaptive limited feedback for MISO wiretap channels with cooperative jamming," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 993–1004, 2014.

[72] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Communications Magazine*, vol. 58, no. 1, pp. 106–112, 2020.

[73] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah, and C. Yuen, "Energy efficient multi-user miso communication using low resolution large intelligent surfaces," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–6.

[74] F. Zhou, Z. Li, J. Cheng, Q. Li, and J. Si, "Robust an-aided beamforming and power splitting design for secure miso cognitive radio with swipt," *IEEE Transactions on Wireless Communications*, vol. 16, no. 4, pp. 2450–2464, 2017.

[75] J. Yang, I.-M. Kim, and D. I. Kim, "Optimal cooperative jamming for multiuser broadcast channel with multiple eavesdroppers," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2840–2852, 2013.

[76] Q. Gao, Y. Huo, T. Jing, L. Ma, Y. Wen, and X. Xing, "An intermittent cooperative jamming strategy for securing energy-constrained networks," *IEEE Transactions on Communications*, vol. 67, no. 11, pp. 7715–7726, 2019.

[77] J. Huang and A. L. Swindlehurst, "Robust secure transmission in miso channels based on worst-case optimization," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1696–1707, 2011.

[78] H. Niu, B. Zhang, Y. Huang, D. Guo, Z. Chu, and Z. Zhu, "Robust secrecy beamforming and power-splitting design for multiuser miso downlink with swipt," *IEEE Systems Journal*, vol. 13, no. 2, pp. 1367–1375, 2018.

[79] G. Zhang, J. Xu, Q. Wu, M. Cui, X. Li, and F. Lin, "Wireless powered cooperative jamming for secure ofdm system," *IEEE transactions on vehicular technology*, vol. 67, no. 2, pp. 1331–1346, 2017.

[80] H. Zhang, H. Xing, J. Cheng, A. Nallanathan, and V. C. Leung, "Secure resource allocation for ofdma two-way relay wireless sensor networks without and with cooperative jamming," *IEEE Transactions on industrial informatics*, vol. 12, no. 5, pp. 1714–1725, 2015.

[81] M. Dehghan, D. L. Goeckel, M. Ghaderi, and Z. Ding, "Energy efficiency of cooperative jamming strategies in secure wireless networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 9, pp. 3025–3029, 2012.

[82] K.-H. Park, T. Wang, and M.-S. Alouini, "On the jamming power allocation for secure amplify-and-forward relaying via cooperative jamming," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 1741–1750, 2013.

[83] Y. Liu, J. Li, and A. P. Petropulu, "Destination assisted cooperative jamming for wireless physical-layer security," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 4, pp. 682–694, 2013.

[84] H. Ma, J. Cheng, X. Wang, and P. Ma, "Robust miso beamforming with cooperative jamming for secure transmission from perspectives of qos and secrecy rate," *IEEE Transactions on Communications*, vol. 66, no. 2, pp. 767–780, 2017.

[85] X. Liu, Y. Gao, L. Wang, N. Sha, and S. Wang, "Distributionally robust optimization for secure transmission with assisting jammer in miso downlink networks," *IEEE Communications Letters*, vol. 23, no. 2, pp. 338–341, 2018.

[86] B. Su, Q. Ni, and B. He, "Robust transmit designs for secrecy rate constrained miso noma system," in *2018 IEEE 29Th annual international symposium on personal, indoor and mobile radio communications (PIMRC)*. IEEE, 2018, pp. 1–5.

[87] Y. Feng, Z. Yang, W.-P. Zhu, Q. Li, and B. Lv, "Robust cooperative secure beamforming for simultaneous wireless information and power transfer in amplify-and-forward relay networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 3, pp. 2354–2366, 2016.

[88] Z. Chu, T. A. Le, H. X. Nguyen, M. Karamanoglu, Z. Zhu, A. Nallanathan, E. Ever, and A. Yazici, "Robust design for miso swipt system with artificial noise and cooperative jamming," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[89] R. Liu, H. Li, M. Li, and Q. Liu, "Symbol-level precoding design for IRS-assisted MU-MISO systems," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.

[90] G. Zhou, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Intelligent reflecting surface aided multigroup multicast MISO communication systems," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3236–3251, 2020.

[91] J. Lyu and R. Zhang, "Spatial throughput characterization for intelligent reflecting surface aided multiuser system," *IEEE Wireless Communications Letters*, vol. 9, no. 6, pp. 834–838, 2020.

[92] J. Lyu and R. Zhang, "Hybrid active/passive wireless network aided by intelligent reflecting surface: System modeling and performance analysis," *arXiv preprint arXiv:2004.13318*, 2020.

[93] X. Guan, Q. Wu, and R. Zhang, "Intelligent reflecting surface assisted secrecy communication: Is artificial noise helpful or not?" *IEEE Wireless Communications Letters*, vol. 9, no. 6, pp. 778–782, 2020.

[94] Y. Pei, Y.-C. Liang, L. Zhang, K. C. Teh, and K. H. Li, "Secure communication over miso cognitive radio channels," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1494–1502, 2010.

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

[96] X. Le and J. Wang, "A two-time-scale neurodynamic approach to constrained minimax optimization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 3, pp. 620–629, 2017.

[97] A. Zappone and E. Jorswieck, "Energy efficiency in wireless networks via fractional programming theory," *Foundations and Trends in Communications and Information Theory*, vol. 11, no. 3-4, pp. 185–396, 2015.

[98] A. Zappone, E. Björnson, L. Sanguinetti, and E. Jorswieck, "Globally optimal energy-efficient power control and receiver design in wireless networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 11, pp. 2844–2859, 2017.

[99] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Transactions on Signal Processing*, vol. 66, no. 10, pp. 2616–2630, 2018.

[100] J. Liu, K. Xiong, Y. Lu, D. W. K. Ng, Z. Zhong, and Z. Han, "Energy efficiency in secure irs-aided swipt," *IEEE Wireless Communications Letters*, vol. 9, no. 11, pp. 1884–1888, 2020.

[101] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Processing Magazine*, vol. 27, no. 3, pp. 20–34, 2010.

[102] Q. Wang and F. Zhou, "Fair resource allocation in an MEC-enabled ultra-dense IoT network with NOMA," in *2019 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2019, pp. 1–6.

[103] Y. Jiang, Y. Zou, J. Ouyang, and J. Zhu, "Secrecy energy efficiency optimization for artificial noise aided physical-layer security in ofdm-based cognitive radio networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 858–11 872, 2018.

[104] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for irs-assisted uplink noma," *IEEE Communications Letters*, vol. 25, no. 1, pp. 234–238, 2020.

[105] W. Ni, X. Liu, Y. Liu, H. Tian, and Y. Chen, "Resource allocation for multi-cell irs-aided noma networks," *IEEE Transactions on Wireless Communications*, 2021.

[106] F. Fang, Y. Xu, Q.-V. Pham, and Z. Ding, "Energy-efficient design of irs-noma networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 14 088–14 092, 2020.

[107] Z. Li, M. Chen, Z. Yang, J. Zhao, Y. Wang, J. Shi, and C. Huang, "Energy efficient reconfigurable intelligent surface enabled mobile edge computing networks with noma," *IEEE Transactions on Cognitive Communications and Networking*, 2021.

[108] F. Zhou, C. You, and R. Zhang, "Delay-optimal scheduling for irs-aided mobile edge computing," *IEEE wireless communications letters*, vol. 10, no. 4, pp. 740–744, 2020.

[109] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Joint power control and beamforming for uplink non-orthogonal multiple access in 5g millimeter-wave communications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 9, pp. 6177–6189, 2018.

[110] Q. Wang and F. Zhou and R. Q. Hu and Y. Qian, "Energy-efficient beamforming and cooperative jamming in IRS-assisted MISO networks," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–7.

CURRICULUM VITAE

Qun Wang

Qun Wang received the M.S. degree from Xidian University, Xi'an, China, in 2016. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Utah State University, Logan, UT, USA. His research interests include mobile edge computing, nonorthogonal multiple access, intelligent reflect surface, MIMO, and machine learning.

Published Journal Articles

- Q. Wang, F. Zhou, R.Q. Hu, and Y. Qian, "Energy efficient robust beamforming and cooperative jamming design for IRS-assisted MISO networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2592-2607, April 2021.
- Q. Wang, L. T. Tan, R. Q. Hu, Y. Qian, "Hierarchical energy efficient mobile edge computing in IoT networks," *IEEE Internet of Things J.*, vol. 7, no. 12, pp. 11626-11639, Dec. 2020.
- H. Hu, Q. Wang, R. Q. Hu, H. Zhu, "Mobility-aware offloading and resource allocation in an MEC-enabled IoT network with energy harvesting," *IEEE Internet of Things J.*, early access.
- Q. Wang, H. Sun, R. Q. Hu, and A Bhuyan, " When machine learning meets spectrum sharing security: Methodologies and challenges", *submitted to IEEE Open Access.*

Published Conference Papers

- Q. Wang, F. Zhou, H. Hu, and R.Q. Hu, "Energy-efficient design for IRS-assisted MEC networks with NOMA", *accepted by IEEE WCSP 2021.*

- X. MA, H. Sun, Q. Wang, "User scheduling for federated learning through over-the-air computation." *Proc. VTC Fall*, 2021.

- Q. Wang, H. Hu, H. Sun and R.Q. Hu, "Secure and energy-efficient offloading and resource allocation in a NOMA-based MEC network," *Proc. IEEE/ACM Symposium on Edge Computing (SEC)*, pp. 420-424, 2020.

- H. Hu, W. Song, Q. Wang, F. Zhou, R. Q. Hu"Mobility-aware offloading and resource allocation in MEC-enabled IoT networks", *Proc. 16th International Conference on Mobility, Sensing and Networking (MSN)*, pp. 554-560, 2020.

- H. Sun, Q. Wang, X. Ma, Y. Xu and R. Q. Hu, "Towards green mobile edge computing offloading systems with security enhancement," *2020 Intermountain Engineering, Technology and Computing (IETC)*, Orem, UT, USA, 2020, pp. 1-6.

- Q. Wang and F. Zhou, R. Q. Hu, and Q. Yi, "Energy-efficient beamforming and cooperative jamming in IRS-assisted MISO networks," *Proc. IEEE ICC* , Dublin, Ireland, 2020.

- Q. Wang and F. Zhou, "Fair resource allocation in an MEC-enabled ultra-dense IoT network with NOMA," *Proc. 2019 IEEE ICC Workshops*, Shanghai, China, 2019, pp. 1-6.

- Q. Wang, L. T. Tan, R. Q. Hu, and G. Wu, "Hierarchical collaborative cloud and fog computing in IoT networks," *Proc. 2018 WCSP*, Hangzhou, 2018, pp. 1-7.

- H. Sun, Q. Wang, S. Ahmed, and R. Q. Hu, "Non-orthogonal multiple access in a mmWave based IoT wireless system with SWIPT," *Proc. 2017 IEEE VTC Spring*, Sydney, NSW, 2017, pp. 1-5.

- H. Sun, Q. Wang, R. Q. Hu, and Y. Qian, "Outage probability study in a NOMA relay system," *Proc. 2017 IEEE WCNC*, San Francisco, CA, 2017, pp. 1-6.