Reinforcement Learning Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon

By
Atefeh Omidkar

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis in whole or part should be addressed to:

Head of the Department of Electrical and Computer Engineering
3B48 Engineering Building 57 Campus Drive
University of Saskatchewan
Saskatoon, SK S7N 5A9 Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

It is anticipated that mobile data traffic and the demand for higher data rates will increase dramatically as a result of the explosion of wireless devices, such as the Internet of Things (IoT) and machine-to-machine communication. There are numerous location-based peer-to-peer services available today that allow mobile users to communicate directly with one another, which can help offload traffic from congested cellular networks. In cellular networks, Device-to-Device (D2D) communication has been introduced to exploit direct links between devices instead of transmitting through a the Base Station (BS).

However, it is critical to note that D2D and IoT communications are hindered heavily by the high energy consumption of mobile devices and IoT devices. This is because their battery capacity is restricted. There may be a way for energy-constrained wireless devices to extend their lifespan by drawing upon reusable external sources of energy such as solar, wind, vibration, thermoelectric, and radio frequency (RF) energy in order to overcome the limited battery problem. Such approaches are commenly referred to as Energy Harvesting (EH) There is a promising approach to energy harvesting that is called Simultaneous Wireless Information and Power Transfer (SWIPT).

Due to the fact that wireless users are on the rise, it is imperative that resource allocation techniques be implemented in modern wireless networks. This will facilitate cooperation among users for limited resources, such as time and frequency bands. As well as ensuring that there is an adequate supply of energy for reliable and efficient communication, resource allocation also provides a roadmap for each individual user to follow in order to consume the right amount of energy. In D2D networks with time, frequency, and power constraints, significant computing power is generally required to achieve a joint resource management design. Thus the purpose of this study is to develop a resource allocation scheme that is based on spectrum sharing and enables low-cost computations for EH-assisted D2D and IoT communication.

Until now, there has been no study examining resource allocation design for EH-enabled IoT networks with SWIPT-enabled D2D schemes that utilize learning techniques and convex optimization. In most of the works, optimization and iterative approaches with a high level of computational complexity have been used which is not feasible in many IoT applications. In order to overcome these obstacles, a learning-based resource allocation mechanism based on the SWIPT scheme in IoT networks is proposed, where users are able to harvest energy from different sources. The system model consists of multiple IoT users, one BS, and multiple D2D pairs in EH-based IoT networks. As a means of developing an energy-efficient system, we consider the SWIPT scheme with D2D pairs employing the time switching method (TS) to capture energy from the environment, whereas IoT users employ the power splitting method (PS) to harvest energy from the BS. A mixed-integer nonlinear programming (MINLP) approach is presented for the solution of the Energy Efficiency (EE) problem by jointly optimizing subchannel allocation, power-splitting factor, power, and time together. As part of the optimization approach, the original EE optimization problem is decomposed into three subproblems, namely: (a) subchannel assignment and power splitting factor, (b) power allocation, and

(c) time allocation. In order to solve the subproblem assignment problem, which involves discrete variables, the Q-learning approach is employed. Due to the large size of the overall problem and the continuous nature of certain variables, it is impractical to optimize all variables by using the learning technique. Instead dealing for the continuous variable problems, namely power and time allocation, the original non-convex problem is first transformed into a convex one, then the Majorization-Minimization (MM) approach is applied as well as the Dinkelbach.

The performance of the proposed joint Q-learning and optimization algorithm has been evaluated in detail. In particular, the solution was compared with a linear EH model, as well as two heuristic algorithms, namely the constrained allocation algorithm and the random allocation algorithm, in order to determine its performance. The results indicate that the technique is superior to conventional approaches. For example, it can be seen that for the distance of $d = 10$ m, our proposed algorithm leads to EE improvement when compared to the method such as prematching algorithm, constrained allocation, and random allocation methods by about 5.26%, 110.52%, and 143.90%, respectively. Considering the simulation results, the proposed algorithm is superior to other methods in the literature. Using spectrum sharing and harvesting energy from D2D and IoT devices achieves impressive EE gains. This superior performance can be seen both in terms of the average and sum EEs, as well as when compared to other baseline schemes.

Acknowledgements

First and foremost I am extremely grateful to my supervisor, Prof. Ha Nguyen who passed away on Sep 4. My sincere appreciation goes out to him for his invaluable advice, continuous support, and patience during my MSc study. It was one of the most bitter experiences for me to lose Prof. Ha. He was an incredible and undoubtedly the most caring supervisor I met. He was always helpful to me, so losing him is one of the hardest things anyone has to deal with. I wish him all the happiness in the universe.

It is my pleasure to thank my new supervisors, who accepts my supervision, Prof. Berscheid, for his consistent support and guidance during the writing of my thesis.

I would like to give special thanks to my husband, Sajjad. I appreciate his support during my study and helping me in all steps of my life. He has been of great assistance to me throughout the entire thesis process, and I am grateful for his help.

Finally, I wish to acknowledge the constant love and support I receive from my mom. I have always been motivated by her to continue my MSc studies.

Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AWGN	Additive White Gaussian Noise
BPSK	Binary Phase-shift keying
BS	Base Station
CU	Cellular Users
D2D	Device-to-Device
DNN	Deep Neural Network
EE	Energy Efficiency
EH	Energy Harvesting
FL	Federated Learning
FDM	Frequency Division Multiplexing
GA	Genetic Algorithm
ID	Information Decoding
IoT	Internet of Things
MIMO	Multiple-Input Multiple-Output
MINLP	Mixed Integer Non-linear Problem
MM	Majorization-Minimization
NLP	Natural Language Processing
PS	Power Splitting
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
RA	Resource Allocation
RF	Radio Frequency
RL	Reinforcement Learning
SE	Spectral Efficiency
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SWIPT	Simultaneous Wireless Information and Power Transfer
TDM	Time Division Multiplexing
TS	Time Switching
UE	User Equipment
VM	Virtual Machines
WIT	Wireless Information Transfer

WPT	Wireless Power Transmission

1 Introduction

1.1 Motivation

As the infrastructure of the Internet of Things (IoT) has developed over the past few years, it has evolved from a theoretical idea into one that is now a major priority for many organizations and industries. Many industries are being revolutionized by IoT, including smart homes, smart cities, self-driving cars, agriculture, traffic management, health care, wearables, IoT-connected factories, and so on [1].

The implementation of IoT in many sectors has led to tremendous growth in the number of IoT devices in use currently. A recent study estimates that there will be more than 30 billion IoT devices in use, three times more than non-IoT devices [2]. As the number of IoT devices continues to rise exponentially, this results in an increase in traffic volume on the global network with different quality of service (QoS) requirements [3]. As an example, when it comes to IoT devices such as CCTV cameras, high bandwidth and a low latency are critical to maintaining video streams, while Voice over IP (VoIP) applications have low bandwidth requirements, but depend heavily on delay. Due to the fact that most IoT devices are connected to cellular networks, direct communication between mobile users is proposed as a means of establishing various location-based peer-to-peer services and thereby alleviating congestion on cellular networks as a result.

There is a concept called device-to-device (D2D) that is used in cellular networks to describe a direct line of communication between two mobile users without having to pass through either a base station (BS) or the core network first. Traditional cellular networks route all communication through the BS, even when users are in range of D2D communication. There is an increasing trend in the use of services today that require high data rates (such as video sharing, online gaming, and proximity-aware social networks), as well as services where users may be able to directly communicate with each other (i.e., D2D). In this respect, D2D communications are able to increase the spectral efficiency of a network significantly when they are used in scenarios such as these. Moreover, D2D communications could potentially enhance throughput, energy efficiency, delay, and fairness [3].

However, it has been demonstrated that the high energy consumption of mobile and IoT devices, whose battery capacity is limited, is the main obstacle to fully utilizing D2D and IoT communications [4]. There are many IoT devices that are designed with the intent of being left in place for a long period of time. This includes devices that track the climate or soil in harsh environments or sensors which monitor preventative maintenance in buildings, such as humidity sensors in buildings. The cost of repairing or replacing batteries across an IoT fleet that is dispersed and difficult to access can be quite high. It would be extremely valuable to extend the lifetime of energy-constrained wireless devices by using renewable energy sources, such as solar, wind, vibrations, thermoelectric, and radio frequency (RF). Therefore, a new paradigm of Energy Harvesting (EH) techniques have been developed in order to capture energy from the environment.

Additionally, resource allocation techniques are essential for managing the energy efficiency of constraint-battery devices in order to facilitate devices competing for limited resources. A resource allocation process is conducted as part of D2D communication after peer discovery and mode selection. This includes assigning frequency spectrum and allocating power levels for each user according to their specific needs. By determining how to distribute the available frequency sources of the network and the amount of power transmitted, it aims to improve D2D communication by maximizing network efficiency. As devices operating in D2D mode and devices communicating in the conventional cellular BS-centric mode (referred to here D2D and IoT devices, respectively) usually share the same bandwidth at the same time, this resource allocation needs to be optimized. There are four primary goals in optimizing D2D communications: 1) increasing energy efficiency, 2) increasing frequency spectrum usage, 3) maximizing total data rate, 4) and minimizing interference [5].

1.2 Research Objectives and Contributions

The aim of this research is to study the problem of EE optimization for D2D users and IoT devices that are supported by the Simultaneous Wireless Information and Power Transfer protocol (SWIPT). Aiming for simple and computationally inexpensive optimization solutions is the primary objective. In recent years, reinforcement learning (RL) has been successfully applied to optimization problems in many domains [6, 7] Since RL techniques are capable of providing a large amount of computing ability, an objective of this study is to devise a system for applying RL in allocating resources to EH-aided IoT networks using SWIPT-enabled D2D schemes in order to achieve EE provisioning through EH. In summary, this research aims to maximize EE by employing learning schemes in EH-aided D2D and IoT networks. The approach and main contributions are summarized as follows.

- The potential application of resource allocation techniques in IoT networks with a limited battery life in order to harvest energy for energy harvesting and energy storage is explored. In order to establish an energy-efficient system, the SWIPT scheme is considered. Since D2D pairs have a limited amount of power and exchange data at a high rate, they need to use a lot more auxiliary energy resources in order to exchange data with each other. There is no doubt that the amount of energy absorbed by the environment (solar, wind, etc.) is much higher than the amount of energy captured by BS. Due to this, D2D pairs use the TS approach in order to harvest energy from their surroundings. The application does not consider the data rate of IoT devices to be a very significant criterion since latency and reliability requirements are more important than data rate. Thus, IoT devices that utilize the PS approach can benefit from the harvesting of energy from the BS, and they can communicate at a lower rate.

- In order to evaluate the performance of the proposed scheme, the mixed-integer nonlinear programming (MINLP) EE problem is formulated by jointly optimizing the subchannel allocation, power splitting factor, power, and time.

- The proposed approach to this problem is to decompose the original EE optimization problem into three subproblems: (i) subchannel assignment and power splitting factor, (ii) power allocation, and (iii) time scheduling. In the first subproblem (which includes discrete variables), the Q-learning technique is applied. It is important to note that for the current size of the problem, optimizing all variables using the learning technique is not feasible. This is because the size of the problem has subsequently increased. Additionally, mini-batch has a limited size and is not applicable to a large number of variables due to its limited size. RL is a technique that can be used for both discrete variables and continuous variables, such as power, as well. Although the power allocation vectors can be discretized and solved by means of RL algorithms, it is possible for the solution to be inaccurate.

- The MM technique and Dinkelbach algorithm are used here to find the locally optimal solution using convex optimization methods in order to solve continuous variable problems including power and time allocation. While power can be discretized and solved with reinforcement learning algorithms, if we choose a poor state space or action space, we may introduce a hidden state into the problem, making it difficult to learn the optimal policy.

- As a result of the simulation results, the proposed RL-based resource allocation approach was able to demonstrate its efficiency for the proposed SWIPT scheme for multiple D2D pairs and IoT users. The results show that the proposed scheme achieves superior performance in terms of both average and sum EE when compared to [8] and other baseline schemes in the research literature.

1.3 Thesis Structure

This thesis is written in manuscript style, and is structured as follows:

- **Chapter 1: Introduction**: This chapter introduces the research's general overview and highlights the motivation of our research and its contribution to the field.

- **Chapter 2: Background and Literature Reviews**: This chapter initially provides the basic conceptual definitions of our system model. Then, an intuitive literature review of several studies regarding EH-assisted D2D communication, EE optimization, learning-based resource allocation analysis, and Simultaneous Wireless Information and Power Transfer (SWIPT) technique. In this chapter, we discuss the shortcomings in previous works, and summarize our design criteria in order to address these shortcomings.

- **Chapter 3: Reinforcement-Learning-Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks**:

  This chapter is the manuscript of our paper published in IEEE Internet of Things journal on Sep 1, 2022. This chapter provides a detailed description of Reinforcement-Learning-Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks. The constraints regarding a joint subchannel assignment, power allocation, and time allocation are described in order to maximize the EE under the spectrum sharing scenario while maintaining minimum data rates for D2D pairs and IoT devices. Then, the solution to the optimization problem is proposed and detailed performance analysis is provided.

- **Chapter 4: Discussion and Conclusion**: This chapter presents the conclusions drawn that show the key contributions of the proposed method and highlight the most significant advantages of our work over previous studies. Then, future possible research topics and improvements are suggested.

References

[1] Daniel Minoli and Benedict Occhiogrosso. “Internet of things applications for smart cities”. In: *Internet of things A to Z: technologies and applications* (2018), pp. 319–358.

[2] Md Mainuddin, Zhenhai Duan, and Yingfei Dong. “Network Traffic Characteristics of IoT Devices in Smart Homes”. In: *2021 International Conference on Computer Communications and Networks (ICCCN).* 2021, pp. 1–11.

[3] Arash Asadi, Qing Wang, and Vincenzo Mancuso. “A Survey on Device-to-Device Communication in Cellular Networks”. In: *IEEE Communications Surveys & Tutorials* 16.4 (2014), pp. 1801–1819.

[4] Shruti Gupta, Rong Zhang, and Lajos Hanzo. “Energy Harvesting Aided Device-to-Device Communication Underlaying the Cellular Downlink”. In: *IEEE Access* 5 (2017), pp. 7405–7413.

[5] Omid Yazdani and Ghasem Mirjalily. “A survey of distributed resource allocation for device-to-device communication in cellular networks”. In: *2017 Artificial Intelligence and Signal Processing Conference (AISP).* 2017, pp. 236–239.

[6] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. “Applications of Deep Reinforcement Learning in Communications and Networking: A Survey”. In: *IEEE Communications Surveys & Tutorials* 21.4 (2019), pp. 3133–3174.

[7] Meisam Maleki, Vesal Hakami, and Mehdi Dehghan. “A model-based reinforcement learning algorithm for routing in energy harvesting mobile ad-hoc networks”. In: *Wireless Personal Communications* 95.3 (2017), pp. 3119–3139.

[8] Haohang Yang, Yinghui Ye, Xiaoli Chu, and Mianxiong Dong. “Resource and Power Allocation in SWIPT-Enabled Device-to-Device Communications Based on a Nonlinear Energy Harvesting Model”. In: *IEEE Internet of Things Journal* 7.11 (2020), pp. 10813–10825.

2 Background

In ordero assist the readers in understanding our research problem and system model this chapter briefly introduces the relevant communication concepts and frameworks. Next, we explain resource allocation techniques for EE enhancement, practical SWIPT system for EH, and learning-based techniques to address the high computational complexity in EE problem optimization.

Section 2.1 contains the conceptual definitions and basic information regarding communication systems, resource allocation, Device-to-Device communication, energy-efficient communications, IoT networks and applications. Section 2.2 introduces resource allocation based on deterministic algorithms, and resource allocation based on heuristic algorithms. Afterwards, in Section 2.3, as SWIPT System for EH, some basic definitions including components of WPT, techniques for SWIPT, and resource allocation for systems with SWIPT are discussed. Following that, in Section 2.4, there is a brief introduction to reinforcement learning, and the Q-Learning algorithm. Finally, in the Section 2.5,literature review of key papers in this area is presented.

2.1 Conceptual Definitions

2.1.1 Communication Systems

During the communication process, information is exchanged between a sender and receiver. An example of a generic communication system can be seen in Fig. 2.1. A transmitter converts information into signal forms that are appropriate to be sent through the communication channel. As the signals propagate through the channel, toward the receiver, noise tends to corrupt the transmitted signal. The job of the receiver is to recover the message signal from the transmitted signal and noise.

As introduced above, the key components in a communication system are the transmitter, channel, and receiver. The following paragraphs provide additional details about each:

Transmitter: Unprocessed information in the form of digital bits cannot be directly transmitted over a communication channel. In order to convert the raw message bits into a format that can be transmitted through the communication channel, the data must be processed by the transmitter unit. The processing typically involves modulation, which superimposes a low frequency information signal on top of a high frequency carrier, and coding, which add redundancy to the message to increase reliability. Depending on the requirements, a variety of modulation and coding techniques can be used.

Communication channel: There is a communication channel between a transmitter and receiver

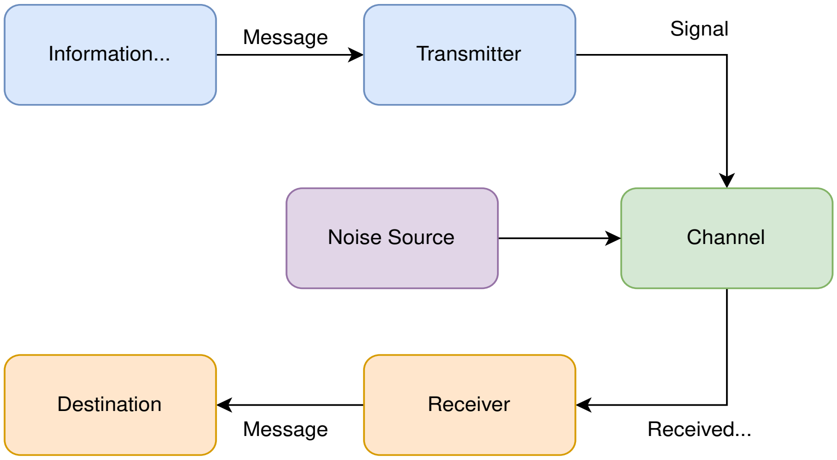


Figure 2.1: Communication systems process

through which signals are sent. The channel represents the physical medium through which the signal travels. For example, in radio communication systems, the channel is air, and in satellite communication systems, it is air and vacuum. Different types of channels have different effects on the transmitted signal. Most commonly, noise is added to the signal. Furthermore, as a result of channel attenuation with distance, the signal strength also decreases.

Receiver: It is necessary to decode or demodulate the signal as it passes through the communication channel so that the original message can be recovered. Using an amplifier, the receiver compensates for the loss of signal power caused by the attenuation of the channel. Obtaining an accurate copy of the message signal requires both a high level of selectivity as well as high levels of sensitivity on the receiver.

2.1.2 Elements of a Wireless Communication System

The communication systems of interest in this research are wireless, which means the communication channel medium is the air. Due the public and shared nature of the wireless channel, additional operations above those shown in Fig. 2.1 are generally required required in wireless networks to ensure secure and reliable communication. Fig. 2.2 shows the structure of a typical wireless communication system.

According to Fig. 2.2, transmission path of wireless communications systems typically includes encoders, encryptions, channel encoding, modulations, and multiplexings. During the signal processing process, the source encoder converts the information into a form suitable for signal processing. As part of this process, redundant information is removed to optimize resource utilization. The resulting data are then encrypted to prevent unauthorized access to the signal and the information. By applying channel encoding to a signal, the effect of impairments such as noise and interference on the received message can be reduced. The signal is made robust against noise by introducing some redundancy during this process. Modulation Techniques (such as quadrature amplitude modulation (QAM), phase-shift keying (PSK), and frequency-shift keying (FSK)) are used to modulate the signal so that it can be transmitted easily using an antenna. To share the wireless channel medium among multiple signals and users while avoiding interference, many multiplexing techniques

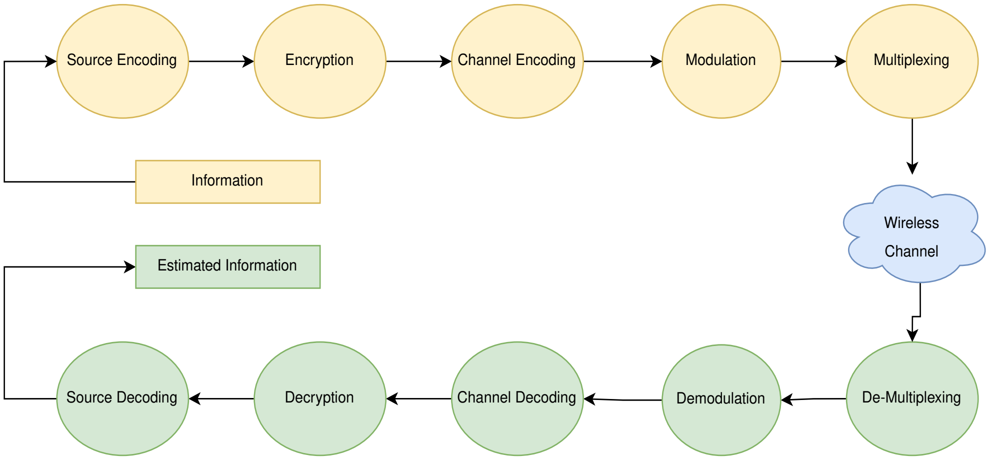


Figure 2.2: Basic elements of wireless communication systems

are used, such as Time Division Multiplexing (TDM) and Frequency Division Multiplexing (FDM). Simply put, in TDM different users transmit at different times, while in FDM different users transmit using different carrier frequencies.

In the reception path, the receiver collects a signal from the channel with the goal of reproducing the source signal. A wireless communication system's reception path is comprised of a number of different steps, which include demultiplexing, demodulation, channel decoding, decryption, and source decoding. As a result of the components of the reception path, the receiver has to do exactly the opposite of the transmitter in order to receive data. During the demultiplexing process, multiple signals are received from different channels and separated from each other. It is necessary to use appropriate demodulation techniques in order to demodulate each signal individually in order to recover the original message signal. In order to correct errors and remove redundant bits from the message, the Channel Decoder is used. Considering the message is encrypted, decryption of the signal will restore the message to its original unencrypted form. Finally, a Source Decoder processes this signal in order to retrieve the original message.

2.1.3 QAM Block Diagram

In order to analyze the performance of wireless networks, it is critical to understand the modulation process and the wireless models. Therefore, section 2.1.3 focuses on modulation and section 2.1.4 and 2.1.5 discuss wireless channels.

As previously mentioned, there are many different modulation techniques. One of the most common and representative techniques is QAM. So, we focus on QAM in this section.

To clarify how bits are converted to signals in wireless communication systems, we examine the 4-QAM transceiver as an example. This includes the transmitter, channel, and receiver. It is first necessary to feed the data stream from the transmitter into an M-QAM mapper in order to covert it to the series of discrete voltages levels. This M-QAM mapping process involves splitting the data bit stream into in-phase (I) and quadrature (Q) bit streams, which are encoded separately and then mapped to complex symbols. As an example, Fig. 2.3 shows the overall diagram of 4-QAM systems.

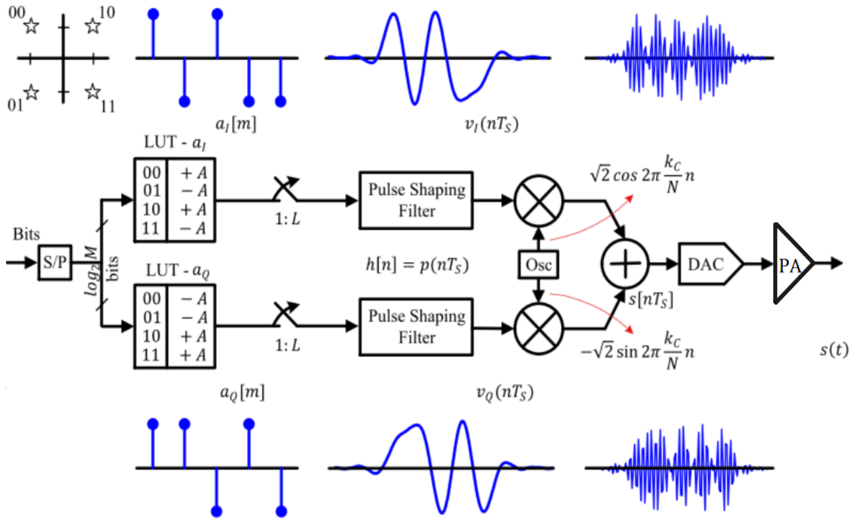


Figure 2.3: Block diagram for 4-QAM systems. [1]

The input receives a new bit every Tb seconds, forming a serial bit stream. Then, a serial-to-parallel (S/P) converter is applied to collect $\log_2 M$ bits every $T_M = \log_2 M \times T_b$ seconds in order to use them as an address for accessing two Look-Up Tables (LUT). A LUT stores 4 symbol values a_I while another stores 4 symbol values a_Q specified by the mapping rule.

In order to produce a QAM waveform, the symbol sequences $a_I[m]$ and $a_Q[m]$ need to be converted into discrete-time impulse trains in two separate arms (one I and the other Q) by upsampling by the amount of sample/symbol defined as a ratio between symbol time and sample time T_M/T_S. After upsampling, zeros are inserted between each symbol. The interpolated intermediate samples are then processed by a pulse shaping filter in each arm of the transmitter. Aside from shaping the spectrum, the pulse shaping filter also suppresses all the replicas of the original spectral data that arise from upsampling with the exception of the original data. The result is two independent I and Q pulse amplitude modulation (PAM) waveforms which are demonstrated as v_I and v_Q, respectively. Then, with carrier frequency F_C, the I PAM waveform v_I is upconverted by mixing with the carrier signal $\cos(2\pi F_C nT_S)$ while the Q PAM waveform v_Q is mixed with -$\sin(2\pi F_C nT_S)$. These mixed signals will be added together to form the QAM signal $s(nT_S)$. There is an oscillator at the Tx that is used to generate the carriers. Consequently, a general QAM waveform can be written as

$$s(nT_S) = v_I(nT_S)\sqrt{2}cos(2\pi\frac{F_C}{F_S}n) - v_Q(nT_S)\sqrt{2}sin(2\pi\frac{F_C}{F_S}n). \tag{2.1}$$

Through the use of a digital to analog conversion (DAC), it is possible to convert the discrete-time signal $s(nT_S)$ into a continuous-time signal $s(t)$, which can be expressed as

$$s(t) = v_I(t)\sqrt{2}cos(2\pi F_C t) - v_Q(t)\sqrt{2}sin(2\pi F_C t). \tag{2.2}$$

Optionally, the signal $s(t)$ may be amplified to increase its strength prior to transmission.

If an amplifier is used, we should place it exactly after the DAC block in the transmitter. Since energy efficiency in wireless networks is a key aspect of this research, it is important to understand the effects of the amplifier. In order to rate the efficiency of a power amplifier, power-added efficiency (PAE) or power amplifier factor, is a metric which considers the effects of the gain on the efficiency of the amplifier. This can be calculated (in percent) as follows:

$$\phi = \frac{p_{out}^{RF} - p_{in}^{RF}}{p_{DC}}, \tag{2.3}$$

where $0 < \phi < 1$, and p_{out}^{RF}, p_{in}^{RF} and p_{DC} represent the output power, input RF power, and DC input power, respectively. As a result of its DC source, an amplifier has the ability to amplify signals. In order to determine how much power is added to an input signal, the power-added efficiency can be used as a convenient parameter that will tell how much of the DC input power has been added to the signal.

Received signal $r(t)$ is a distorted version of $s(t)$, based on the channel characteristics. In this section, we simply assume $s(t)$ is corrupted by additive white Gaussian noise (AWGN) $w(t)$. AWGN is discussed in detail in section 2.1.4. As illustrated in Fig. 2.4, the symbols are detected through a series of steps.

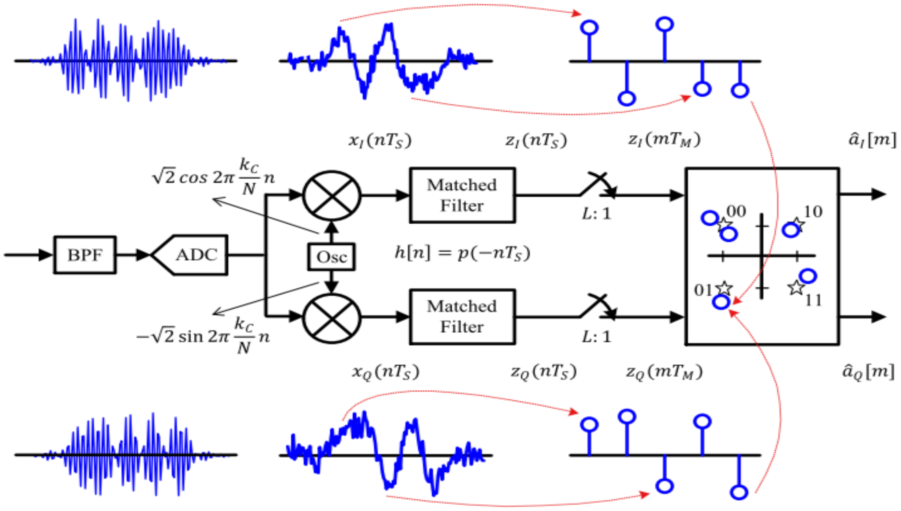


Figure 2.4: Block diagram for receiver of 4-QAM detector.[1]

In order to select the desired signal from the overall frequency space and reject any other signals that are present, a Bandpass Filter (BPF) is used. A digital to analog converter (ADC) is used to sample the signal at a rate of F_S samples to produce a sequence of Ts-spaced samples, which is then processed by the receiver. The next step is to obtain a complex signal $x(nT_S)$ by downconverting the ADC output by mixing it with two carriers, $\cos(2\pi F_C nT_S)$ and $\sin(2\pi F_C nT_S)$ created through the use of an oscillator at the receiver.

Two matched filters are then used to process the resulting complex waveform $x(nT_S)$ in the I and Q arms. As a result, $z_I(nT_S)$ and $z_Q(nT_S)$ will be generated. The matched filters impulse response is a time-reversed version of the pulse shaping filter, so its output is a correlation between the received signal and the original pulse shape. A downsampling of I and Q outputs is performed by L at the optimal sampling instants $n = mL = m\frac{T_M}{T_S}$ to recover T_M-spaced samples $z_I(mT_M)$ and $z_Q(mT_M)$. Ideally, the downsampler outputs should match the values which were originally read from the transmitter LUTs, allowing the receiver to recover the original bit streams through an inverse mapping process.

2.1.4 Effect of Additive White Gaussian Noise

Additive white Gaussian noise (AWGN) is one of the most common models of noise because it mimics natural random processes in nature including the random agitation of electrons within a conductor. For a better understanding of this noise model, the following discuss its key properties.

- AWGN is additive noise, which means that the received signal is the sum of the transmitted signal plus the random noise signal as shown in Fig. 2.5.

$$r(t) = s(t) + w(t). \tag{2.4}$$

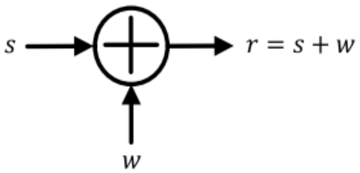


Figure 2.5: Additive white Gaussian noise

Additionally, it should be noted that this noise is statistically independent of the signal itself.

- There is a similarity between white noise and the white colour in the sense that white light is composed frequencies (colors) in the visible spectrum, just as white noise appears to have a uniform frequency across the whole range of frequencies. As a result, for all frequencies between $-\infty$ to ∞, the Power Spectral Density (PSD) $S_w(f)$ of white noise is constant

$$S_w(f) = \frac{N_0}{2}. \tag{2.5}$$

- AWGN has a normal (Gaussian) distribution over time and has zero as the average value over time. Noise has a random nature, which can cause distortion of signals and reduce the reliability of communication systems. In other words, the value of the noise at any particular instant in time is random, but its probability distribution over a long period of time is assumed to be known.

2.1.5 Wireless Channels

An example of a noisy constellation of 4-QAM modulation is shown in Fig. 2.6 with transmitted power 10 dB, and noise power 4 dB. The downsampled outputs from the matched filters are distorted by AWGN noise and mapped back to the constellation which was previously shown in the QAM detector block diagram. As can be seen, this plot of received constellation points is two dimensional. It can be interpreted as a plot of

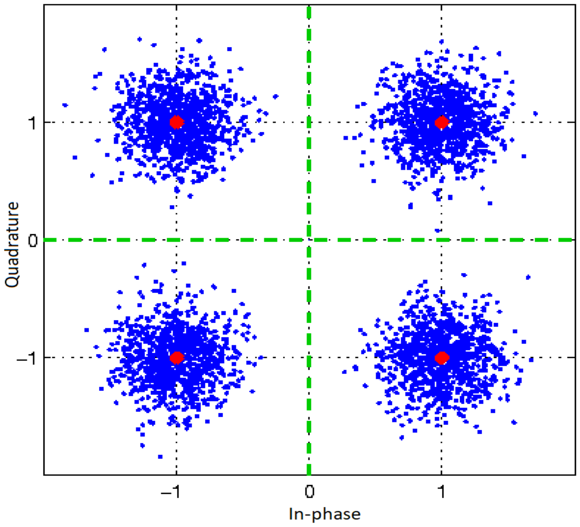


Figure 2.6: Noisy 4-QAM constellation.

Q versus I for each received value. Notice that the received points (blue) are distributed in "cloud" around the ideal points (red). The diameter of these clouds is determined by the noise power that is corrupting the signal. In the ideal case where there is no noise, this diameter will be zero, and all the optimally timed samples will coincide with the red points in the constellation.

Recall that, there is a medium in wireless networks between the transmitting and receiving antennas called "channel". As the signal travels from the transmitter to the receiver in a wireless transmission, its characteristics change in response to the environment. As a result of several factors, the signal characteristics can be determined: 1) Line of Sight (LOS) between the antennas, 2) reflection, refraction and diffraction caused by objects between the antennas (illustrated in Fig. 2.8), 3) relative motion between the transmitter and receiver, 4) attenuation of the signal while it travels through the medium, and 5) noise. As long as the channel between the antennas can be accurately modeled, the received signal can be derived from the transmitter signal. However, in practice, it is quite difficult to design wireless channel models that can accurately reflect the real world environments.

Let's take an example of a transmitter (base station) and receiver (mobile) in a city environment. Several buildings, trees, and other objects obstruct the medium between the transmitter and receiver. During transmission, the receiver may continue to move away from the transmitter. As the distance between devices varies, the signal power at the receiver can be measured. As an example, power ratio of the received signal to the transmitted signal is plotted against distance as shown in the Fig. 2.7 on the left half. There are three key factors influencing the received signal, which are highlighted in the right half of the Fig. 2.7. These three components include 1) Propagation Path Loss 2) Shadowing or slow fading 3) Multipath fading.

Propagation Path Loss: Typically, path losses include propagation losses as a result of the natural expansion of the radio wave front in free space. In addition to absorption losses (also referred to as penetration losses), which occur when a signal passes through a medium that is not transparent to electromagnetic waves,

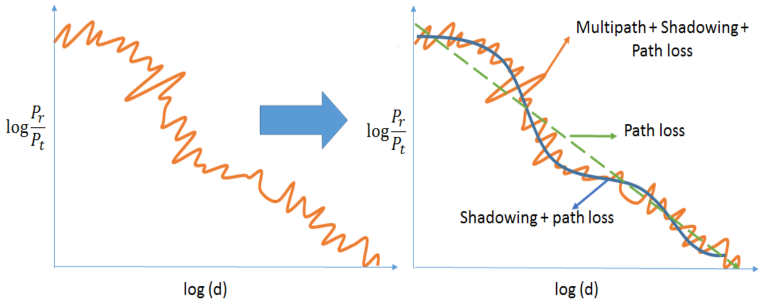


Figure 2.7: Components of wireless channels [2]

diffraction losses are also caused by opaque obstacles that obstruct part of the radiowave front. As can be seen in Fig. 2.8, an LOS path can be defined as a clear line between a transmitter and a receiver that is unobstructed and unimpeded. The free space propagation path loss through the obstacle-free, LOS path and free space (usually air) environment can be calculated as

$$P_r = P_t G_t G_r \left(\frac{\lambda}{4\pi d} \right)^2, \tag{2.6}$$

in which $P_t, P_r, G_t, G_r, d, \lambda$ are transmitted power, received power, transmitter gain, receiver gain, distance of transmitter to receiver, and wavelength in meters, respectively. In particular, note that the received power varies as $\frac{1}{d^2}$.

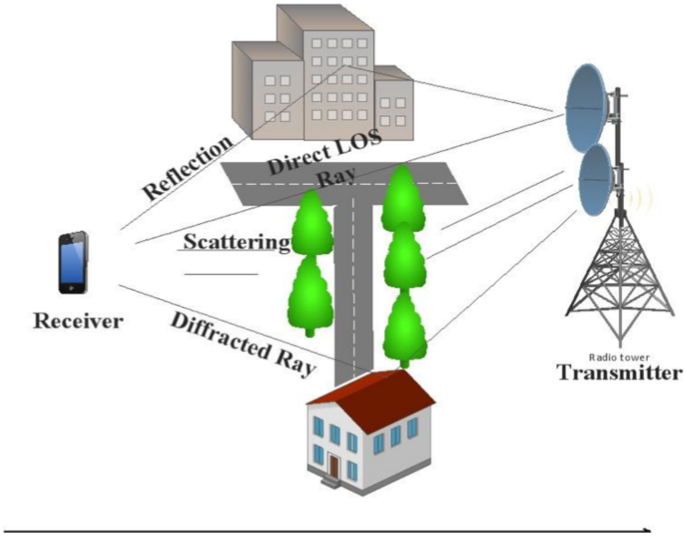


Figure 2.8: Propagation Path Loss [3]

Shadowing: Shadowing occurs when the transmitted signal is affected by absorption, reflection, diffraction and scattering as a result of objects such as buildings and trees. Shadowing is also known as slow fading or long-term fading.

Multipath fading: When a signal travels through an environment, it is likely to be reflected by a number of various objects along the way. As a result, several reflected signals are generated. Due to the fact that the reflected signals arrive at the receiver at various time instants and with differing intensities, multipath propagation of the signal can occur. The power of the received signal differs according to the phase of each individual reflected signal. It is possible for the various signal copies to add constructively or destructively, depending on their phases. There can be a significant difference in the total received power depending upon a small variation in the phase of each reflected signal from each multipath, changing the interference from constructive to destructive. It has also been referred to as fast fading or short-term fading.

2.1.6 Signal to Noise and Interference Ratio

The fundamental question that one must ask in the case of a wireless communication system is "What is the maximum performance that can be accomplished for a given channel?"As a means of measuring the performance of a communication link, we use the concept of channel capacity, C. Capacity can be simply defined as the maximum rate at which information can be transmitted over a channel reliably. A channel in the real world is essentially continuous both in time and in the space of the signal in both directions. The fundamental limitation of physical channels is that they have a limited bandwidth. Furthermore, practical transmitters can only output a limited amount of power. The following will look at the analysis of a real AWGN channel with a limited bandwidth and continuous in time. In the case of a continuous AWGN channel with a bandwidth B of Hz and a power average P of Watts, the capacity in $bits/s$ is defined as follow:

$$C = B \log_2 \left(1 + \frac{P}{N_0 B} \right), \tag{2.7}$$

where P is the average power spectrum of the transmitted signal, and $N_0/2$ is the power spectrum density of the noise.

The quantity $\Gamma = \frac{P}{N_0 B}$ is referred to the signal to noise ratio (SNR). As a result, the equation can be rewritten as follows:

$$C = B \log_2 (1 + \Gamma). \tag{2.8}$$

As explained above, SNR ratio is the ratio of the power of a desired signal (meaningful input) to the power of background noise (undesired input). In the same way, Signal-to-interference-plus-noise (SINR) ratio is a quantity that has been used in information theory to determine theoretical upper bounds on the capacity of a channel (or the rate of information transfer) in wireless communication systems.

The term interference is used to describe any event or process that disrupts or modifies a transmitted signal as it travels along a channel between a transmitter and a receiver in relation to wireless communication. When it comes to traditional wireless communication systems, interference is considered to be a limiting factor that prevents the achievement of substantial level of quality of experience (QoE) and limits the overall performance of the system. There is an amount of variance in the transmitted signal caused by interference, which degrades the final result. An important objective of traditional communication systems is to ensure that there is minimal or no interference. Therefore, in order to avoid, mitigate or cancel interferences, a great deal of effort needs to be made.

Let us consider a simple example of the instantaneous interference that may occur in a two-user transmission scheme, as shown in Fig. 2.9. In this example, we are using the symbols $u1$ and $u2$ to illustrate the symbols for user 1 and user 2 respectively. The Binary Phase Shift Keying (BPSK) modulation technique is used for the simplicity of the model, such that $u1 = 1$, $u2 = -1$. Further, it has been assumed that the noise at the receiver has been ignored, and $h_{i,k}$ represents the channel attention between transmitter i, and receiver k. Based on these basic parameters, it can be expressed that the received signal y is as follows:

$$y = u_1 h_{i,3} + u_2 h_{2,3} \tag{2.9}$$

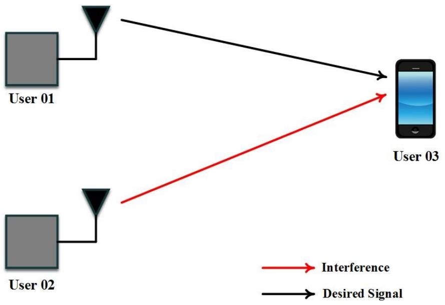


Figure 2.9: Interference between two users

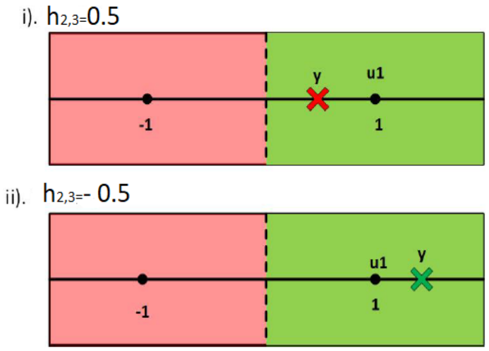


Figure 2.10: Destructive and constructive interference scenario in the Two-user transmission example

Fig.2.10 illustrates two distinct cases of interference, which both have the same transmit power of 1. In the first scenario, we have destructive interference at $h_{2,3} = 0.5$ and according to Fig.2.10, the received signal will be $y = 0.5$. In order to recover the data, the receiver will compare the received signal to a decision threshold of $y = 0$, the midpoint between the ideal constellation points. As a result of destructive interference from user 2, it appears that the received symbol of user 1 has shifted towards the decision threshold in BPSK constellations, as indicated by the dotted line. Consequently, the power received by user 1 has been reduced, and the performance of the system has been affected. In the presence of noise, the receiver will be much more likely to misidentify the transmitted point due to this interfernce. In the second scenario, the interference from the user 2 will be constructive if $h_{2,3} = -0.5$. Accordingly, the receiver sees $y = 1.5$. Due to this constructive interference from user 2 in this case, it has now led to the increasing of the received power of user 1. In this case the interference has decreased the likelihood of the receiver making an error. Symbol detection is tolerant of constructive interference rather than to the situation in which interference is not present. Clearly the SNR metric is not sufficient to analyze transmission reliability when interference is present. Instead of SNR parameter, SINR can be used. Basically, the SINR is calculated by dividing the power of a certain signal of interest by the sum of all the interference signals (from all the other signals interfering with the desired signal) and the power of some background noise that is around the desired signal. Here, received SINR, Γ, at receiver k is calculated as follows:

$$\Gamma_k = \frac{p_m \left|h_{m,k}\right|^2}{\sum_{i=1}^{i \neq m} p_i \left|h_{i,k}\right|^2 + N_0}, \tag{2.10}$$

where p_m and $\left|h_{m,k}\right|^2$ are the transmitter power and channel gain of user m, respectively. Note that the denominator 2.10 the summation of the interference from all other neighboring users i. The SINR will be converted to the signal-to-interference ratio (SIR) if the power of the noise term is zero. Conversely, if there are no interferences, the SINR is reduced to the SNR. In the example of Fig. 2.9, the SINR at receiver 3 will be calculated as follows

$$\Gamma_3 = \frac{p_1 \left|h_{1,3}\right|^2}{p_2 \left|h_{2,3}\right|^2 + N_0}, \tag{2.11}$$

where p_1 and $h_{1,3}$ are transmit power of user 1 and transmission channel gains between the user 1 and user 3 respectively, while p_2 and $h_{2,3}$ are defined as transmit power of user 2 and the interference channel spanning from the user 2 to user 3.

In IoT networks, the transmission from each user become interference to the other users, reducing the reliability of their transmissions. Therefore, assigning the transmission opportunities and power levels in the network is a complex multidimensional problem. It should be noted that in our system model, all transmission channels as well as interference channels are considered to be independent and identically Rayleigh distributed channels. Rayleigh fading models assume that the magnitude of a signal that has passed through a channel will vary randomly, or fade, according to a Rayleigh distribution, the radial component of the sum of two uncorrelated Gaussian random variables. In situations where there are many objects in the environment that scatter the radio signal before it reaches the receiver, Rayleigh fading is a commonly used model. The central limit theorem indicates that, if there is enough scatter, the impulse response of the channel will be well-modeled as a Gaussian process, regardless of how the components are distributed across the channel. This means that in the case of a scatter process where there are no dominant components, then the mean and phase of the process will be zero and equally distributed between 0 and 2π radians. As a result, the envelope of the channel response will have a Rayleigh distribution.

These characters provide a good rational reason to use the Rayleigh fading model in heavily congested city centers since a large number of buildings and other objects attenuate, reflect, refract, and diffract the signal. Additionally there is often no line of sight between the transmitter and receiver, and the signal is attenuated, reflected, refracted, and diffracted by many buildings and other objects.

2.1.7 Data Rates in Wireless Communication Channels

The term data rate refers to the speed at which data is transmitted in wireless communication. There can be many definitions of the term ”data rate”, but the simplest one is describing how many bits are transferred per second from one device to another over a network. In most cases, data rates are expressed in terms of bits per second or bytes per second. The data transmitted in wireless networks consists of not only the useful data bits, but also additional number of bits for the purpose of signaling, error correction, and addressing.

There are several factors which can affect the maximum data transmission rate of a wireless communication channel. Among these factors are the bandwidth of the channel, the number of discrete levels in the digital signal, and the amount of noise in the signal. As seen previously, the amount of noise or the quality of the signal will have an impact on how fast data can be reliably received. According to Nyquist's theorem for noiseless communication channels, and Shannon's theorem for noisy communication channels, there is a maximum data rate for communication channels in both noiseless and noisy situations.

Nyquist Theorem: It has been proven that according to the Nyquist theorem, the maximum data rate that can be obtained (bits per second) by a noiseless communication channel with bandwidth 'B' that transmits data with 'M' number of discrete levels is given

$$R_{max} = 2\ Blog_2\ M. \tag{2.12}$$

When a communication channel has a fixed bandwidth, then the only variable in the equation is the discrete level of the signal that appears in the channel. As the number of levels in a signal increases, the data rate will increase as well.

Shannon's Theorem: In practice, it is impossible to design a communication channel that is noiseless. Despite the fact that communication channels are always noisy, Shannon's theorem gives an answer to the question of how to achieve the highest data rate in noisy communication channels. Based on Shannon's theorem, the maximum possible data transmission rate in bits per second can be calculated by using the equation given below

$$R_{max} = B\ log_2(1 + \Gamma). \tag{2.13}$$

Note that Γ is SNR. The relationship between maximum data rate and Shannon's capacity is shown in the Fig. 2.11. Theoretically, it is possible to transmit R_{max} information at any rate $R_{max} \leq C$, with an arbitrary small error probability by using a sufficiently complicated coding scheme. While for an information rate $R_{max} > C$, it is not possible to find a code that can achieve an arbitrary small error probability. So, transmission at more than Shannon's rate will result in a high bit error rate.

2.1.8 Resource Allocation

When a wireless system is operated on the basis of a downlink, Base stations (BSs) provide data transmission between the pool of users, whose channels vary according to the type of wireless system and are typically time and frequency dependent. The BS would like to make maximum use of its bandwidth and power resources by

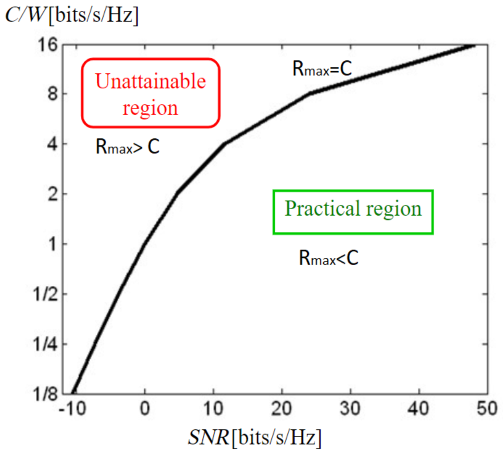


Figure 2.11: Shannon's capacity limit

pairing users up with strong subchannels and distributing power to them in the most efficient manner. The BS wants to make the most of its limited bandwidth and power resources. Moreover, the BS should impose a specific quality-of-service (QoS) constraint for each user, for example, maintaining a minimum rate per user. In order for the BS to achieve the desired efficiency-related quantities (e.g. throughput) under specific constraints (e.g., power limits), it must allocate resources in such a way that the efficiency-related quantities are maximized [4].

2.1.9 Device-to-Device communication

It has been four generations since the cellular network was invented. During this forward journey, the primary motivation has been the need to exchange multimedia-rich data quickly and make high-quality voice calls. In the era of new, more demanding applications and exponentially increasing subscriber bases, new techniques for boosting data rates and reducing latency are urgently needed. The concept of device-to-device communication has been adopted as an innovative paradigm in the cellular network field [5]. User equipment (UE) that are in close proximity and capable of direct communication can communicate directly instead of having radio signals travel through the core network or BS. As a result of its shorter signal traversal path, its main advantage is very low latency communication. There are a number of short-range wireless technologies that can be used for D2D communication, including Bluetooth [6], WiFi Direct [7], and LTE Direct [8] (defined by the Third Generation Partnership Project (3GPP)). There are a number of differences between them, including data rates, distance between one-hop devices, device discovery mechanisms, and typical applications they support. The maximum data rate of Bluetooth is 50 Mbps with a range of up to 240 meters, the maximum data rate of WiFi Direct is 250 Mbps and the range is up to 200 meters, while the maximum data rate of LTE Direct is 13.5Mbps with a range of 500 meters [9]. The D2D connection will allow operators to offload traffic from their core network to their edge network, increasing the efficiency of the spectrum and reducing the network's energy consumption. Here in Fig.2.12 is a diagram which illustrates the operation of cellular communication as well as D2D communication.

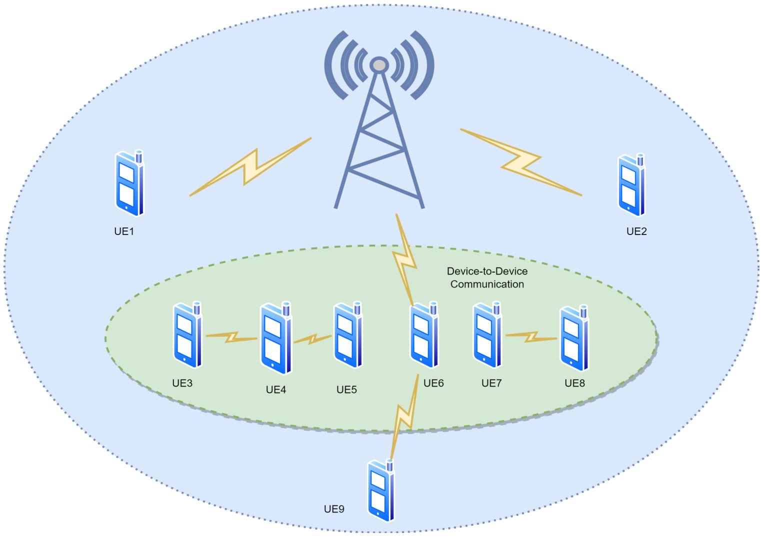


Figure 2.12: D2D communication versus cellular communication. This figure shows single-hop and multi-hop networks derived from D2D links.

2.1.10 Energy-Efficient Communications

The deployment widespread of next generation communication systems will result in a significant increase in the energy consumption of mobile communications, especially in developing countries (such as China and India). If dedicated actions are not taken, mobile communications will consume extremely large amounts of energy in terms of energy consumption, it is being estimated that the radio access part consumes more than 5% of the total energy, while the power amplifier uses to 50-80% [10–13]. Due to this, energy efficiency (EE) is not only important to tackling climate change from the perspective of operators, but it also has significant economic benefits as well. Therefore, when planning wireless networks, it is imperative to shift the focus from optimizing capacity and spectral efficiency to optimizing energy efficiency. The energy efficiency of a wireless communication system can be defined as the ratio between the maximum rate of data transfer that can be achieved and the total amount of energy consumed by the system:

$$\eta_{EE} = \frac{R^{\text{total}}}{P^{\text{total}}}, \tag{2.14}$$

where R^{total} and P^{total} are the total data rate and total power of the network, respectively.

2.1.11 IoT Networks and Applications

IoT refers to a network of physical objects known as "things", embedded with sensors, software, and other technologies (cloud computing, 5G, WiFi, and etc), forming the Internet of Things (IoT). Moreover, these devices are capable of communicating over the Internet with other devices and systems to exchange information [14]. This allows the objects to communicate with each other and exchange data with each other. Devices range in sophistication from simple household objects to complex industrial tools. Nowadays, embedded devices enable seamless communication between people, processes, and objects, such as cars, thermostats, and baby monitors, thanks to the Internet of Things. It is now possible, through low-cost computing, cloud computing, big data, analytics, and mobile technologies, to share and collect data from physical things with limited human involvement. Digital systems can be used in this hyperconnected world to record, monitor, and adjust the interactions between every thing that is connected.

Although the concept of IoT network has existed a long time, a number of technological advances in recent years have made IoT networks practical. Some of these technologies are listed below:

- **Low-cost, low-power technology**

  The availability of low-cost and reliable sensors is enabling more manufacturers to take advantage of IoT.

- **Connectivity**

  There are a wide variety of network protocols that have been developed in recent years that allow us to connect sensors to devices on the cloud and to other devices for efficient data transmission.

- **Cloud computing**

  Simply put, cloud computing is the delivery of computing services including servers, storage, databases, networking, software, analytics, and intelligence—over the Internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale. Cloud computing relies on sharing of resources to achieve coherence and typically uses a ”pay as you go” model, which can help in reducing capital expenses. In the future, cloud platforms will become more available to businesses and consumers, so they will be able to have access to the infrastructure they need to expand and scale without having to worry about managing it all themselves. This trend is encouraging the adoption of IoT systems.

- **Machine learning and data mining**

  Due to advances in machine learning and data mining, as well as the ability to access a large and diverse amount of data stored in the cloud, business owners are now able to gather insights faster and more easily than ever before.

- **Artificial intelligence (AI)**

  Due to progress in the field of neural networks, the use of natural-language processing (NLP) has become increasingly incorporated into IoT devices making these devices more appealing and useful. IoT powered by AI generates intelligent technologies that mimic intelligent behaviour and assist in decision-making with little or no human intervention.

It is expected that IoT applications will provide connectivity and intelligence to billions of everyday objects in the near future. There are already a variety of IoT applications in use, including smart homes, health care, wearables, smart cities, agriculture, industrial automation, vehicle-to-vehicle traffic management, and vehicle-to-person traffic management[14].

2.2 Resource Allocation in IoT Networks

Due to the rapid growth of IoT devices as well as the enormous amount of data generated by these devices, IoT resource management faces a number of major challenges [15]. Each IoT device can act as an edge node, cooperate and make its resources available dynamically to achieve a complex goal. Even though IoT devices are often constrained by limited resources, including exchangeable energy, storage capacity, and processing power, they can provide a wide range of services. IoT would perform better if these IoT resources were allocated efficiently. It is, however, not an easy task to allocate resources optimally in IoT due to its distributed and heterogeneous nature [16]. By deploying and integrating new devices, resources can be discovered automatically with minimum human involvement [17].

In order to share resources in the IoT, Kang *et al.* [18] proposed a rating system for IoT devices. Any object that can be allocated in an IoT network is considered a resource. A major challenge with IoT resource allocation is the limitation of QoS and the Service Level Agreement (SLA). The IoT system must provide an SLA, which identifies the level of QoS provided to the end user in addition to guaranteeing QoS [19]. Resources can be categorized into two main categories. The first category of resources consists of nodes/things, which include storage capacity, computational power, and energy capacity. The second category of IoT resources relates to communication channels or network resources, including the channel capacity, the load balancer, and the traffic analyzers that are used for communicating. Li *et al.* classified the resource allocation task as shown in Fig2.13 [20].

In order to solve resource allocation problems in IoT networks, there are two important classes of algorithms to be considered. First of all, we have deterministic algorithms. Second, there are heuristics and evolutionary algorithms. Given the dynamic nature of the resource allocation problems, heuristic and evolutionary algorithms have been gaining popularity in recent years. Throughout the following subsections, we will examine each of these types of algorithms separately.

2.2.1 Resource Allocation based on Deterministic Algorithms

Many resource allocation problems can be classified as NP-hard problems. Deterministic algorithms allocates resources based on a set of rules. It is important that these rules strike a balance between efficiency and effectiveness. While searching the whole solution space requires a lot of time, fewer searches generally lead to poor resource allocation. Simple implementation is one of the main advantages of deterministic algorithms. The resource management module in the middleware layer is proposed as a comprehensive approach to performing resource allocation. However, this approach uses a centralized architecture that is not appropriate for IoT's distributed structure. The relationship between resources is also assumed to be static and reliable, which may not be the case in IoT networks.

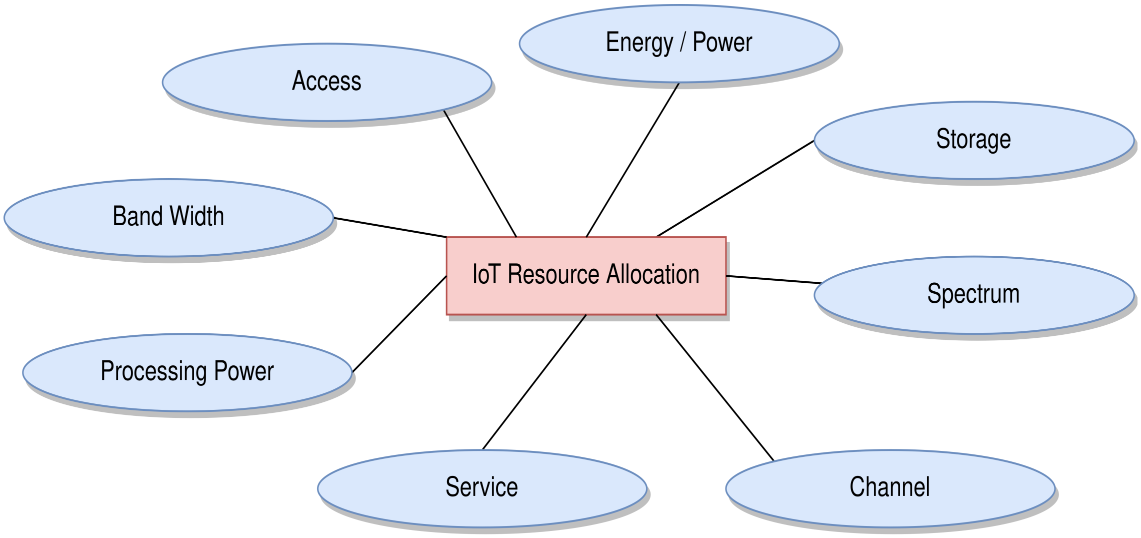


Figure 2.13: Resource allocation in IoT networks.

As a result, some previous studies have not attempted to reduce data sent over the network [21]. A rule-based approach has been proposed in other research, where rules are designed based on the service size, task completion time, and Virtual Machine (VM) capacity. According to the design rules, the VM's capacity is assigned to IoT services [21]. Another research project proposes two deterministic algorithms [22], each performing RA according to its specific rules. Using the first method, the average cost of each source is calculated and assigned to the service with the highest demand. As a second method, the resource that is most in demand is randomly assigned. Some papers have used game theory to allocate resources. Using game theory, Huang *et al.* proposed a tool-to-tool communication approach [23]. This method allocates resources based on a response function. The purpose of this function is to increase the overall profitability of the resource allocation method. IoT resource allocation can be managed using a game-based approach, according to Kim *et al.* [24].

As a result of this method, resources are allocated in such a way that the total performance of the system is increased. Computational complexity is the main problem with these methods. It would be challenging to solve these problems using a deterministic algorithm since most resource allocation problems are NP-complete and do not have a polynomial solution. The main limitation of most of the above techniques is the size of the problem. Therefore, in recent years, there have been a lot of studies focused on heuristic algorithms, such as genetic algorithms and particle swarm optimizations.

2.2.2 Resource Allocation based on Heuristic Algorithm

Unlike deterministic algorithms, heuristic algorithms conduct a partial search of the solution space instead of a full search. It is for this reason that heuristic algorithms have become more popular in recent years. Genetic algorithms (GA) are evolutionary algorithms that can be used to design heuristic methods for resource allocation. Resource allocation models are represented by each chromosome in this method. Scheduling quality is determined by the fitness of each chromosome. Another GA method used by Kim *et al.* [25] divides each chromosome into gateways and resources. As a result, generated solutions are also able to reduce the communication costs between gateways [25]. The authors in [26] illustrate their idea with resource allocation in the internet of things. A new resource allocation method was proposed for an IoT-driven healthcare system. Based on cost and latency criteria, two existing systems were benchmarked to evaluate the proposed system, called IoTR4Healthcare. To provide adaptive allocation and anti-jamming transmission, a novel automatic control allocation (ACA) model was developed [27]. Spread-time technology is also used to meet a node's power and bandwidth needs. To allocate resources in anti-jamming situations, a new joint stepwise recursive function is used. The key to the ACA model is that the optimal allocation solution is mapped into an orthogonal frequency division multiplexing (OFDM) waveform. Utilizing its assignment flexibility [15], OFDM mainly allocates subcarriers and power to users based on various fading characteristics of different users on the same subcarrier.

2.3 SWIPT System for EH

The integration of energy harvesting technologies into communication networks has attracted considerable interest in recent years. Several studies have examined optimal resource allocation techniques for different objective functions and topologies when using conventional renewable energy resources, such as solar and wind. For applications requiring a high level of QoS, energy harvesting is essential due to the intermittent and unpredictable nature of these energy sources. Most conventional harvesting technologies are restricted to specific environments. It is possible to overcome the above limitations through a technology called wireless power transfer (WPT), which uses electromagnetic radiation to charge the nodes' batteries.

The WPT technology allows the harvesting of green energy from ambient signals or from a dedicated source in a fully controlled way; in the latter case, the transfer of green energy can occur from more powerful nodes (e.g., base stations) utilizing conventional renewable energy sources. The initial focus of WPT has been on long-distance and high-power applications. In spite of this, both the low efficiency of the transmission process as well as health concerns associated with such high-power applications prevented their further development [28]. Thus, the majority of recent WPT research has been directed toward inductive coupling for transmitting near-field energy (e.g., for charging smart phones, medical equipment, and electrical vehicles). Furthermore, new developments in silicon technology have significantly reduced the energy requirements of simple wireless devices. There is currently a lot of interest in WPT among academia and industry. Several experimental results for different WPT scenarios have been published [28].

In the future, as sensors and wireless transceivers become smaller and more energy-efficient, radio waves will not only be the source of energy for operating these devices, but will also be used for transmitting information and energy. By superimposing information and power transfer, simultaneous wireless information and power transfer (SWIPT) can improve spectral efficiency, time delay, energy consumption, and interference management. For instance, wireless implants can be charged and calibrated simultaneously with the same signal, and wireless sensor nodes can be charged using the access point's control signals. SWIPT technologies have a vital role to play in the era of the Internet of Things. As we move into the era of the Internet of Things, SWIPT technologies can be of fundamental importance for the provision of energy supply to and the exchange of information with numerous ultra-low-power sensors, which support a wide range of heterogeneous sensing applications. In addition, future cellular systems that will employ small cells, massively multiple-input multiple-output (MIMO), and millimeter-wave technologies will be able to overcome the current path loss effect; in such a scenario, SWIPT could be integrated as a cost-effective way to jointly provide high throughput and energy efficiency [29].

A brief overview of SWIPT technology is provided in this subsection, along with a discussion of recent advances and future research challenges associated with the technology. In particular, this subsection will describe the rectifying antenna (rectenna) circuit, which is an essential part of the implementation of WPT/SWIPT technology because it converts microwave energy into direct current (DC) electricity. In order to achieve SWIPT, the received signal must be split into two orthogonal parts. An overview of recent SWIPT techniques has been provided which separate received signals based on their power, time, antenna, and spatial characteristics [29]. A SWIPT system, however, introduces fundamental changes to the operation of a communication system, and motivates new applications. Then, the impact of SWIPT on radio resource allocation problems will be examined. Also, we will investigate advanced cognitive radio scenarios that enable information and energy cooperation between primary and secondary networks.

2.3.1 Components of WPT

There are three main types of wireless electromagnetic power exchange [29]:

- Near-field power transfer: the transfer of power over short distances of up to one meter using either inductive, capacitive, or resonant coupling.

- Directional power beaming at far distances: requires directive antennas with a range of several milliwatts at distances of up to a few meters in both indoor and outdoor environments

- Low-power, far-field ambient RF power scavenging: receivers that use power transmitted from random transmitters (cell phone base stations, TV broadcast stations) to communicate with peer nodes.

If the power density is adequate, the range of communications can be several kilometers in this last case, assuming the collected power is several microWatts. Despite the fact that there are several applications of near field wireless charging, such as the charging of electric cars, cell phones, or other handheld devices, this study will focus primarily on far field wireless charging, which uses antennas that communicate far from the device.

It has been assumed in early information theory studies on SWIPT that the same signal will transmit both information and energy without loss, revealing a fundamental trade-off between power and information [30]. In practice, this simultaneous transfer is not possible due to the fact that the energy harvesting operation in the RF domain destroys the information content during the energy harvesting process. For SWIPT to work in practice, the received signal has to be divided into two separate parts, one that is used for harvesting the energy and the other that is used to decode the information. As can be seen from the generic model of the system depicted in Fig.2.14 , each user terminal is generally capable of both harvesting energy and decoding information simultaneously (by implementing the power splitting scheme discussed later) or they can choose harvesting mode or information transfer mode.

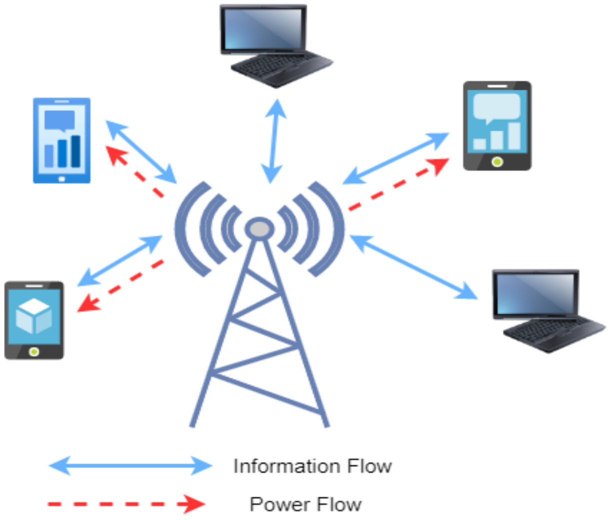


Figure 2.14: An energy and information-transfer wireless network

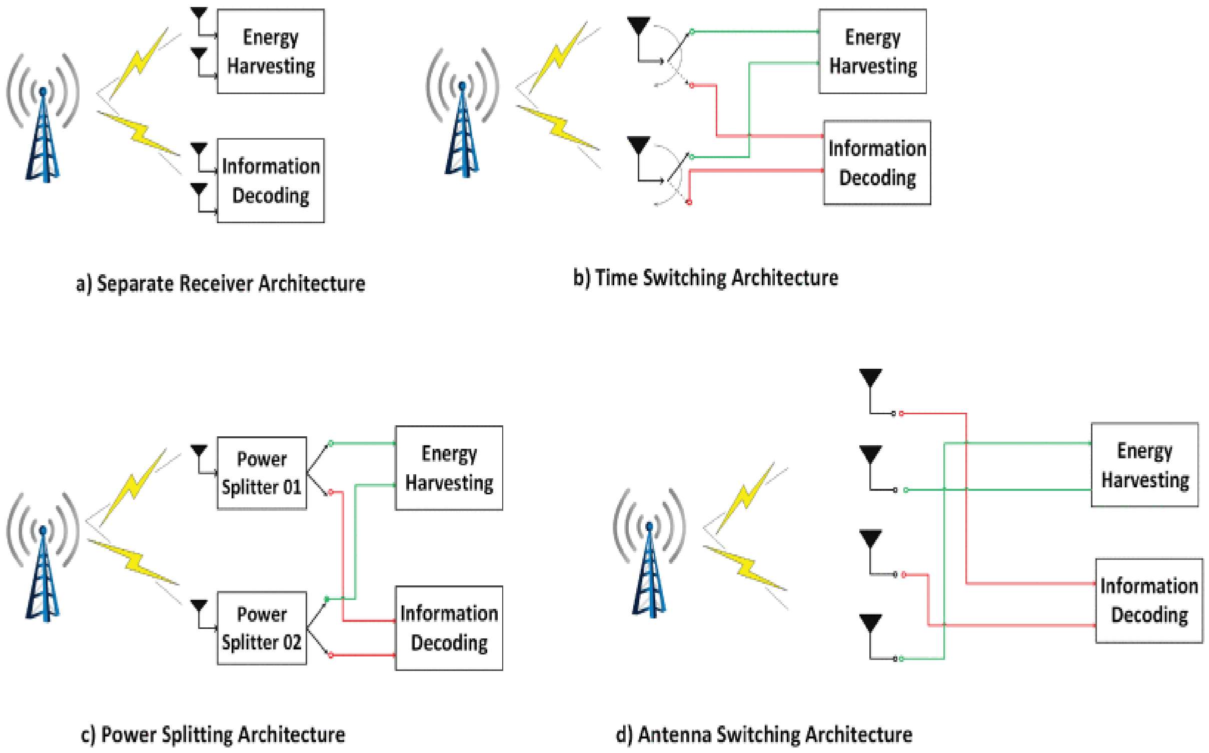
From the point of view of implementation, a design whereby each user is either an information receiver or an energy receiver at any given time might be desirable, which is known as a time switching design. A practical advantage of this scheme is that state-of-the-art wireless information and energy receivers usually operate at very different power sensitivity levels (e.g., -50dBm for information receivers vs. -10dBm for energy receivers).

2.3.2 Techniques for SWIPT

According to Fig. 2.15. a comparison of several SWIPT enabled receiver architectures is presented in the following section, including Separate Receiver, Time Switching (TS), Power Splitting (PS) and Antenna Switching (AS) architectures [31].

1. **Separate Receiver**

   In this separate receiver antenna architecture, the ID and EH circuits are handled by two separate receivers with separate antennas, which are served by a transmitter with multiple antennas. A further feature of these two antennas is that they observe different channels. It is easy to implement this



a) Separate Receiver Architecture

b) Time Switching Architecture

c) Power Splitting Architecture

d) Antenna Switching Architecture

Figure 2.15: Architectures for integrated SWIPT receivers. (a) Separated receiver architecture for energy decoding and information decoding. (b) Time-switching architecture: Each connected antenna will be allocated a specific time for harvesting energy and decoding information based on a predetermined time factor. (c) Power Splitting (PS) receiver: according to a PS ratio, it divides the received signal into two power streams. (d) Antenna switching receiver: A receiver that switches between harvesting energy and decoding information, based on an algorithm that optimizes antenna switching. [31]

receiver architecture using off-the-rack components for the EH and ID receivers. EH and ID can be performed concurrently and independently through this architecture. The trade-off between achievable EH and information rate can be optimized using channel state information (CSI) and receiver feedback.

2. **Time Switching**

In TS architectures, also known as co-located receiver architectures, the EH antenna is used for both information reception and EH. There are several components in the receiver used in this architecture, including an RF energy harvester, information decoder, as well as a switch that allows the user to switch between different types of receiving antennas. The antenna or antennas of the receiver are periodically switched between ID circuit and EH circuit dependant on the sequence of the TS. TS receivers require time synchronization, as well as accurate scheduling, of the information and energy. When receiver j operates in EH mode, it can harvest the following amount of energy from the source i [31]:

$$P_{i,j} = \eta P_i [h_{ij}]^2, \tag{2.15}$$

where η represents the amplifier's efficiency factor of the EH process, P_i represents the transmit power at source i, and h_{ij} represents the channel gain between source i and receiver j during transmission. It is possible to calculate the ID rate using the equation Eq 2.15 when the same receiver is set to ID mode, where B represents the transmission bandwidth and σ represents the noise power, respectively [31]:

$$R_{i,j} = B \log \left(1 + \frac{P_i h_{ij}^2}{\sigma^2} \right). \tag{2.16}$$

3. **Power Splitting**

   A PS receiver divides the received signal into two streams of different power levels with a certain PS ratio before signal processing can begin. It is then necessary to send both of the power streams into an information decoder and an energy harvester in order to make simultaneous Information Decoding (ID) and EH possible [31]. In other words, at the transmitter side, the energy signal is sent through high power unmodulated continuous wave centered at the carrier frequency, while the information signal is sent via low power subcarriers around the carrier frequency. As a result of such power allocation, harvestable power can be increased, while interference between external wireless networks is decreased. It is possible to optimize the PS ratio in each receiver antenna. Additionally, the PS ratio can be varied according to system requirements to balance information rate and harvesting energy. It is also possible to improve overall performance by optimizing the PS ratio in combination with the signal ratio. Let α_j be the power-splitting factor value for the receiver j. According to the formula below, we can calculate how much power is harvested at PS receiver j from source i:

$$P_{j,i} = \eta P_i |h_{ij}|^2 \alpha_j. \tag{2.17}$$

   When the power of the signal processing noise is expressed as σ_{sp}^2, the maximum ID rate as it is decoded from the source i can be written as:

$$R_{j,i} = B \log \left(1 + \frac{(1 - \alpha_i) P_i h_{ij}^2}{\sigma^2 + \sigma_{sp}^2} \right). \tag{2.18}$$

4. **Antenna Switching**

   It is possible to enable SWIPT by switching low complexity antennas between EH and ID. For example, there can be a subset of antennas at the receiver that work on ID, and the rest of the antennas can be used on EH. We can also adopt the dual antenna receiver architecture described in [32]. Furthermore, this architecture can be expanded easily to include many more antennas with a proper antenna switching protocol. PS architectures are sometimes viewed as separate cases of antenna switching. It is also possible to optimize a separate receiver architecture by using antenna switching architecture.

A brief overview of SWIPT technology and four different receiver architectures was provided in this section. As discussed, communication networks can benefit from SWIPT's ability to simultaneously transfer information and power. Different receiver architectures are needed to facilitate SWIPT because ID and EH operations have different sensitivity levels. Two distinct parts of the received signal must be separated, one for ID and one for EH. To accomplish this signal splitting, we provided the techniques that have been proposed for time, power, and antenna domains.

2.3.3 Resource Allocation for Systems with SWIPT

For SWIPT systems, the following aspects must be taken into consideration when designing the resource allocation algorithm:

1. **User Scheduling and Power Control**

   RF signals serve as both information carriers and power carriers at the same time. Although SWIPT can be realized in theory, a wide dynamic range of power sensitivity for EH (-10 dBm) and information decoding (-50 dBm) pose obstacles [29]. Thus, to facilitate SWIPT in practice, we need joint power control and user scheduling. In order to extend the lifetime of a communication network, idle users who have high channel gains can be scheduled for power transfer. Also, opportunistic power control can improve energy and information transfer efficiency by exploiting channel fading.

2. **Information and Energy Scheduling**

   It is only possible for passive receivers, such as small sensor nodes, to transmit uplink data once they have harvested enough energy from the downlink. Using a harvest-then-transmit design is motivated by the physical constraints on energy usage. When more time is allocated to energy harvesting in the downlink, more energy is harvested, which can be used in the uplink. Nevertheless, this may result in lower data transmission rates, since there is less time for uplink transmission. Thus, the system throughput can be optimized by varying the amount of time allocated for energy harvesting and information transmission.

3. **Interference Management**

   The co-channel interference that limits the performance of traditional communication networks can be suppressed or avoided through resource allocation. It should be noted, however, that in SWIPT systems, the receivers will embrace strong interference as it can provide the receivers with energy. Furthermore, by concentrating and aggregating multicell interference in a certain area, a "wireless charging zone" may be created by using interference alignment and/or interference coordination.

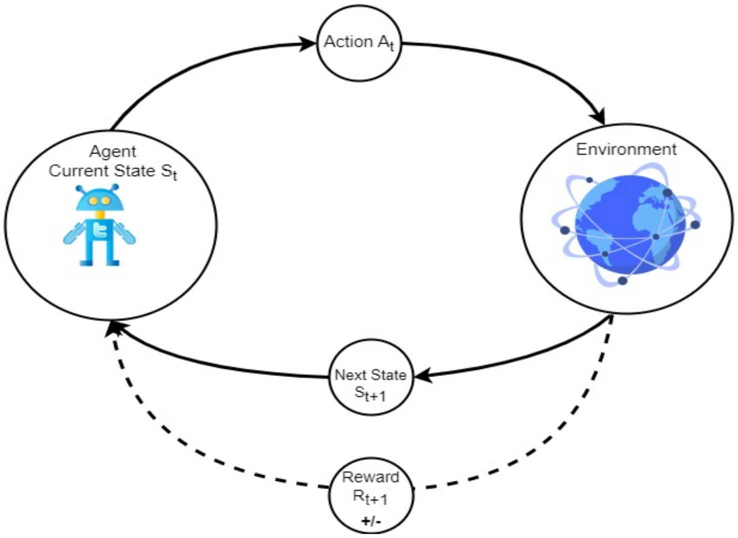


Figure 2.16: Reinforcement learning model

2.4 Reinforcement learning

2.4.1 Introduction to Reinforcement Learning

In machine-learning problems, a major objective is to produce intelligent programs or intelligent agents that adapt to changing environments through learning and adaption. One class of algorithms which can help to achieve this goal is Reinforcement Learning (RL) [33]. This method of learning involves direct interaction with the environment by learners or software agents. Even if there is no complete model or information about the environment, the agent can still learn through interaction and experience. A positive reward or negative reward is given to an agent in response to its actions. An environment's responses to the user's actions are mapped to each situation during the learning process, as shown in Fig 2.16. Algorithms based on reinforcement learning aim to maximize rewards during interactions with the environment and establish a mapping between states and actions to determine what actions to take. Policies can be set once or can adapt to a changing environment.

Reinforcement learning differs from supervised learning, which is perhaps the most commonly used method of learning. A supervised learning environment provides examples provided by a knowledgeable external supervisor. The examples are used to train a parameterized function approximator. In supervised learning, there is no interaction involved. Rather than learning from within a situation or environment, it is more like learning from external guidance. A significant problem with supervised learning is that obtaining examples of desired behavior that are both accurate and representative of all the situations in which the agent must act is often impractical in interactive problems. It is essential that agents can learn both from their own experience and from their environment in uncharted territory. By combining dynamic programming and supervised learning, reinforcement learning generates a machine-learning system that approaches human learning very closely [33].

The trade-off between exploration and exploitation is one of the challenges of reinforcement learning. A reinforcement-learning agent may aim to collect rewards by preferring actions that has is tried in the past and found effective for producing rewards. However, such actions can be discovered only by experimenting with actions it has never chosen before. Agents must exploit their existing knowledge to obtain reward, but they also must explore to make better future decisions. Exploration and exploitation cannot be pursued exclusively and must be balanced. In order to find the best action, the agent must try a variety of options and then prefer the ones that seem most promising. To gain a reliable estimate of the reward of an action on a stochastic task, it is necessary to try each one many times. Generally speaking, no issues related to the balance between exploration and exploitation arise as a result of supervised learning, as it is traditionally defined. As well as that, supervised learning does not address the issue of exploration, and the responsibility for exploration is given to experts [33].

The other important feature of reinforcement learning is the explicit consideration of the entire problem of a goal-directed agent interacting with an uncertain environment. In contrast, many other approaches focus on subproblems without thinking about how they fit into a broader context. As an example, most machine-learning research is concerned with supervised learning without stating explicitly how such an ability would be useful in practice. Theoretical studies of planning with general goals have not addressed planning's role in real-time decision making, nor have they addressed the question of where the predictive models necessary for planning would come from. In spite of their usefulness, these approaches are restricted to isolated subproblems. Since real-time interaction is not possible and active learning is not available, these limitations are present.

It is important to note that reinforcement learning differs from supervised learning in several ways. A key difference is the absence of input-output pairs. Instead, the agent is told the immediate reward and the subsequent state after selecting an action, but not which action is the most beneficial long-term. Agents must actively learn about possible system states, actions, transitions, and rewards in order to act optimally. As another difference from supervised learning, online performance plays an important role; the assessment of the system can often occur concurrently with learning.

Reinforcement learning begins with a goal-seeking agent that is complete, interactive, and complete with goals. In addition to having explicit goals, reinforcement-learning agents can sense aspects of their environments and take action to affect their environments. Furthermore, it is typically assumed that the agent has to work in an uncertain environment. Reinforcement learning involves both planning and real-time action selection, as well as the acquisition and improvement of environmental models when it involves planning. As long as reinforcement learning involves supervised learning, it does so for specific reasons that determine which capabilities are critical to the success of a learning process and which are not [33].

RL is a special branch of Artificial intelligence (AI) algorithms that is composed of three key elements: an environment, agents, and rewards. A graph of states is searched by AI search algorithms in order to find a satisfactory trajectory. Based on informed or uninformed methods, search algorithms seek a goal state. Combining informed and uninformed methods is similar to exploring and exploiting knowledge. Similar to graphs, planning involves composing states from logical expressions instead of atomic symbols, but is typically carried out within a more complex construct. In most cases, these methods are constrained by predefined models and well-defined constraints. As opposed to conventional search algorithms, reinforcement learning assumes that, at least for discrete cases, the entire state space can be enumerated and stored in memory.

In reinforcement learning, the agent learns from the environment through a trial-error process because there is no supervisors to tell the agent which actions are right or wrong, as would be the case in supervised learning, where a supervisor would guide the agent. In order to solve this problem, there are two main strategies that can be used. One strategy is to conduct a search in behavioral space to identify a combination of action and behavior in the given environment that would perform well in that environment. A second strategy is based on statistical techniques and dynamic programming to estimate the utility of actions and the chances of achieving a specific goal.

2.4.2 Q-Learning Algorithm

The Q-learning method has been widely used in the literature as one of the most effective methods for reinforcement learning. The following section will present a brief description of the Q-learning algorithm and its extensions for advanced Markov decision process (MDP) models to be considered. In an MDP, each agent starts at a given initial state $s_0 \in \mathcal{S}$. Once the current state is observed, each agent selects one action among I actions, $a = \left\{a^1, ..., a^I\right\}$, and receives its corresponding rewards along with the subsequent observations. Meanwhile, the system will move to a new state $s' \in \mathbb{S}$ based on the probability $p(s' \mid s, a)$. Upon each new state, the procedure is repeated, and the step-by-step process can continue for a finite or infinite number of times. As a result of this method, each agent tries to come up with the optimal policy to maximize its own expected long-term average rewards, i.e., $\sum_{t=0}^{\infty} \gamma r_t \left(s_t, \pi^*(s_t)\right)$. Since the agent aims to maximize the expected long-term reward function, a value function must be defined as $\mathcal{V}^\pi : \mathcal{S} \to \mathbb{R}$, which is the expected value obtained by applied policy π from each state $s \in \mathcal{S}$ [34]. Accordingly, the value function $\mathcal{V}$ computes the overall quality of the policy through an infinite horizon and discounted MDP, which can be expressed as follows:

$$\mathcal{V}^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma r_t \left(s_t, a_t\right) | s_0 = s\right] = \mathbb{E}_\pi \left[r_t \left(s_t, a_t\right) + \gamma \mathcal{V}^\pi \left(s_{t+1}\right) | s_0 = s\right]. \tag{2.19}$$

If we define Q as the optimal Q-function that applies to all pairs of states and actions, then we can write the definition of the optimal value function as follows:

$$\mathcal{V}^*(s) = \max_a \left\{\mathcal{Q}^*(s, a)\right\}. \tag{2.20}$$

For all state-action pairs, it is possible to reduce the problem in order to determine the optimal value of the Q function, i.e., $\mathcal{Q}^*(s, a)$, through an iterative process. Following is the rule for updating the Q-function:

$$\mathcal{Q}_{t+1}(s,a) = \mathcal{Q}_t(s,a) + \alpha_t \Big[r_t(s,a) + \gamma \max_{a'} \mathcal{Q}_t(s,a') - \mathcal{Q}_t(s,a) \Big]. \tag{2.21}$$

In this update, the central idea is to find the Temporal Difference (TD) between the predicted Q-value, i.e., $r_t(s,a) + \gamma \max_{a'} \mathcal{Q}_t(s,a')$, and its current value, i.e., $\mathcal{Q}_t(s,a)$. In 2.21, the learning rate α_t indicates how new information affects the current Q-value. It is possible to choose a constant learning rate or to adjust it dynamically during the learning process.

2.5 Literature Reviews

The concept of EH-powered wireless communications has attracted the attention of both industry and academia. Many studies have been conducted in the area of resource allocation for IoT networks considering EH and D2D communications [35–43].

An energy efficient resource allocation strategy for a machine-to-machine cellular network was explored in [35], which focused on two different multiple access strategies: 1) nonorthogonal multiple access (NOMA) and 2) time division multiple access (TDMA), which utilized nonlinear energy harvesting. Through joint power control and time allocation, authors aimed to minimize total network energy consumption while taking into account circuit power consumption. Iterative power control and time allocation algorithms were used to convert the original optimization problem for NOMA into an equivalent suboptimal problem.

EH-powered D2D Communication underlaying a Cellular Network (EH-DCCN) was examined as a resource allocation problem in terms of spectrum and energy [36]. The D2D pairs powered by EH were permitted to utilize the spectrum resources formerly occupied by Cellular Users (CUs). A sum-rate maximization problem of the whole cellular network was formulated in order to investigate the resource allocation problems. This problem accounted for QoS and energy constraints.

Saleem *et al.* investigated resource allocation for D2D communication using energy harvesting underlaying a cellular network [37]. To maximize the overall sum rate subject to quality of service, subcarrier reuse, power, and energy harvesting constraints, they formulated a joint subcarrier assignment and power allocation problem. Resource allocation problem for energy harvesting-powered D2D communications [38], examined solution where D2D pairs first harvest energy and then send the information signals. Through joint time scheduling and power control, they aimed to maximize the sum throughput while satisfying the Signal-to-Interference-plus-Noise Ratio (SINR) of cellular users and considering energy consumption. A joint spectrum resource matching, power allocation, and transmission time optimization problem for EH-assisted D2D communications on cellular networks, assumed that one D2D link can reuse multiple cellular resource blocks (RBs) to increase spectral efficiency (SE) [39]. The objective was maximizing the sum SE of D2D communications while taking into account the QoS of cellular users and the energy harvested by D2D links.

As a means of enhancing the performance of the traditional D2D model in terms of energy preservation, Yu *et al.* considered the introduction of an energy harvesting mechanism (EH) [40]. Considering EH-aided communications, the sum throughput for D2D users was maximized without compromising QoS for cellular users. In further detail, the authors stated that D2D transmissions would only be enabled at the start of a time slot if the batteries of the D2D users still had enough power (which is set as a lower threshold,) for one-slot data transmission. When the batteries run out of energy, it would switch to energy harvesting mode until they reach an upper threshold. A joint resource block and power resource allocation algorithm was proposed as a suboptimal solution. It should be noted that these mentioned works aim to maximize sum-rate or throughput rather than EE [35–40].

With the introduction of EH technology, D2D communications are faced with new resource management challenges. In D2D pairs, for example, when energy consumption increases, the energy level drops, and the CUs cannot reuse it. Therefore, it is necessary to solve the resource allocation problem in EH-based D2D communications to maximize EE. For EH-based D2D communication heterogeneous networks (EH-DHNs), the problem of multiplexing D2D user with cellular user was investigated [41]. As long as the quality of service of cellular users is guaranteed and the EH constraints of D2D links are met, their goal was to maximize the average energy efficiency of all D2D links. EH time slot allocations of D2D users, power and spectrum resource block allocations were all part of the resource allocation problem. For the scenario when multiple EH-based D2D communication links multiplex the uplink channel resources of a single CU, it was focused on improving the energy efficiency of EH-based D2D communication [42]. Based on the variation of transmission requests over time slots and available energy, a short-term sum energy efficiency maximization problem was formulated for EH-based D2D communication to maintain a fixed transmission rate requirement for both CU and D2D links while integrating transmission scheduling and power allocation. To obtain a feasible suboptimal solution, they proposed a two-layer convex approximation iteration algorithm (CAIA). An energy-efficient resource allocation problem is studied for energy harvesting-powered D2D communications that are underlying cellular networks using the harvest-then-transmit protocol [43]. Using joint time allocation and power control, the authors aimed to maximize the EE of D2D communications while satisfying the QoS requirements of CUs and energy causality constraints. To solve the formulated nonconvex optimization problem, an iterative resource allocation scheme based on the Dinkelbach method was proposed by utilizing nonlinear fractional programming. As a part of each iteration, the optimization problem was decomposed into two subproblems: the power control subproblem and the time allocation subproblem.

Simultaneous wireless information and power transfer (SWIPT) is a promising approach which can be used to address EH methods [44–47]. In this technique, a receiver harvests energy from the RF signals and processes information simultaneously with either a power splitting (PS) or TS scheme or a combination of both. Kuang *et al.* [41], Luo *et al.* [42], and Wang *et al.* [43] concentrated on maximising EE as a single objective optimization problem, but without considering SWIPT or the notion of spectrum sharing. It should be noted that Zhou *et al.* [48] have taken advantage of SWIPT technique in an attempt to allocate power and spectrum resources jointly so as to maximize the amount of energy harvested by D2D pairs as well as CUs, even though switching time and the cellular's power were not optimized. An algorithm that utilizes the prematching of RF energy harvesters is proposed [49] to help deal with the nonlinear behavior of RF energy harvesters and to improve the amount of energy harvested by SWIPT-enabled D2D links by optimizing the distribution of spectrum resources between D2D links and uplink communication units. Afterwards, an iterative two-layer algorithm was proposed in order to optimise the D2D transmission power, as well as the PS ratio of each D2D enabled link jointly, through the use of an iterative procedure.

In order to address the high computational complexity of wireless networks with limited resources, a number of studies have investigated the use of learning-based methods for resource management [50–52]. Chen *et al.* aimed to to examine the problem of training federated learning (FL) algorithms over a realistic wireless network [50]. Based on the considered model, wireless users execute an FL algorithm while training their local FL models on their own data and transmitting the local FL models to a base station (BS). The base station then generates a global FL model and sends it back to the users. In wireless training, all parameters are transmitted over wireless links, and packet errors and wireless resource availability may have an adverse effect on training quality. To build an accurate global FL model, the BS needs to select an appropriate subset of users based on the limited wireless bandwidth. The goal of this joint learning and wireless resource allocation and user selection problem is to minimize an FL loss function that measures the FL algorithm's performance. In order to quantify the impact of wireless factors on FL, a closed-form expression for the expected convergence rate of FL is derived first. The optimal transmit power for each user is then calculated based on the expected convergence rate of the FL algorithm.

Lee *et al.* proposed a deep neural network (DNN)-based transmit power control strategy for underlay D2D communications where D2D user equipment (DUE) shares radio resources with cellular user equipment (CUE) [51]. Using a newly proposed DNN structure, a transmit power control strategy for DUE is found. By taking into account both the spectral efficiency of the DUE and the amount of interference at the CUE, the SE of the DUE can be improved while the cellular transmission is not deteriorated. The authors investigated a method of controlling transmit power over underlaid D2D in their future work [52]. Using a DNN, it was proposed that the transmit power of a D2D users can be learned autonomously, such that the maximum weighted sum rate of D2D users can be achieved by taking into account interference from cellular user equipment when learning the transmit power. This proposal differs from conventional transmit power control schemes in which complex optimization problems must be solved iteratively, which may require a considerable amount of computation time to solve, whereas in our proposed scheme, the transmit power can be determined in a relatively short time without complex optimization problems.

In particular, Lee *et al.* proposed a method of resource management based on deep learning system, that controls both the transmit power and the power splitting ratio in order to maximize the sum rate with low computational complexity in D2D networks that have energy harvesting requirements [53]. D2D networks are complicated by the introduction of energy harvesting requirements, and this makes it difficult to design an effective resource management system, as interference signals need to be treated differently from conventional resource management that only aims to maximize the rate of resource utilization. As a response to drawbacks of deep learning methods, a new training algorithm was proposed that controls both transmit power and PS ratio.

It is worth mentioning that the idea of this work was initially inspired by research conducted by Yang *et al.* [49]. They improved the energy efficiency while considering Device-to-Device (D2D) communication and cellular users. It was assumed that all cellular users and D2D pairs can harvest energy from ambient radio-frequency (RF) signals. To this end, a prematching algorithm was proposed to divide D2D links into a SWIPT-enabled group and a non-EH group that cannot meet the EH circuit sensitivity. Using a two-layer iterative algorithm, they optimized the D2D transmission power and the power splitting ratio in order to maximize the EE for each SWIPT-enabled D2D link. Based on the D2D link model we have assumed, the transmit power of the link is taken into account under the realistic energy causality constraint. Energy causality is an important aspect to consider, which means that the total energy consumed by the device during a transmission session should not exceed the total energy harvested at that particular time. As well, the transmit power of both the BS and D2D pairs must be feasible. Last but not least, we were able to optimize subchannel assignment, power splitting ratios, power, and timing allocation in a joint manner that was not considered in previous work. The high computational complexity of this problem, therefore, makes it difficult to solve using optimization techniques or iterative methods so it cannot be easily resolved.

However, to the best of our knowledge, no study has been conducted on the investigation of learning-based resource management in a D2D network that is incorporated with EH. Considering the large state space and dynamic environment we are dealing with in this research, it is difficult to specify an exact state transition model in this dynamic environment. This is why the RL technique was chosen to solve part of the optimization problem of interest in this article. Additionally, since the system can adapt to new environments automatically, there is no need to retrain it or to have prior knowledge of its actions since it can adapt to new situations without any training. It is important to train the network for each time slot when the users' locations change in the system model. This means that RL is a suitable solution for the networks that have different observations and that need to be analyzed. Due to the fact that learning-based techniques can provide powerful problem-solving capabilities, this study attempts to develop EE-based resource management that will maximize EE through joint power, time, and subchannel allocation in EH-based D2D communication systems using learning-based techniques.

References

[1] Qasim Chaudhari. *Quadrature Amplitude Modulation (QAM)*. 2021. URL: `https://wirelesspi.com/quadrature-amplitude-modulation-qam/`.

[2] Huda Al-Khafaji and Haider Alsabbagh. "Simple Analysis of BER Performance for BPSK and MQAM Over Fading Channel". In: *IJCDS Journal* 6 (Sept. 2017), pp. 303–310.

[3] Muhammad Ammar Saeed, Muhammad Zeeshan Khan, Asim Khan, Muhammad Umer Saeed, Muhammad Arshad Shehzad Hassan, and Talal Javed. "Impact of Propagation Path Loss by Varying BTS Height and Frequency for Combining Multiple Path Loss Approaches in Macro-Femto Environment". In: *Arabian Journal for Science and Engineering* 47.2 (2022), pp. 1227–1238.

[4] Zhu Han and KJ Ray Liu. *Resource allocation for wireless networks: basics, techniques, and applications.* Cambridge university press, 2008.

[5] Arash Asadi, Qing Wang, and Vincenzo Mancuso. "A Survey on Device-to-Device Communication in Cellular Networks". In: *IEEE Communications Surveys & Tutorials* 16.4 (2014), pp. 1801–1819.

[6] 3GPP. *3G Release 99*. 1999. URL: `https://www.3gpp.org/specifications-technologies/releases/release-1999`.

[7] 3GPP. *Overview of 3GPP Release 6*. 2009. URL: `https://www.3gpp.org/specifications-technologies/releases/release-6`.

[8] 3GPP. *LTE-Advanced Pro Ready to Go*. 2015. URL: `https://www.3gpp.org/specifications-technologies/releases/release-13`.

[9] Michael Haus, Muhammad Waqas, Aaron Yi Ding, Yong Li, Sasu Tarkoma, and Jörg Ott. "Security and Privacy in Device-to-Device (D2D) Communication: A Review". In: *IEEE Communications Surveys & Tutorials* 19.2 (2017), pp. 1054–1079.

[10] E Calvanese Strinati and Laurent Hérault. "Holistic approach for future energy efficient cellular networks". In: *e and i Elektrotechnik und Informationstechnik* 127.11 (2010), pp. 314–320.

[11] Tomas Edler and Susanne Lundberg. "Energy efficiency enhancements in radio access networks". In: *Ericsson review* 81.1 (2004), pp. 42–51.

[12] P Grant. "MCVE Core 5 Programme, Green radio-the case for more efficient cellular basestations". In: *GLOBECOM*. 2010, pp. 6–8.

[13] David Lister. "An operator's view on green radio". In: *Keynote Speech, GreenComm* (2009).

[14] Abishi Chowdhury and Shital A Raut. "A survey study on internet of things resource management". In: *Journal of Network and Computer Applications* 120 (2018), pp. 42–60.

[15] Massimo Villari, Antonio Celesti, Maria Fazio, and Antonio Puliafito. "AllJoyn Lambda: An architecture for the management of smart environments in IoT". In: *2014 International Conference on Smart Computing Workshops*. 2014, pp. 9–14.

[16] Vangelis Angelakis, Ioannis Avgouleas, Nikolaos Pappas, Emma Fitzgerald, and Di Yuan. "Allocation of Heterogeneous Resources of an IoT Device to Flexible Services". In: *IEEE Internet of Things Journal* 3.5 (2016), pp. 691–700.

[17] CP Vandana and Ajeet A Chikkamannur. "Study of resource discovery trends in Internet of Things (IoT)". In: *International Journal of Advanced Networking and Applications* 8.3 (2016), p. 3084.

[18] Hyunjoong Kang, Marie Kim, MyungNam Bae, Hyo-Chan Bang, and Hyun Yoe. "A conceptual device-rank based resource sharing and collaboration of smart things". In: *Multimedia Tools and Applications* 75.22 (2016), pp. 14569–14581.

[19] Zahra Ghanbari, Nima Jafari Navimipour, Mehdi Hosseinzadeh, and Aso Darwesh. "Resource allocation mechanisms and approaches on the Internet of Things". In: *Cluster Computing* 22.4 (2019), pp. 1253–1282.

[20] Xuemei Li and Li Da Xu. "A Review of Internet of Things—Resource Allocation". In: *IEEE Internet of Things Journal* 8.11 (2021), pp. 8657–8666.

[21] Giuseppe Colistra, Virginia Pilloni, and Luigi Atzori. "The problem of task allocation in the Internet of Things and the consensus-based approach". In: *Computer Networks* 73 (2014), pp. 98–111.

[22] Vangelis Angelakis, Ioannis Avgouleas, Nikolaos Pappas, Emma Fitzgerald, and Di Yuan. "Allocation of heterogeneous resources of an IoT device to flexible services". In: *IEEE Internet of Things Journal* 3.5 (2016), pp. 691–700.

[23] Jun Huang, Ying Yin, Qiang Duan, and Huifang Yan. "A game-theoretic analysis on context-aware resource allocation for device-to-device communications in cloud-centric internet of things". In: *2015 3rd International Conference on Future Internet of Things and Cloud*. IEEE. 2015, pp. 80–86.

[24] Sungwook Kim. "Asymptotic shapley value based resource allocation scheme for IoT services". In: *Computer Networks* 100 (2016), pp. 55–63.

[25] Minhyeop Kim and In-Young Ko. "An efficient resource allocation approach based on a genetic algorithm for composite services in IoT environments". In: *2015 IEEE international conference on web services*. IEEE. 2015, pp. 543–550.

[26] Thar Baker, Emir Ugljanin, Noura Faci, Mohamed Sellami, Zakaria Maamar, and Ejub Kajan. "Everything as a resource: Foundations and illustration through Internet-of-things". In: *Computers in industry* 94 (2018), pp. 62–74.

[27] Zheng Dou, Guangzhen Si, Yun Lin, and Meiyu Wang. "An adaptive resource allocation model with anti-jamming in IoT network". In: *IEEE Access* 7 (2019), pp. 93250–93258.

[28] Naoki Shinohara. “Development of rectenna with wireless communication system”. In: *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP)*. IEEE. 2011, pp. 3970–3973.

[29] Ioannis Krikidis, Stelios Timotheou, Symeon Nikolaou, Gan Zheng, Derrick Wing Kwan Ng, and Robert Schober. “Simultaneous wireless information and power transfer in modern communication systems”. In: *IEEE Communications Magazine* 52.11 (2014), pp. 104–110.

[30] Pulkit Grover and Anant Sahai. “Shannon meets Tesla: Wireless information and power transfer”. In: *2010 IEEE International Symposium on Information Theory*. 2010, pp. 2363–2367.

[31] Tharindu D. Ponnimbaduge Perera, Dushantha Nalin K. Jayakody, Shree Krishna Sharma, Symeon Chatzinotas, and Jun Li. “Simultaneous Wireless Information and Power Transfer (SWIPT): Recent Advances and Future Challenges”. In: *IEEE Communications Surveys & Tutorials* 20.1 (2018), pp. 264–302.

[32] Kaibin Huang and Erik Larsson. “Simultaneous information and power transfer for broadband wireless systems”. In: *IEEE Transactions on Signal Processing* 61.23 (2013), pp. 5972–5986.

[33] Parag Kulkarni. *Reinforcement and systemic machine learning for decision making*. Vol. 1. John Wiley & Sons, 2012.

[34] Nguyen Cong Luong, Dinh Thai Hoang, Shimin Gong, Dusit Niyato, Ping Wang, Ying-Chang Liang, and Dong In Kim. “Applications of Deep Reinforcement Learning in Communications and Networking: A Survey”. In: *IEEE Communications Surveys & Tutorials* 21.4 (2019), pp. 3133–3174.

[35] Zhaohui Yang, Wei Xu, Yijin Pan, Cunhua Pan, and Ming Chen. “Energy Efficient Resource Allocation in Machine-to-Machine Communications With Multiple Access and Energy Harvesting for IoT”. In: *IEEE Internet of Things Journal* 5.1 (2018), pp. 229–245.

[36] Ying Luo, Peilin Hong, Ruolin Su, and Kaiping Xue. “Resource Allocation for Energy Harvesting-Powered D2D Communication Underlaying Cellular Networks”. In: *IEEE Transactions on Vehicular Technology* 66.11 (2017), pp. 10486–10498.

[37] Umber Saleem, Sobia Jangsher, Hassaan Khaliq Qureshi, and Syed Ali Hassan. “Joint Subcarrier and Power Allocation in the Energy-Harvesting-Aided D2D Communication”. In: *IEEE Transactions on Industrial Informatics* 14.6 (2018), pp. 2608–2617.

[38] Haichao Wang, Guoru Ding, Jinlong Wang, Le Wang, Theodoros A. Tsiftsis, and Prabhat K. Sharma. “Resource allocation for energy harvesting-powered D2D communications underlaying cellular networks”. In: *2017 IEEE International Conference on Communications (ICC)*. 2017, pp. 1–6.

[39] Yue Meng, Zhi Zhang, Yuzhen Huang, and Ping Zhang. “Resource Allocation for Energy Harvesting-Aided Device-to-Device Communications: A Matching Game Approach”. In: *IEEE Access* 7 (2019), pp. 175594–175605.

[40] Shuo Yu, Waleed Ejaz, Ling Guan, and Alagan Anpalagan. “Resource Allocation for Energy Harvesting Assisted D2D Communications Underlaying OFDMA Cellular Networks”. In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. 2017, pp. 1–7.

[41] Zhufang Kuang, Gang Liu, Gongqiang Li, and Xiaoheng Deng. “Energy Efficient Resource Allocation Algorithm in Energy Harvesting-Based D2D Heterogeneous Networks”. In: *IEEE Internet of Things Journal* 6.1 (2019), pp. 557–567.

[42] Ying Luo, Peilin Hong, and Ruolin Su. “Energy-Efficient Scheduling and Power Allocation for Energy Harvesting-Based D2D Communication”. In: *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 2017, pp. 1–6.

[43] Ke Wang, Wei Heng, Jinming Hu, Xiang Li, and Jing Wu. “Energy-Efficient Resource Allocation for Energy Harvesting-Powered D2D Communications Underlaying Cellular Networks”. In: *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. 2018, pp. 1–5.

[44] Hongwu Liu, Kyeong Jin Kim, Kyung Sup Kwak, and H. Vincent Poor. “Power Splitting-Based SWIPT With Decode-and-Forward Full-Duplex Relaying”. In: *IEEE Transactions on Wireless Communications* 15.11 (2016), pp. 7561–7577.

[45] Yanqing Xu, Chao Shen, Zhiguo Ding, Xiaofang Sun, Shi Yan, Gang Zhu, and Zhangdui Zhong. “Joint Beamforming and Power-Splitting Control in Downlink Cooperative SWIPT NOMA Systems”. In: *IEEE Transactions on Signal Processing* 65.18 (2017), pp. 4874–4886.

[46] Jie Tang, Arman Shojaeifard, Daniel K. C. So, Kai-Kit Wong, and Nan Zhao. “Energy Efficiency Optimization for CoMP-SWIPT Heterogeneous Networks”. In: *IEEE Transactions on Communications* 66.12 (2018), pp. 6368–6383.

[47] Yongjun Xu, Guoquan Li, Yang Yang, Miao Liu, and Guan Gui. “Robust Resource Allocation and Power Splitting in SWIPT Enabled Heterogeneous Networks: A Robust Minimax Approach”. In: *IEEE Internet of Things Journal* 6.6 (2019), pp. 10799–10811.

[48] Zhenyu Zhou, Caixia Gao, Chen Xu, Tao Chen, Di Zhang, and Shahid Mumtaz. “Energy-Efficient Stable Matching for Resource Allocation in Energy Harvesting-Based Device-to-Device Communications”. In: *IEEE Access* 5 (2017), pp. 15184–15196.

[49] Haohang Yang, Yinghui Ye, Xiaoli Chu, and Mianxiong Dong. “Resource and Power Allocation in SWIPT-Enabled Device-to-Device Communications Based on a Nonlinear Energy Harvesting Model”. In: *IEEE Internet of Things Journal* 7.11 (2020), pp. 10813–10825.

[50] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H. Vincent Poor, and Shuguang Cui. “A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks”. In: *IEEE Transactions on Wireless Communications* 20.1 (2021), pp. 269–283.

[51] Woongsup Lee, Minhoe Kim, and Dong-Ho Cho. “Transmit Power Control Using Deep Neural Network for Underlay Device-to-Device Communication”. In: *IEEE Wireless Communications Letters* 8.1 (2019), pp. 141–144.

[52] Woongsup Lee, Minhoe Kim, and Dong-Ho Cho. “Deep Learning Based Transmit Power Control in Underlaid Device-to-Device Communication”. In: *IEEE Systems Journal* 13.3 (2019), pp. 2551–2554.

[53] Kisong Lee, Jun-Pyo Hong, Hyowoon Seo, and Wan Choi. “Learning-Based Resource Management in Device-to-Device Communications With Energy Harvesting Requirements”. In: *IEEE Transactions on Communications* 68.1 (2020), pp. 402–413.

3 Reinforcement-Learning-Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks

A. Omidkar, A. Khalili, H. H. Nguyen and H. Shafiei, ”Reinforcement-Learning-Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks,” in IEEE Internet of Things Journal, vol. 9, no. 17, pp. 16521-16531, 1 Sept.1, 2022, doi: 10.1109/JIOT.2022.3151001.

In this paper, we have investigated and solved the problem of joint resource allocation and time allocation for D2D communications underlying IoT networks to maximize the total EE of the system. In particular, we considered energy harvesting for D2D and IoT users in which the IoT users harvest energy from the macro BS based on the PS technique, while the D2D users harvest energy from ambient resources based on the TS technique. The considered problem is nonconvex MINLP, which is challenging to solve. To tackle it, we decompose the original problem into three subproblems 1) joint subchannel assignment and PS; 2) power control; and 3) time allocation. The Q-learning method is employed to solve the first subproblem when the MM approach, as well as the Dinkelbach method, are applied to solve the power control subproblem. Simulation results demonstrate that our proposed algorithm could improve average and sum EE compared to the other baseline schemes.

Reinforcement Learning Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networkse

A. Omidkar, A. Khalili *Member, IEEE*, H. H. Nguyen *Senior Member, IEEE* and H. Shafiei

Abstract

This paper proposes a novel approach to improve the Energy Efficiency (EE) of an Energy-Harvesting (EH)-enabled IoT network supported by Simultaneous Wireless Information and Power Transfer (SWIPT). More specifically, the Device-to-Device (D2D) users harvest ambient energy throughout their communication with the time switching (TS) technique, while the Internet of Things (IoT) users harvest energy from the Base Station (BS) based on the power splitting (PS) method. We study the EE optimization problem that takes into account transmit power feasibility conditions for D2D users and IoT users, the minimum data rate requirements for D2D and IoT users, joint spectrum sharing block, and time allocation for the D2D links. The underlying problem is highly non-convex mixed integer non-linear problem (MINLP) and the global optimal solution is intractable. To handle it, we decompose the original problem into three subproblems: (i) joint subchannel allocation and power splitting, (ii) power control, and (iii) time allocation. Since it is difficult to find an exact state model approach in a dynamic environment with a large state space, we exploit a Q-learning method based on reinforcement learning (RL) to solve the first sub-problem. To solve the second sub-problem, we apply a conventional convex optimization technique based on the majorization-minimization (MM) approach and Dinkelbach method. Simulation results not only demonstrate the superiority of our proposed algorithm as compared to other methods in the literature but also confirm impressive EE gains through spectrum sharing and harvested energy from D2D and IoT users.

Index Terms

Device-to-device (D2D) communications, energy harvesting, Internet of Things (IoT) networks, majorization– minimization (MM), mixed-integer nonlinear problem (MINLP), reinforcement learning (RL), resource management, spectrum sharing.

3.1 Introduction

3.1.1 Motivations and State of the Art

With, the explosive growth in the number of wireless devices and various types of communications, including Internet of Things (IoT) and machine-type communications, it is expected that the amount of mobile data traffic and demand for higher data rates will increase drastically [1]. In this context, direct communications between mobile users can create various location-based peer-to-peer services to help offload traffic from congested cellular networks. Device-to-Device (D2D) communication underlaying cellular networks have been introduced to exploit better direct links instead of transmitting via the Base Station (BS). Furthermore, D2D communication has the potential to improve spectral efficiency while reducing power consumption [2]. However, the extra energy consumption of mobile and IoT devices whose battery capacity is limited is the main obstacle to benefit from D2D and IoT communications completely. In this regard, harvesting energy from reusable external sources such as solar, wind, vibrations, thermoelectric, and Radio Frequency (RF) could help to prolong the lifetime of energy-constraint wireless devices [3, 4].

Given the increasing number of wireless users, employing resource allocation techniques is essential in modern wireless networks to facilitate devices competing for limited resources, such as time and frequency bands. It provides road maps for appropriate energy consumption at individual users for interference minimization, as well as ensuring an adequate supply of energy for reliable and efficient communications between devices. Unfortunately, obtaining a jointly resource management design for the D2D network that includes such numerous aspects is typically not easy and often requires significant computing power in order to achieve near-optimal solutions. Thus, the main goal of this paper is to develop an acceptable resource allocation scheme with low computing cost for EH-assisted D2D and IoT communication that depends on spectrum sharing.

3.1.2 Related Works

There are numerous works on resource allocation in D2D systems with Energy Harvesting (EH) to enhance the energy efficiency (EE) performance, and hence battery lifetime of wireless devices [5–13]. The problem of minimizing the total energy consumption of the network via joint power control and time allocation concerning circuit power consumption is studied in [5] with a focus on multiple access techniques. In [6], the authors investigated spectrum resource matching and power allocation to maximize a sum transmission rate with consideration of the available energy and the Quality of Service (QoS) requirements of both cellular users (CUs) and D2D. The authors in [7] formulated a joint power allocation and subcarrier assignment for EH-assisted D2D communication while maximizing the overall sum rate for EH-based D2D links and multiple cellular users. In [8], a resource allocation problem was investigated for EH-powered D2D communication to maximize the throughput by joint time switching and power control subject to Signal-to-Interference-plus-Noise Ratio (SINR) criterion and energy constraint. A joint spectrum, power, and time allocation in EH-aided D2D communications is proposed in [9] to maximize Spectral Efficiency (SE) of D2D communications subject to QoS of cellular users and the harvested energy of D2D links. In [10], the throughput for D2D pairs was maximized in an EH-aided communications model where a heuristic solution was suggested. It should be considered that all these mentioned works aim to maximize the sum-rate or throughput rather than EE [6–10].

The incorporation of EH technology presents new challenges to resource management in D2D communications. For example, when the energy consumption is increased in D2D pairs, the energy level becomes very low, and there is no energy to be reused by CUs. Thus, solving the resource allocation problem to maximize

EE in EH-based D2D communications is necessary. In terms of maximizing EE, the work in [11] maximizes the average EE of the D2D links in a Heterogeneous Network (HetNet) by taking into the account time slot, power control, and spectrum allocation. Furthermore, EE maximization was investigated in [12] under both the transmission rate constraint and energy restriction of cellular and D2D links. Moreover, maximizing EE via join time allocation and power control was investigated in [13] while satisfying the energy constraint and QoS for CUs.

To address EH methods, an exciting and promising approach is Simultaneous Wireless Information and Power Transfer (SWIPT) [14–17]. In this technique, a receiver harvests energy from the RF signals and processes information simultaneously with either a power splitting (PS) or time switching (TS) scheme or a combination of both. The authors in [11–13] focused on maximizing EE as a single objective optimization problem, but without considering SWIPT or spectrum sharing concept. By employing the SWIPT technique, the authors in [18] addressed joint spectrum resource and power allocation to maximize EE of D2D pairs and the harvested energy by CUs while the switching time and cellular's power were not optimized. To deal with the non-linear behavior of RF energy harvesters, a pre-matching algorithm is proposed in [19] to improve the sum EE of SWIPT-enabled D2D links by optimizing spectrum resource sharing between D2D links and uplink CUs. Then a two-layer iterative algorithm was proposed to jointly optimize the D2D transmission power and the power splitting ratio for each enabled D2D link.

Regarding wireless networks operating with restricted resources, various studies have investigated learning-based techniques for resource management to address the high computational complexity of iterative methods [20–22]. An optimal resource allocation scheme and user selection via joint learning techniques is proposed in [20] to minimize a federated learning loss function. A few deep learning approaches have recently been applied to D2D networks for optimizing the transmit power, but without considering EH [21, 22]. In particular, the authors in [23] proposed a resource management method based on deep learning, which controls both the transmit power and the power splitting ratio. Moreover, for cognitive D2D-based IoT networks, the authors in [24] presented a QoS-driven social-aware D2D communication network model for social and cognitive IoT by optimizing the network performance under different QoS requirements.

However, to the best of our knowledge, there is no study on the investigation of learning-based resource management in a D2D network with EH. One of the most critical advantages of RL is that there is no need to have access to large labeled datasets, which is a significant benefit because when the volume of data increases, labeling becomes more and more complex. Unlike supervised learning, which is mostly employed in an input-output manner, RL is goal-oriented and can be applied for a sequence of actions. We choose the RL technique to solve part of the optimization problem of interest in this paper because it is difficult to specify an exact state transition model in a dynamic environment with a huge state space. Furthermore, retraining and previous knowledge of the system's actions are not required because it can adapt to new environments automatically. When the users' locations change in our system model, the network should be trained in each time slot. Thus, RL is an appropriate solution for the networks with different observations.

Given that learning-based techniques have a powerful problem-solving capability, this paper aims to develop learning-based resource management for maximizing EE via joint power, time, and subchannel allocation in EH-based D2D communication.

3.1.3 Contributions

To the best of our knowledge, resource allocation design for EH-aided IoT network with the help of SWIPT-enabled D2D scheme by means of learning techniques and convex optimization has not been studied in the literature. Thus, it motivates us to study the potential benefits of resource allocation in IoT systems with EH via the SWIPT scheme. In particular, we are interested to find out whether employing learning schemes is beneficial in EH-aided D2D and IoT network as far as EE maximization is concerned. Our main contributions are summarized as follows.

- We investigate a potential deployment of resource allocation techniques to harvest energy in the networks with limited battery lifetime. Specifically, to establish EE system, we consider the SWIPT scheme when D2D pairs employ the TS approach to harvest energy from the environment, while IoT users use the PS approach to harvest energy from the BS.

- To assess the performance, a mixed-integer nonlinear programming (MINLP) EE problem is formulated by jointly optimizing allocation of subchannel, power splitting factor, power, and time.

- We decompose the original EE optimization problem into three subproblems: (i) subchannel assignment and power splitting factor, (ii) power allocation, and (iii) time scheduling. For the first subproblem which includes discrete variables, the Q-learning technique is applied. In fact, optimizing all variables using the learning technique is not practically feasible since the problem's size becomes very large. Furthermore, mini-batch has a restricted size and cannot be utilized for a large number of variables. On the other hand, RL is an appropriate technique for discrete variables, but it may also be utilized for continuous variables, such as power allocation. It is worthwhile to note that although the power allocation vectors can be discretized and solved via RL algorithms, the solution can be inaccurate [25].

- In order to handle the above problem for continuous variables, i.e., power and time allocation, we use the MM approach as well as the Dinkelbach method to transform the original non-convex problem into a convex problem and obtain the locally-optimal solution.

- Simulation results demonstrate the effectiveness of the proposed RL-based resource allocation approach for the considered EH-aided D2D and IoT network with multiple D2D devices under the SWIPT scheme. The superior performance is demonstrated in terms of the average and sum EE and compared to [19] and other baseline schemes.

The rest of the paper is organized as follows. Section 3.2 described the system model. The EE optimization problem is formulated in Section 3.3. Section 3.4 presents the proposed solution, including principles

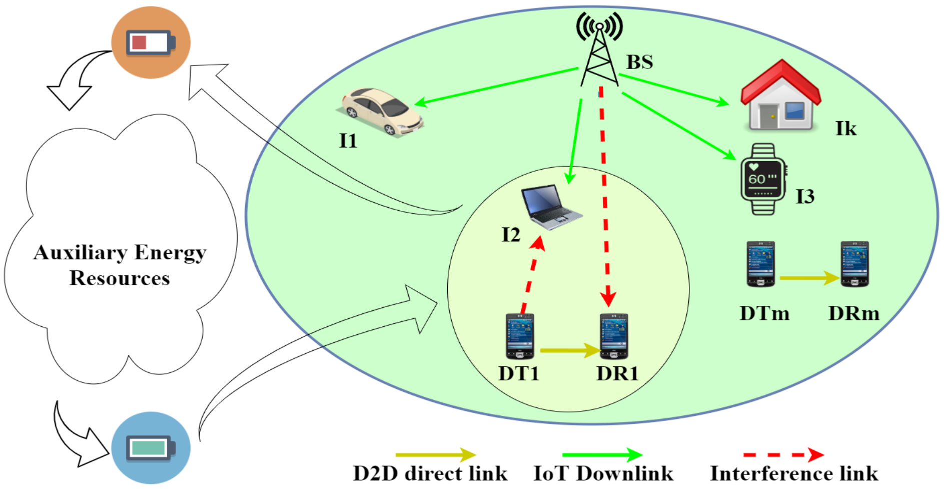


Figure 3.1: D2D communications in an IoT network.

of reinforcement learning, subchannel and power splitting learning algorithm, power, and time allocation. Section 3.5 presents complexity analysis. Simulation results and discussion are evaluated in Section 3.6. Section 3.7 concludes the paper.

3.2 System Model

In our system model, the EH-based D2D communication is underlaying an IoT network, which includes one BS, multiple IoT users, and numerous D2D pairs. In this system model, we assume that users are moving and they are not in the same location vicinity. Therefore, the wireless channels are not constant and vary with users' locations. Thus, the considered resource allocation depends on channel gains, and hence it is a dynamic problem and requires online learning. In this scenario, spectrum sharing is used for both D2D and IoT users in which each IoT device has the same subchannel with a D2D pair. As illustrated in Fig. 3.1, D2D pairs (here DT means a D2D transmitter and DR means a D2D receiver) can reuse the channels occupied by IoT users. Moreover, similar to [4], it is assumed that the D2D transmitters can harvest energy from the environment based on the TS scheme. On the other hand, all IoT users use the PS scheme to harvest energy from the BS. Such assumptions are quite reasonable in scenarios when the data rate is not a critical requirement compared with latency, reliability, and availability for IoT users, and when IoT users need to operate at locations that that might not allow to harvest energy from the environment in a reliable manner (e.g., IoT sensors deployed in factories or in basement of buildings). Thus, IoT users can operate with low energy harvested from the BS. On the other hand, D2D users might communicate with high data rates and need to rely on more powerful auxiliary energy resources, i.e., environment, to harvest sufficient power.

For D2D pairs, in the first phase, energy would be harvested in τ_h, and in the second phase, information will be transferred in τ_i [26]. Consequently, $\tau_h + \tau_i \leq T$ should be satisfied where T is the periodicity of the two-phase transmission. Concerning power splitter for IoT users' receivers, $0 \leq \alpha_k \leq 1$ is the power splitting factor, which specifies the fraction of the received signal power used for information decoding, whereas the remaining portion, i.e., $1-\alpha_k$, is used for harvesting energy.

A macrocell is considered with $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ IoT users and $m \in \mathcal{M} = \{1, 2, \ldots, M\}$ D2D pairs employing $n \in \mathcal{N} = \{1, 2, \ldots, N\}$ subchannels for transmission. The variable $p_{m,k}^{(n)}$ is the transmit power from the mth D2D user toward the kth IoT user on the same subchannel n, whereas $p_k^{(n)}$ is the transmit power from the BS to the kth IoT user. Moreover, $h_m^{(n)}$, $h_k^{(n)}$, $h_{m,k}^{(n)}$, and $h_{k,m}^{(n)}$ are the channel gains between the mth D2D transmitter and its receiver, between the BS and the kth IoT user, between the mth D2D transmitter and the kth IoT user, and between the kth IoT user and the mth D2D transmitter, respectively. The binary variable $\rho_{m,k}^{(n)} \in \{0, 1\}$ indicates the subchannel assignment which is shared between the mth D2D user and the kth IoT. Moreover, each D2D transmitter is equipped with a fixed battery and an extra EH device.

The EE of the system, defined as the total (normalized) data rate[1] of the IoT network per total power consumption [11, 27], is

$$\mathrm{EE} = \frac{R_{\text{tot}}}{P_{\text{tot}}}, \tag{3.1}$$

where R_{tot} and P_{tot} are given by

$$R_{\text{tot}} = \sum_{m=1}^{M} r_m + \sum_{k=1}^{K} r_k, \tag{3.2}$$

$$P_{\text{tot}} = p_{\text{tot}-\text{D2D}} + p_{\text{tot}-\text{BS}}, \tag{3.3}$$

in which r_m, r_k, $p_{\text{tot}-\text{D2D}}$ and $p_{\text{tot}-\text{BS}}$ are the sum rate of D2D pair, sum rate of IoT users, the total power of D2D pairs, and total power of BS, respectively. They are defined as follows:

$$r_m = \sum_{n=1}^{N} \tau_i \rho_{m,k}^{(n)} \log\left[1 + \Gamma_m^{(n)}\right], \tag{3.4}$$

$$r_k = \sum_{n=1}^{N} \rho_{m,k}^{(n)} \log\left[1 + \Gamma_k^{(n)}\right], \tag{3.5}$$

$$p_{\text{tot}-\text{D2D}} = \sum_{m=1}^{M} \sum_{n=1}^{N} \tau_i p_{m,k}^{(n)} + \sum_{m=1}^{M} p_m^{(\text{circuit})}, \tag{3.6}$$

$$p_{\text{tot}-\text{BS}} = \sum_{k=1}^{K} \sum_{n=1}^{N} p_k^{(n)} + p_{\text{BS}}^{(\text{circuit})}, \tag{3.7}$$

where $\Gamma_m^{(n)}$, $\Gamma_k^{(n)}$, $p_m^{(\text{circuit})}$, and $p_{\text{BS}}^{(\text{circuit})}$ are the received SINR of the mth D2D user on subchannel n, and received SINR of the kth IoT user on subchannel n, the circuit power of the mth D2D pairs, and circuit power of the BS, respectively.

[1] In this paper, the data rates referred to normalized data rates, i.e., normalized to 1 Hz of bandwidth. As such, the units of data rates are bits/s/Hz.

$$\Gamma_m^{(n)} = \frac{(p_{m,k}^{(n)} + \psi_m \tau_h) \left| h_m^{(n)} \right|^2}{\sum_{k=1}^{K} \rho_{m,k} \alpha_k p_k^{(n)} \left| h_{k,m}^{(n)} \right|^2 + N_0}, \tag{3.8}$$

$$\Gamma_k^{(n)} = \frac{\alpha_k p_k^{(n)} \left| h_k^{(n)} \right|^2}{\alpha_k \sum_{m=1}^{M} \rho_{m,k}^{(n)} (p_{m,k}^{(n)} + \psi_m \tau_h) \left| h_{m,k}^{(n)} \right|^2 + N_0}, \tag{3.9}$$

here, $\psi_m \tau_h$,and N_0 are the harvested energy for the D2D transmitter, and power of additive white Gaussian noise, respectively. The harvested energy by the IoT users $\mathrm{EH}_{\mathrm{IoT}}$ is computed as

$$\mathrm{EH}_{\mathrm{IoT}} = \sum_{n=1}^{N} \sum_{k=1}^{K} \phi_k (1 - \alpha_k) p_k^{(n)} \left| h_k^{(n)} \right|^2, \tag{3.10}$$

where ϕ_k is the amplifier's efficiency factor for the kth IoT user, which also includes a time constant to convert power to energy.

3.3 Problem Formulation

In this section, we are interested in finding a joint subchannel assignment, power, and time allocation to maximize the EE while considering minimum data rate requirement for the D2D and IoT users under a spectrum sharing scenario. The optimization problem is formally stated as follows:

$$\max_{\rho_{m,k}, \alpha_k, p_{m,k}, p_k, \tau_i, \tau_h} \mathrm{EE} \tag{3.11a}$$

$$\text{s.t.} : r_m \geq R_{\min}^{(\mathrm{D2D})}, \quad \forall m \in \mathcal{M}, \tag{3.11b}$$

$$r_k \geq R_{\min}^{(\mathrm{IoT})}, \quad \forall k \in \mathcal{K}, \tag{3.11c}$$

$$\tau_i, \tau_h \geq 0, \tag{3.11d}$$

$$\tau_i + \tau_h \leq T, \tag{3.11e}$$

$$\mathrm{EH}_{\mathrm{IoT}} \geq E_{\min}, \tag{3.11f}$$

$$\rho_{m,k}^{(n)} \in \{0, 1\}, \tag{3.11g}$$

$$\sum_{m=1}^{M} \rho_{m,k}^{(n)} \psi_m \tau_h \leq E_{\max}, \quad \forall k \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{3.11h}$$

$$\sum_{n=1}^{N} \sum_{k=1}^{K} \rho_{m,k}^{(n)} p_k^{(n)} \leq p_{\max}^{(\mathrm{BS})}, \quad \forall m \in \mathcal{M}, \tag{3.11i}$$

$$\sum_{n=1}^{N} \sum_{m=1}^{M} \rho_{m,k}^{(n)} p_{m,k}^{(n)} \leq p_{\max}^{(\mathrm{D2D})}, \quad \forall k \in \mathcal{K}, \tag{3.11j}$$

$$\sum_{k=1}^{K} \sum_{m=1}^{M} \rho_{m,k}^{(n)} \leq 1, \quad \forall n \in \mathcal{N}, \tag{3.11k}$$

$$\rho_{m,k}^{(n)} + \rho_{m,\acute{k}}^{(\acute{n})} \leq 1,$$

$$\forall m \in \mathcal{M},\ \forall n, \acute{n} \in \mathcal{N},\ \forall k, \acute{k} \in \mathcal{K} \text{ and } k \neq \acute{k}, \tag{3.11l}$$

where $R_{\min}^{(\mathrm{D2D})}$, $R_{\min}^{(\mathrm{IoT})}$, and $E_{\min}$ are the minimum value for the sum rate of D2D pairs, the minimum value for the sum rate of IoT users, and the minimum value for the harvested energy of each IoT user, respectively. In Eq. (3.11), constraints (3.11b) and (3.11c) guarantee the minimum QoS enjoyed by each D2D user and each IoT user, respectively. Constraint (3.11f) shows that each IoT user should meet the minimum energy harvesting while (3.11h) is an energy causality constraint imposed on the amount of the transmit energy of each D2D link. The feasibility of transmit powers of the BS and D2D pairs are indicated by constraints (3.11i) and (3.11j), respectively, in which $p_{\max}^{(\mathrm{BS})}$ and $p_{\max}^{(\mathrm{DoD})}$ are the maximum transmit powers of the BS and D2D transmitters, respectively. Finally, (3.11k) assigns each subchannel between D2D user and IoT user, while (3.11l) denotes the D2D-IoT matching constraint, which determines that only a single D2D pair can reuse a shared subchannel of one IoT user.

3.4 Proposed Solution

Observe that the optimization problem (3.11) includes continuous variables for the powers of IoT and D2D users, the time durations for information processing and energy harvesting, and the power splitting ratio for the IoT users. Moreover, the subcarrier assignment is represented by discrete variables. Therefore, the optimization problem (3.11) is a mixed integer nonlinear programming (MINLP), which is a non-convex optimization problem [27] and difficult to solve. Additionally, in practice, the channel qualities of D2D and IoT links could change dynamically. As such, traditional optimization solutions are generally not suitable to obtain the resource management policy under consideration in this work. On the other hand, Reinforcement Learning (RL) model is a dynamic programming strategy that can be adapted to the environment and learn the optimal solution over a dynamic condition.

Considering that numerous devices want to access a shared spectrum resource, we represent the optimization problem as a multi-agent formulation for learning optimization policy under energy causality constraint, various QoS requirements, and the feasibility of transmit powers of the BS D2D users. We decompose the original problem into three subproblems, corresponding to (i) subchannel and power splitting factor, (ii) power allocation, and (iii) time scheduling. It should be mentioned that trying to optimize all the variables via the learning method is very difficult since the size of the problem would increase significantly. Moreover, RL cannot be used in the problem with the large number of variables because the size of mini-batch is limited and also it is more appropriate for discrete variables [25]. Despite the fact that the power allocation vectors may be discretized and solved using RL methods, the results might be inaccurate. Thus, we utilize the MM technique and Dinkelbach algorithm to find the locally optimal solution using the convex optimization method for solving the power and time allocation problem. In particular, power is a continuous variable that should be discretized in reinforcement learning algorithms. If we choose a poor discrete state or action space,

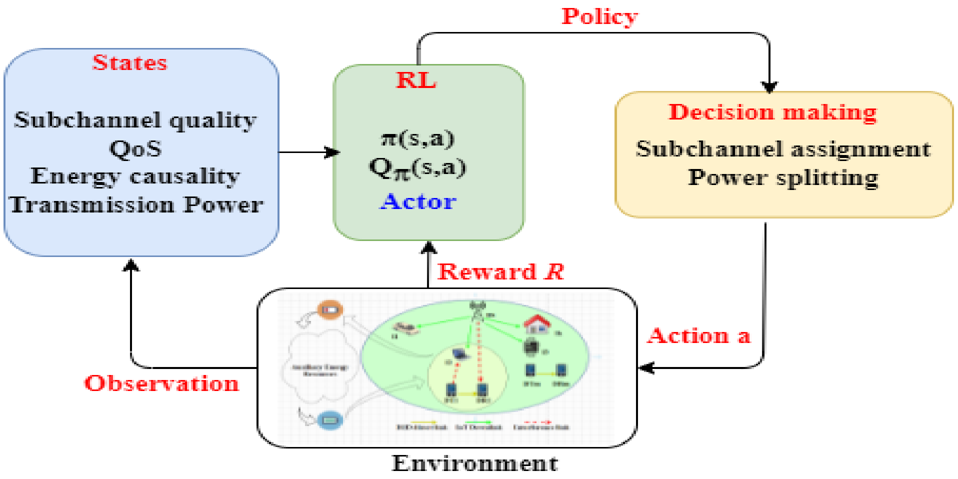


Figure 3.2: Architecture of the multi-agent RL-based resource management.

we may need to bring a hidden state into the problem, making the learning process of the optimal policy impossible. Here, we employ a Q-learning algorithm to train and update the subchannel and power splitting factor using the feedback signal received from the environment. At the higher level, we optimize power and time allocation as the reward function for the lower level. In the following subsection, we briefly describe the RL technique and its application in solving dynamic problems.

3.4.1 Reinforcement Learning

RL is a type of machine learning technique that enables agents to learn from their own actions and experiences in an interactive environment through the trial and error method. In particular, the agent, or agents in the multi-agent RL models, can interact with their environment at a sequence of discrete-time step, signified by $t = \{0, 1, \dots, \}$, and state s_t is a situation which is occupied by the agent at time t. Then, the agent takes action $a_t \in \mathcal{A}(s_t)$ where $\mathcal{A}(s_t)$ is the set of actions available at state s_t. At the next time slot, the agent obtains reward r_{t+1} in the feedback signal and moves to a new state s_{t+1}. At each time step, the policy is utilized by the agent to identify the next action based on the current state. This is illustrated in Fig. 3.2. The policy map gives the probability of taking action a when in state s i.e.,

$$\begin{aligned} &\pi : A \times S \to [0, 1] \\ &\pi\left(s, a\right) = \Pr\left(s_t = s \mid a_t = a\right). \end{aligned} \tag{3.12}$$

The policy should be updated at each time slot based on the previous experience and reward. In the RL algorithm, the primary purpose is to maximize the reward function received from the environment. Moreover, the return function is a sum of future discounted rewards, which is expressed as:

$$R_t \triangleq \sum_{k=0}^{T-1} \gamma^k r_{t+k+1}, \tag{3.13}$$

where T is the final time step and $\gamma \in [0.1, 0.9]$ is the discount factor that denotes how much the function Q in the state s depends on future actions. If γ is close to 0, the agent will only consider immediate rewards.

If γ is close to 1, the agent will consider rewards with greater weight in the future and be ready to delay rewards. Then, the RL algorithm can be improved by defining the action-value function $Q_{\pi}(s,a)$ and finding a policy that maximizes the return function by maintaining a set of estimates of expected returns for some policy. The action-value function, or Q-function, can be denoted as

$$Q_{\pi}(s,a) = \mathbb{E}_{\pi}\left\{R_t \mid s_t = s, a_t = a\right\}. \tag{3.14}$$

Using the return function (3.13), the Q-function can be divided into an immediate reward and the discounted Q-function of the successor state as

$$Q_{\pi}(s,a) = \mathbb{E}_{\pi}\left\{r_{t+1} + \gamma Q_{\pi}(s_{t+1}, a_{t+1}) \mid s_t = s, a_t = a\right\}. \tag{3.15}$$

The optimal Q-function which satisfies the Bellman optimally equation [28] is given by

$$Q^{*}(s,a) = \mathbb{E}_{\pi}\left\{r_{t+1} + \gamma \max_{a'} Q^{*}(s_{t+1}, a') \mid s_t = s, a_t = a\right\}. \tag{3.16}$$

It should be pointed out that the optimality of the Bellman equation is non-linear and there is no closed-form solution for it. Thus, to solve (3.16), an iterative method can be used. As reported by this method, the learned Q-function is recursively updated as

$$Q(s_t, a_t) \leftarrow Q^{*} + \alpha\left[r_{t+1} + \gamma \max_{a} Q(s_{t+1}, a) - Q(s_{t+1}, a)\right], \tag{3.17}$$

where $\alpha \in [0,1]$ is the learning rate. One of the significant advantages of this method is that the optimal action-value function $Q^{*}(s,a)$ is directly approximated by the learned Q-function derived by (3.17), independent of the policy followed by the agent [28].

3.4.2 Subchannel Allocation and Power Splitting as a Q-Leaning Algorithm

The key concept of Q-learning is to use an episode as a training period, going from the beginning to the end of states. The agent goes on to the next episode to learn when each episode is completed. As a result, an episode is the outer loop of Q-learning, while every phase of the episode is the inner loop. We reset the state at the end of every episode. In each episode, the policy should be applied for the states, actions, and rewards. As indicated in the return function (3.13), the policy can be used to achieve accumulative rewards from the environment.

In the following, we solve the EE optimization problem to find subchannel allocation and power splitting based on the multi-agent Q-learning. In this model, each user is considered as an intelligent agent that monitors the behavior of the network state s_t and takes its associated action a_t at the current time slot t. At the end of each time slot, the reward will be generated and the negative or positive effect of the executed action will be evaluated. In what follows, we describe the agents, states, actions, and rewards associated with our Q-learning model:

- **Agents**: D2D and IoT users

- **States**: The channel gain between the D2D transmitter and its receiver, channel gains between BS and IoT users, and channel gains between D2D and IoT users are considered as states in our model.
- **Actions**: Available actions at each state are the decisions corresponding to subchannel allocation and power splitting factor.
- **Reward**: Generally, reward is related to objective function which is EE in our paper. If the users at time t select optimal channels, a reward value will be obtained. In our model, the reward function is the EE defined in (3.11). All of the constraints associated with subchannel allocation and power splitting should be satisfied as the statements of reward function. Thus, the reward function is applicable only if constraints (3.11b), (3.11c), (3.11f), (3.11h), (3.11i), (3.11j), (3.11k), (3.11l) are satisfied. Otherwise, it is set to 0. Mathematically, the reward function is expressed as

$$R_t = \begin{cases} \mathrm{EE}_t(s,a), & \text{if (3.11b),(3.11c), (3.11f)-(3.11l),} \\ 0, & \text{otherwise,} \end{cases} \tag{3.18}$$

  where EE_t reflects the utility of EE.

Algorithm 1 Q-Learning for subchannel allocation and power splitting.

```
Initialize: For all D2D and IoT users, initialize their Q-function $Q_\pi(s,a)$ for all $s \in S$, $\forall a \in \mathcal{A}(s)$, $\forall m \in \mathcal{M}$ and $\forall k \in \mathcal{K}$, arbitrarily.
for each episode to $EP$ episodes do do
    Initialize the states by sending pilot signal,
    for each step of an episode to T steps do do
        Each agent, i.e. users, selects an action $a$ at the associated state $s$ based on policy $Q_\pi(s,a)$,
        evaluate the state $s = s_t$ true
        Check constraints (3.11b), (3.11c), (3.11f), (3.11h), (3.11i), (3.11j), (3.11k), (3.11l) ;
        if constraints are met then
            The users obtain the immediate reward $R_t(s,a)$.
        else
            Otherwise, the BS and D2D transmitter will not reply anything and the users will achieve a negative reward, $R_t(s,a) = 0$.
        end if
        Update $Q(s_t, a_t)$ according to (3.17)
        $s = s_{t+1}$
        $t = t + 1$
    end for  T steps
end for  EP episodes
```

Algorithm 1 explains the Q-learning for the optimization of power. At the initial of each training, the state, i.e., the channels, should be initialized. Then, the D2D and IoT users' received power can be obtained by transmitting a pilot signal from the associated D2D user and BS via the randomly selected channel. After that, each D2D and IoT user report their own current state to their associated transmitters. Thus, the general state information of all users will be achieved. Then, the D2D transmitter and BS forward the general state information s to all assigned users. It should be mentioned that each episode includes T steps ($T = 100$ is used in obtaining the numerical results in Section 3.6). According to our policy, $Q_\pi(s, a)$, the action a_t will be selected in each step of an episode. In each step t, according to the condition of the available channels (states), users estimate the immediate reward R_t if $EE_t(s, a)$ is maximized based on all constraints, and then each user updates Q-values $Q_\pi(s, a)$. Therefore, the subchannels and power splitting factor will be obtained based on all the constraints and states. Each episode ends when all constraints of all users are satisfied or when the maximum number of steps T is reached. Based on (3.13), the total episode reward is the accumulation of immediate rewards at all steps within an episode.

As discussed earlier, Q-learning focuses on maximizing the EE of the whole network, while the power and time allocation will be obtained from solving an optimization problem.

3.4.3 Power and Time Allocation

In this section, we propose a solution for time and power allocation subject to subchannel assignment, power splitting factor, QoS for both IoT and D2D pairs, and energy causality constraint. To this end, we first solve a power allocation subproblem, and then we solve the time allocation problem for the obtained results.

Power Allocation

In this section, we employ an iterative algorithm known as Dinkelbach method [29] in order to solve problem (3.11). Specifically, the optimization problem is solved with respect to power vectors $\mathbf{p}^* = (p_{1,K}^{*(n)}, p_{2,K}^{*(n)}, \ldots, p_{M,K}^{*(n)})$ and $\acute{\mathbf{p}}^* = (p_1^{*(n)}, p_2^{*(n)}, \ldots, p_K^{*(n)})$, which represent the transmit powers of D2D pairs and IoT users, respectively. After substituting (3.4), (3.5), (3.8), and (3.9) in (3.2), the total rate R_{tot} is given as (3.19).

$$
\begin{aligned}
R_{\text{tot}} = {} & \sum_{m=1}^{M} \underbrace{\sum_{n=1}^{N} \tau_i \log \left[\sum_{k=1}^{K} \alpha_k p_k^{(n)} \left| h_{k,m}^{(n)} \right|^2 + N_0 + p_{m,k}^{(n)} \left| h_m^{(n)} \right|^2 \right]}_{f_{1,m}} + \sum_{k=1}^{K} \underbrace{\sum_{n=1}^{N} \log \left[\alpha_k \sum_{m=1}^{M} p_{m,k}^{(n)} \left| h_{m,k}^{(n)} \right|^2 + N_0 + \alpha_k p_k^{(n)} \left| h_k^{(n)} \right|^2 \right]}_{f_{2,k}} \\
& - \sum_{m=1}^{M} \underbrace{\tau_i \log \left[\sum_{k=1}^{K} \alpha_k p_k^{(n)} \left| h_{k,m}^{(n)} \right|^2 + N_0 \right]}_{g_{1,m}} - \sum_{k=1}^{K} \underbrace{\log \left[\alpha_k \sum_{m=1}^{M} p_{m,k}^{(n)} \left| h_{m,k}^{(n)} \right|^2 + N_0 \right]}_{g_{2,k}}. \qquad (3.19)
\end{aligned}
$$

To solve the optimization problem (3.11), we first rewrite the objective function EE as the difference of two concave functions as follows:

$$\begin{aligned}R_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p}) &= \left[\sum_{m=1}^{M} f_{1,m}(\acute{\mathbf{p}},\mathbf{p}) + \sum_{k=1}^{K} f_{2,k}(\acute{\mathbf{p}},\mathbf{p})\right] \\ &\quad - \left[\sum_{m=1}^{M} g_{1,m}(\acute{\mathbf{p}}) + \sum_{k=1}^{K} g_{2,k}(\mathbf{p})\right], \\ \text{EE}(\mathbf{p},\acute{\mathbf{p}}) &= \frac{R_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p})}{P_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p})},\end{aligned} \tag{3.20}$$

where the terms $f_{1,m}$,$f_{2,k}$,$g_{1,m}$, and $g_{2,k}$ defined in (3.19), are positive with respect to τ and $\rho_{m,k}$. Thus the optimization problem (3.11) is reformulated as

$$\begin{aligned}&\max_{\acute{\mathbf{p}},\mathbf{p}} \frac{R_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p})}{P_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p})} \\ &\text{s.t.} : f_{1,m}(\acute{\mathbf{p}},\mathbf{p}) - g_{1,m}(\acute{\mathbf{p}}) \geq R_{\text{min}}^{(\text{D2D})}, \\ &\qquad f_{2,k}(\acute{\mathbf{p}},\mathbf{p}) - g_{2,k}(\mathbf{p}) \geq R_{\text{min}}^{(\text{IoT})}, \\ &\qquad (3.11\text{f}), (3.11\text{h}), (3.11\text{i}), (3.11\text{j}).\end{aligned} \tag{3.21}$$

Although each term in (3.21) is a concave function, the sum or difference of two or more concave functions is not necessarily concave. Therefore, we employ the majorization-minimization (MM) approach to create a sequence of surrogate functions using the first-order Taylor approximation to make it convex at the neighborhood of $(\mathbf{p}^{t-1})$ and $(\acute{\mathbf{p}}^{t-1})$ as follows [30, 31]:

$$\begin{aligned}&\tilde{g}_{1,m}(\acute{\mathbf{p}}) \triangleq g_{1,m}(\acute{\mathbf{p}})^{t-1} + \nabla g_{1,m}(\acute{\mathbf{p}})^{t-1}\Big((\acute{\mathbf{p}}) - (\acute{\mathbf{p}})^{t-1}\Big) \geq g_{1,m}(\acute{\mathbf{p}}), \\ &\tilde{g}_{2,k}(\mathbf{p}) \triangleq g_{2,k}(\mathbf{p})^{t-1} + \nabla g_{2,k}(\mathbf{p})^{t-1}\Big((\mathbf{p}) - (\mathbf{p})^{t-1}\Big) \geq g_{2,k}(\mathbf{p}).\end{aligned} \tag{3.22}$$

Now, we can employ the Dinkelbach algorithm to solve the problem with subtractive form objective function at iteration t as:

$$\begin{aligned}&\max_{\acute{\mathbf{p}},\mathbf{p}} \tilde{R}_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p}) - q_{t_i} P_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p}) \\ &\text{s.t.} : f_{1,m}(\acute{\mathbf{p}},\mathbf{p}) - \tilde{g}_{1,m}(\acute{\mathbf{p}}) \geq R_{\text{min}}^{(\text{D2D})}, \\ &f_{2,k}(\acute{\mathbf{p}},\mathbf{p}) - \tilde{g}_{2,k}(\mathbf{p}) \geq R_{\text{min}}^{(\text{IoT})}, \\ &(3.11\text{f}), (3.11\text{h}), (3.11\text{i}), (3.11\text{j}),\end{aligned} \tag{3.23}$$

where $\tilde{R}_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p}) = \Big[\sum_{m=1}^{M} f_{1,m}(\acute{\mathbf{p}},\mathbf{p}) + \sum_{k=1}^{K} f_{2,k}(\acute{\mathbf{p}},\mathbf{p})\Big]$ $- \Big[\sum_{m=1}^{M} \tilde{g}_{1,m}(\acute{\mathbf{p}}) + \sum_{k=1}^{K} \tilde{g}_{2,k}(\mathbf{p})\Big]$. Besides, $q_{t_i}^{*} = \max_{\mathbf{p},\acute{\mathbf{p}}} \frac{\tilde{R}_{\text{tot}}(\acute{\mathbf{p}},\mathbf{p})^{*}}{P_{\text{tot}}(\mathbf{p},\acute{\mathbf{p}})^{*}}$ is the EE maximization that has to be found via optimization.

Optimal and Feasible Conditions for Time Allocation

To obtain the optimal time allocation for IoT users in problem (3.11), we consider the following lemmas.

Lemma 1: With respect to problem (3.11), the maximum EE can always be achieved when $\tau_i + \tau_h = T$.

Proof: As previously mentioned, $q_{t_i}^*$ is the optimal solution for problem (3.11) providing the maximum EE, i.e., EE_t^*, and satisfies $\tau_i^* + \tau^*{}_h < T$ [32]. Then, we construct a new solution in which $q_{t_i}^* = (\alpha_k^*, p_{m,k}^*, p_k^*, \bar{\tau}_i, \bar{\tau}_h)$, where $\bar{\tau}_i = \frac{\tau_i^* T}{\tau_i^* + \tau_h^*}$, $\bar{\tau}_h = \frac{\tau_h^* T}{\tau_i^* + \tau_h^*}$ according to $\bar{\tau}_i + \bar{\tau}_h = T$. With this new solution, constraints (3.11b) to (3.11h) are still satisfied and $\overline{\text{EE}} = \text{EE}^*$, which means that $\overline{\text{EE}}$ is the new EE. Therefore, the maximum EE can always be achieved at $\tau_i + \tau_h = T$.

To make the mathematics and algorithms easy to understand, Fig. 3.3 shows the flowchart of the problem-solving process.

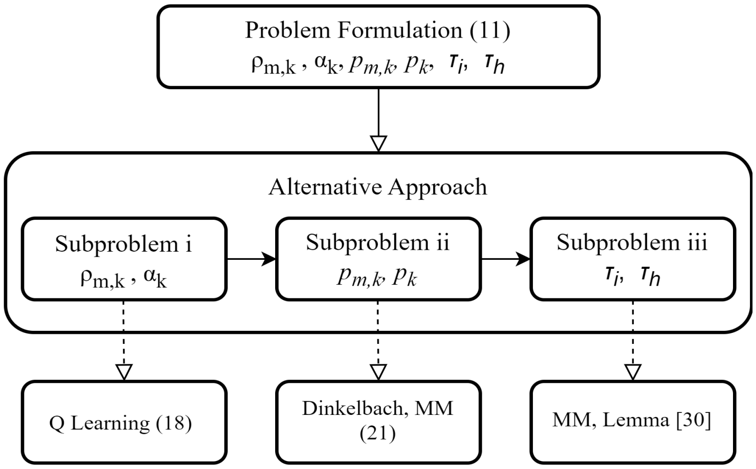


Figure 3.3: Flowchart of the problem-solving process.

3.5 Complexity Analysis

The computational complexity of our proposed approach is determined by both solution approaches adopted in the paper: one is related to the RL method, and the other is related to the MM approach and Dinkelbach algorithm. For the RL algorithm, the computational complexity strongly depends upon the number of users, size, and structure of the state space. It should be noted, however, that all channel gains are engaged in the states, so the number of users has an indirect influence on the RL complexity via the state space. According to the analysis in [33], it requires updating the Q-function of $\forall s_t \in S$ and for each state. Thus, the computational complexity of RL is defined as $\mathcal{O}(|\mathcal{A}|\,|\mathcal{S}|)$. Because we have T time steps in the RL algorithm, i.e., $t \in \{1, 2, \ldots, T\}$, the computational complexity can be simplify expressed as $\mathcal{O}(T^2)$. However, it should be noted that the numbers of D2D pairs, IoT users, and the distance parameters are not considered in the RL complexity analysis. For solving the power control problem based on the MM approach and Dinkelbach algorithm, the complexity for the MM approach is in the order of $\mathcal{O}(MKN)^3$ [27, 34] and the total complexity for the MM approach and Dinkelbach algorithm is $\mathcal{O}(I_{\text{Dinkelbach}} I_{\text{MM}} (MKN)^3)$, where $I_{\text{Dinkelbach}}$ and I_{MM} are the numbers of iterations required for reaching convergence in Dinkelbach and MM methods, respectively.

Table 3.1: Simulation parameters.

Parameter	**Value**
Cell radius R	200 m
Number of D2D pairs, M	$10 \sim 50$
Number of IoT users, K	$10 \sim 50$
D2D transmission distance range, d	$10 \sim 80$ m
Total time duration for IoT users, T	10 s
Max D2D transmission power, $p_{\max}^{(\mathrm{D2D})}$	23 dBm
Max BS transmission power, $p_{\max}^{(\mathrm{BS})}$	26 dBm
Noise power, N_0	-100 dBm
Circuit power consumption of BS, $p_{\mathrm{BS}}^{(\mathrm{circuit})}$	20 dBm
Circuit power consumption of D2D pairs, $p_m^{(\mathrm{circuit})}$	20 dBm
Harvested energy coefficient, $\psi_m \tau_h$	0.8
Amplifier's efficiency factor, ϕ_k	0.35
QoS requirement for D2D users, $R_{\min}^{(\mathrm{D2D})}$	2 bit/s/Hz
QoS requirement for IoT users, $R_{\min}^{(\mathrm{IoT})}$	1 bit/s/Hz
Max harvestable energy of D2D pairs per unit time (i.e., power), $E_{\max}$	-5 dBm
Min harvested energy of IoT users per unit time (i.e., power), $E_{\min}$	-10 dBm

3.6 Numerical Results

In this section, numerical results are provided to evaluate the performance of the proposed joint Q-learning and optimization algorithm. Then, we compared it with the performance of the linear EH model presented in [19], and also two heuristic algorithms, namely constrained and random allocations with considering maximum D2D and IoT transmission power, $p_{\max}^{(\mathrm{D2D})}$ and $p_{\max}^{(\mathrm{BS})}$, respectively. The constrained allocation is based on fixed power across subchannels with optimized scheduling, whereas the random allocation is based on fixed power across subchannels and random scheduling. The parameters used for obtaining the results are summarized in Table 3.1 [18, 19].

Fig. 3.4 shows the EE performance of D2D links versus the number of Dinkelbach iterations for different numbers of D2D pairs and IoT users. With the proposed power allocation, the initial value of EE is a small positive value, but with the increasing number of iterations, q_{t_i} gradually converges to a unique optimum value $q_{t_i}^*$. As a matter of fact, only 4 to 5 iterations are needed to reach the optimum EE values, which depicts the effectiveness of the proposed iterative algorithm in solving the optimization problem at hand. Moreover, depending on the channel condition, the optimal value of EE can vary for different D2D pairs. It can also be seen that increasing the number of D2D pairs and IoT users has a negligible effect on the convergence performance of the proposed algorithm.

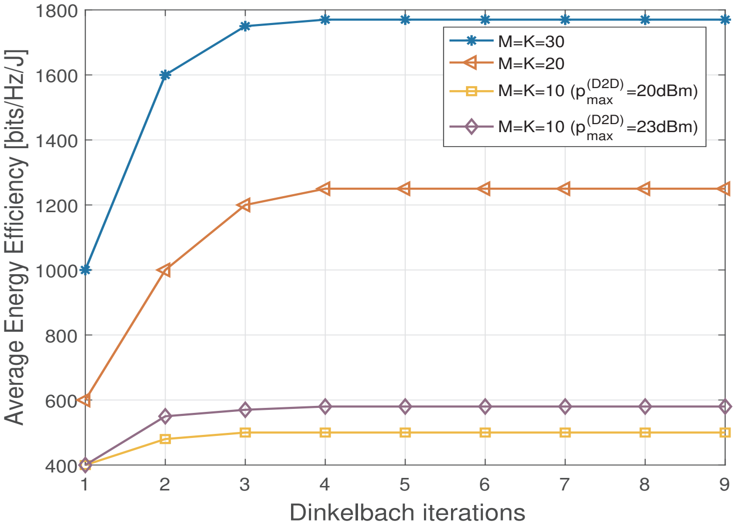


Figure 3.4: Average EE of D2D pairs versus number of Dinkelbach iterations ($d = 50$ m).

Fig. 3.5 illustrates the sum EE versus the number of D2D pairs or IoT users for different algorithms under consideration. As expected, adding more D2D pairs to the system contributes to higher EE. Our proposed algorithm obtains significantly higher EE values as compared to other heuristic algorithms. In particular, there is an increasing EE gap between our proposed RL algorithm and the scheme in [19].

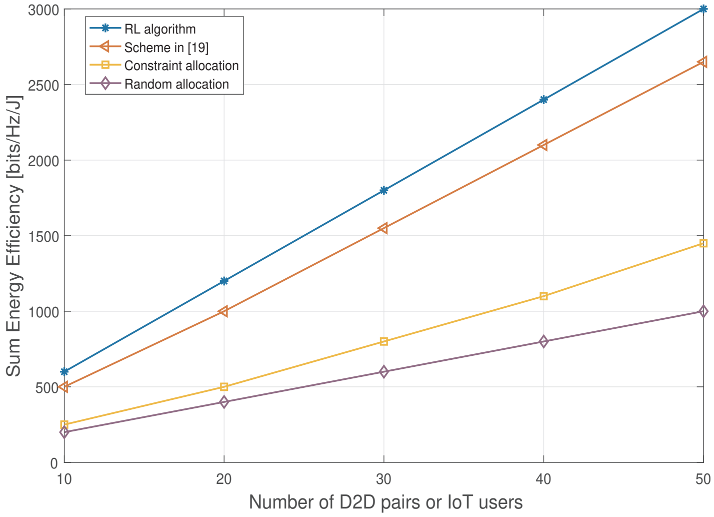


Figure 3.5: Sum EE versus number of D2D pairs or IoT users, $d = 20$ m.

Fig. 3.6 demonstrates the sum EE versus the transmission distance $d = 10 \sim 80$ m with a fixed number of D2D pairs and IoT users, namely $M = K = 30$. Observe that when the transmission distance of D2D pairs increases, the EE performance decreases. The reason for this reduction is that higher transmission power is required for D2D pairs that are separated by a long distance to meet the QoS requirements in (3.11b) and (3.11c). It can be seen that, for the case $d = 10$ m, our proposed algorithm leads to EE improvement when compared to the method in [19], the constrained allocation, and random allocation by about 5.26%, 110.52 %, and 143.90%, respectively.

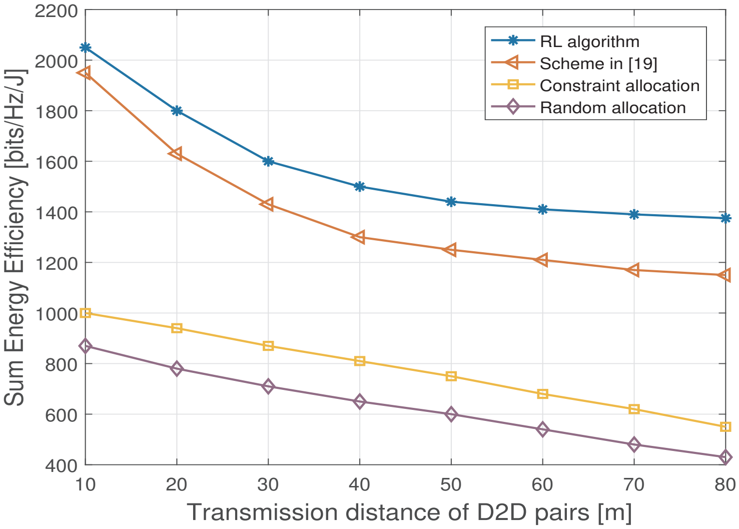


Figure 3.6: Sum EE versus distance of D2D pairs ($M, K = 30$ and $d = 10 \sim 80$ m).

It is pointed out that in terms of EE, the proposed algorithm performs only slightly better than the method in [19] when the transmission distance is short. This is because, with small distances, D2D pairs can find their IoT partners very quickly. But when the transmission distances are large, the number of D2D pairs that fail to match and communicate with their partners increases with the method in [19]. Hence the EE gap between our proposed algorithm and the method in [19] becomes more pronounced. Nevertheless, the EE performance of the method in [19] is still better than that of the constrained and random allocations when $d = 80$ m. Moreover, the EE values of the two heuristic algorithms are approximately constant, which is because the maximum D2D transmission power is previously defined. In general, because of decreasing throughput with increasing distances, the EE reduction suffered by the two heuristic algorithms becomes more severe with increasing distance.

Fig. 3.7 shows the sum EE versus the maximum transmit power in the small scale system model for our proposed algorithm, exhaustive search, constrained allocation, and random allocation. As can be seen, the sum EE first improves with the increase of BS transmission power. This is because a higher BS transmission power can produce more energy for IoT users to harvest and meet their QoS requirement, which results in a higher probability for IoT users to be matched with a D2D pair. It is also obvious that the increase of BS transmission power to around 35 dBm, achieves the maximum EE. Beyond that point, the sum EE of IoT users starts to decline. This is because the EE improvement provided by the larger-than-harvested energy cannot compensate for the loss of EE due to the increased power from the co-channel IoT users. Moreover, the difference between exhaustive search and RL algorithm is negligible. Although exhaustive search achieves better EE, it is an extremely time-consuming technique because in the practical system model with 40-50 users and many subchannels, searching over all possible choices of subchannels/users is a very time-consuming process task and simply impractical.

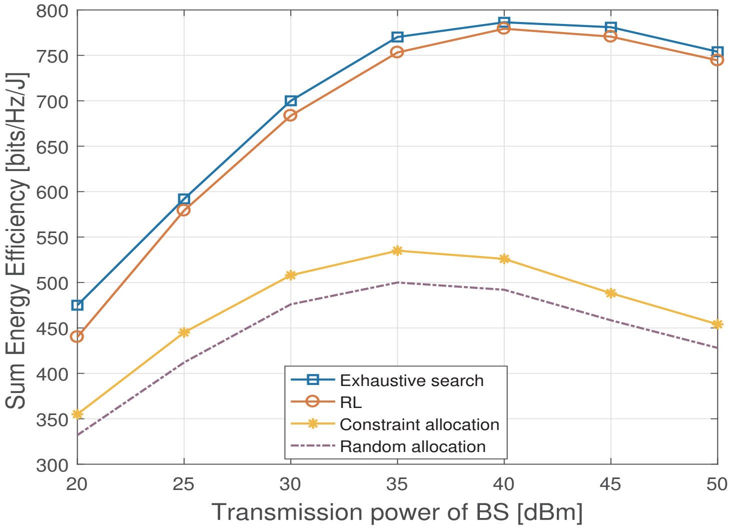


Figure 3.7: Sum EE versus transmit power level of the BS ($M, K = 3$ and $d = 50$ m).

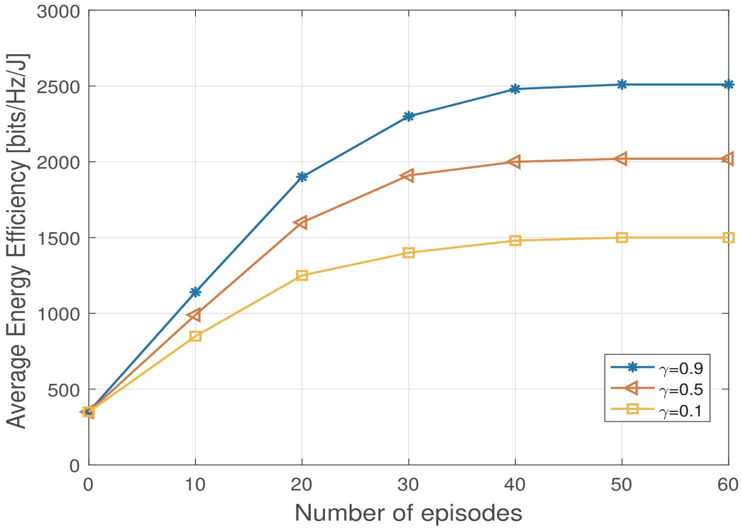


Figure 3.8: Average EE versus number of episodes with different discount factors γ. ($M, K = 40$ and $d = 10$ m)

Fig. 3.8 demonstrates the convergence of the average EE process as the number of episodes increases with different discount factors γ. As can be seen, at the beginning of the learning process (less than 10 episodes) the average EE is approximately the same for all values of γ. This is because the agents (D2D pairs and IoT users) are trained based on the feedback signals received from the interactive environment by trial and error (from their own actions and experiences). Therefore, in the first few episodes, there is very little knowledge about the environment. However, after a sufficient number of episodes, about 40 episodes, the agents acquire adequate knowledge of the network's topology. Therefore, the average EE converges to its optimal value. Moreover, γ also changes the average EE. When γ increases, a larger value of the average EE will be expected. Thus, we can see that the value of γ affects the network lifetime as well as and the average EE value.

Finally, Fig. 3.9 illustrates the overall convergence behavior of the system for different numbers of D2D pairs and IoT users with $M = K = 10$, 20, and 30. In general, the Q-learning process is highly dependent on hyperparameters. Unfortunately, there are no known analytical schemes to set the ideal hyperparameter configuration for a given problem. This is a common problem in the literature concerning the optimality and convergence of Q-learning-based techniques. As such the trial-and-error approach and presenting numerical data is used in related literature [35–38] to study convergence, and it is also used in our paper. As can be seen, several iterations are sufficient for the system to reach convergence. As expected, a larger number of users yields a higher EE.

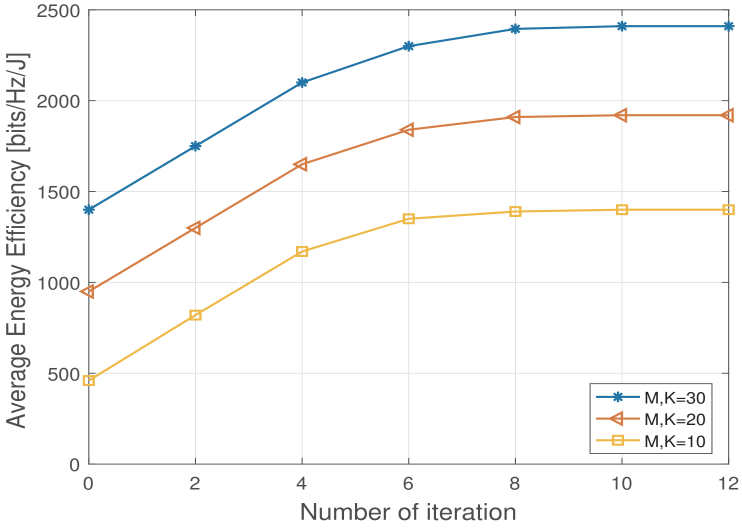


Figure 3.9: Average EE versus number of iterations with different number of D2D and IoT users ($\gamma = 0.5$ and $d = 20$ m)

3.7 Conclusions

In this paper, we have investigated and solved the problem of joint resource allocation and time allocation for D2D communications underlying IoT networks to maximize the total EE of the system. In particular, we considered energy harvesting for D2D and IoT users in which the IoT users harvest energy from the macro BS based on the power splitting technique, while the D2D users harvest energy from ambient resources based on the time switching technique. The considered problem is non-convex MINLP which is challenging to solve. To tackle it, we decompose the original problem into three sub-problems (i) joint subchannel assignment and power splitting, (ii) power control, and (iii) time allocation. The Q-learning method is employed to solve the first sub-problem when the majorization-minimization approach, as well as the Dinkelbach method, are applied to solve the power control sub-problem. Simulation results demonstrate the effectiveness of our proposed algorithm as compared to other methods.

References

[1] Soraya Sinche, Duarte Raposo, Ngombo Armando, André Rodrigues, Fernando Boavida, Vasco Pereira, and Jorge Sá Silva. "A Survey of IoT Management Protocols and Frameworks". In: *IEEE Communications Surveys & Tutorials* 22.2 (2020), pp. 1168–1190.

[2] Arash Asadi, Qing Wang, and Vincenzo Mancuso. "A Survey on Device-to-Device Communication in Cellular Networks". In: *IEEE Communications Surveys & Tutorials* 16.4 (2014), pp. 1801–1819.

[3] Tharindu D. Ponnimbaduge Perera, Dushantha Nalin K. Jayakody, Shree Krishna Sharma, Symeon Chatzinotas, and Jun Li. "Simultaneous Wireless Information and Power Transfer (SWIPT): Recent Advances and Future Challenges". In: *IEEE Communications Surveys & Tutorials* 20.1 (2018), pp. 264–302.

[4] Shruti Gupta, Rong Zhang, and Lajos Hanzo. "Energy Harvesting Aided Device-to-Device Communication Underlaying the Cellular Downlink". In: *IEEE Access* 5 (2017), pp. 7405–7413.

[5] Zhaohui Yang, Wei Xu, Yijin Pan, Cunhua Pan, and Ming Chen. "Energy Efficient Resource Allocation in Machine-to-Machine Communications With Multiple Access and Energy Harvesting for IoT". In: *IEEE Internet of Things Journal* 5.1 (2018), pp. 229–245.

[6] Ying Luo, Peilin Hong, Ruolin Su, and Kaiping Xue. "Resource Allocation for Energy Harvesting-Powered D2D Communication Underlaying Cellular Networks". In: *IEEE Transactions on Vehicular Technology* 66.11 (2017), pp. 10486–10498.

[7] Umber Saleem, Sobia Jangsher, Hassaan Khaliq Qureshi, and Syed Ali Hassan. "Joint Subcarrier and Power Allocation in the Energy-Harvesting-Aided D2D Communication". In: *IEEE Transactions on Industrial Informatics* 14.6 (2018), pp. 2608–2617.

[8] Haichao Wang, Guoru Ding, Jinlong Wang, Le Wang, Theodoros A. Tsiftsis, and Prabhat K. Sharma. "Resource allocation for energy harvesting-powered D2D communications underlaying cellular networks". In: *2017 IEEE International Conference on Communications (ICC)*. 2017, pp. 1–6.

[9] Yue Meng, Zhi Zhang, Yuzhen Huang, and Ping Zhang. "Resource Allocation for Energy Harvesting-Aided Device-to-Device Communications: A Matching Game Approach". In: *IEEE Access* 7 (2019), pp. 175594–175605.

[10] Shuo Yu, Waleed Ejaz, Ling Guan, and Alagan Anpalagan. "Resource Allocation for Energy Harvesting Assisted D2D Communications Underlaying OFDMA Cellular Networks". In: *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. 2017, pp. 1–7.

[11] Zhufang Kuang, Gang Liu, Gongqiang Li, and Xiaoheng Deng. "Energy Efficient Resource Allocation Algorithm in Energy Harvesting-Based D2D Heterogeneous Networks". In: *IEEE Internet of Things Journal* 6.1 (2019), pp. 557–567.

[12] Ying Luo, Peilin Hong, and Ruolin Su. "Energy-Efficient Scheduling and Power Allocation for Energy Harvesting-Based D2D Communication". In: *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 2017, pp. 1–6.

[13] Ke Wang, Wei Heng, Jinming Hu, Xiang Li, and Jing Wu. "Energy-Efficient Resource Allocation for Energy Harvesting-Powered D2D Communications Underlaying Cellular Networks". In: *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*. 2018, pp. 1–5.

[14] Hongwu Liu, Kyeong Jin Kim, Kyung Sup Kwak, and H. Vincent Poor. "Power Splitting-Based SWIPT With Decode-and-Forward Full-Duplex Relaying". In: *IEEE Transactions on Wireless Communications* 15.11 (2016), pp. 7561–7577.

[15] Yanqing Xu, Chao Shen, Zhiguo Ding, Xiaofang Sun, Shi Yan, Gang Zhu, and Zhangdui Zhong. "Joint Beamforming and Power-Splitting Control in Downlink Cooperative SWIPT NOMA Systems". In: *IEEE Transactions on Signal Processing* 65.18 (2017), pp. 4874–4886.

[16] Jie Tang, Arman Shojaeifard, Daniel K. C. So, Kai-Kit Wong, and Nan Zhao. "Energy Efficiency Optimization for CoMP-SWIPT Heterogeneous Networks". In: *IEEE Transactions on Communications* 66.12 (2018), pp. 6368–6383.

[17] Yongjun Xu, Guoquan Li, Yang Yang, Miao Liu, and Guan Gui. "Robust Resource Allocation and Power Splitting in SWIPT Enabled Heterogeneous Networks: A Robust Minimax Approach". In: *IEEE Internet of Things Journal* 6.6 (2019), pp. 10799–10811.

[18] Zhenyu Zhou, Caixia Gao, Chen Xu, Tao Chen, Di Zhang, and Shahid Mumtaz. "Energy-Efficient Stable Matching for Resource Allocation in Energy Harvesting-Based Device-to-Device Communications". In: *IEEE Access* 5 (2017), pp. 15184–15196.

[19] Haohang Yang, Yinghui Ye, Xiaoli Chu, and Mianxiong Dong. "Resource and Power Allocation in SWIPT-Enabled Device-to-Device Communications Based on a Nonlinear Energy Harvesting Model". In: *IEEE Internet of Things Journal* 7.11 (2020), pp. 10813–10825.

[20] Mingzhe Chen, Zhaohui Yang, Walid Saad, Changchuan Yin, H. Vincent Poor, and Shuguang Cui. "A Joint Learning and Communications Framework for Federated Learning Over Wireless Networks". In: *IEEE Transactions on Wireless Communications* 20.1 (2021), pp. 269–283.

[21] Woongsup Lee, Minhoe Kim, and Dong-Ho Cho. "Transmit Power Control Using Deep Neural Network for Underlay Device-to-Device Communication". In: *IEEE Wireless Communications Letters* 8.1 (2019), pp. 141–144.

[22] Woongsup Lee, Minhoe Kim, and Dong-Ho Cho. "Deep Learning Based Transmit Power Control in Underlaid Device-to-Device Communication". In: *IEEE Systems Journal* 13.3 (2019), pp. 2551–2554.

[23] Kisong Lee, Jun-Pyo Hong, Hyowoon Seo, and Wan Choi. “Learning-Based Resource Management in Device-to-Device Communications With Energy Harvesting Requirements”. In: *IEEE Transactions on Communications* 68.1 (2020), pp. 402–413.

[24] Helin Yang, Wen-De Zhong, Chen Chen, Arokiaswami Alphones, and Xianzhong Xie. “Deep-Reinforcement-Learning-Based Energy-Efficient Resource Management for Social and Cognitive Internet of Things”. In: *IEEE Internet of Things Journal* 7.6 (2020), pp. 5677–5689.

[25] Ata Khalili, Ehsan Mohammadi Monfared, Shayan Zargari, Mohammad Reza Javan, Nader Mokari Yamchi, and Eduard Axel Jorswieck. “Resource Management for Transmit Power Minimization in UAV-Assisted RIS HetNets Supported by Dual Connectivity”. In: *IEEE Transactions on Wireless Communications* 21.3 (2022), pp. 1806–1822.

[26] Arooj Mubashara Siddiqui, Leila Musavian, and Qiang Ni. “Energy efficiency optimization with energy harvesting using harvest-use approach”. In: *2015 IEEE International Conference on Communication Workshop (ICCW)*. 2015, pp. 1982–1987.

[27] Ata Khalili, Mohammad Robat Mili, Mehdi Rasti, Saeedeh Parsaeefard, and Derrick Wing Kwan Ng. “Antenna Selection Strategy for Energy Efficiency Maximization in Uplink OFDMA Networks: A Multi-Objective Approach”. In: *IEEE Transactions on Wireless Communications* 19.1 (2020), pp. 595–609.

[28] Parag Kulkarni. “Introduction to Reinforcement and Systemic Machine Learning”. In: *Reinforcement and Systemic Machine Learning for Decision Making*. Wiley-IEEE Press, 2012, pp. 1–21.

[29] Werner Dinkelbach. “On nonlinear fractional programming”. In: *Management science* 13.7 (1967), pp. 492–498.

[30] H.H. Kha, H. D. Tuan, and Ha H. Nguyen. “Fast Global Optimal Power Allocation in Wireless Networks by Local D.C. Programming”. In: *IEEE Transactions on Wireless Communications* 11.2 (2012), pp. 510–515.

[31] Ata Khalili, Mohammad Robat Mili, and Derrick Wing Kwan Ng. “Performance Trade-off Between Uplink and Downlink in Full-Duplex Communications”. In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020, pp. 1–6.

[32] Lu Pei, Zhaohui Yang, Cunhua Pan, Wenhuan Huang, Ming Chen, Maged Elkashlan, and Arumugam Nallanathan. “Energy-Efficient D2D Communications Underlaying NOMA-Based Networks With Energy Harvesting”. In: *IEEE Communications Letters* 22.5 (2018), pp. 914–917.

[33] Meisam Maleki, Vesal Hakami, and Mehdi Dehghan. “A model-based reinforcement learning algorithm for routing in energy harvesting mobile ad-hoc networks”. In: *Wireless Personal Communications* 95.3 (2017), pp. 3119–3139.

[34] Ata Khalili, Shayan Zargari, Qingqing Wu, Derrick Wing Kwan Ng, and Rui Zhang. “Multi-Objective Resource Allocation for IRS-Aided SWIPT”. In: *IEEE Wireless Communications Letters* 10.6 (2021), pp. 1324–1328.

[35] Ursula Challita, Li Dong, and Walid Saad. “Proactive Resource Management for LTE in Unlicensed Spectrum: A Deep Learning Perspective”. In: *IEEE Transactions on Wireless Communications* 17.7 (2018), pp. 4674–4689.

[36] Nan Zhao, Ying-Chang Liang, Dusit Niyato, Yiyang Pei, Minghu Wu, and Yunhao Jiang. “Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Cellular Networks”. In: *IEEE Transactions on Wireless Communications* 18.11 (2019), pp. 5141–5152.

[37] Ying He, F. Richard Yu, Nan Zhao, Victor C. M. Leung, and Hongxi Yin. “Software-Defined Networks with Mobile Edge Computing and Caching for Smart Cities: A Big Data Deep Reinforcement Learning Approach”. In: *IEEE Communications Magazine* 55.12 (2017), pp. 31–37.

[38] Shayan Zargari, Ata Khalili, and Rui Zhang. “Energy Efficiency Maximization via Joint Active and Passive Beamforming Design for Multiuser MISO IRS-Aided SWIPT”. In: *IEEE Wireless Communications Letters* 10.3 (2021), pp. 557–561.

4 Discussions and Conclusion

4.1 Summary

The growth of devices and connections around the world is outpacing the growth of both the population and the number of Internet users. New devices are introduced in the market every year with increased capabilities and intelligence. Devices and connections are increasing as a result of a growing number of M2M applications, including smart meters, video surveillance, healthcare monitoring, transportation, and package or asset tracking.

A growing number of IoT and M2M devices will result in a significant increase in data traffic. Additionally, as social media and mobile phones became increasingly popular, people became aware of the amount of data generated daily. There are around a million terabytes of data generated per day by each of these billions of social media users. Data traffic is expected to grow significantly as a result of all of these factors.

In order to reduce the amount of data traffic in the core networks and improve user data rate, Device-to-Device (D2D) communication in cellular networks was introduced as direct communication between two mobile users without traversing the Base Station (BS) or core network. Hence, D2D communications in such scenarios can greatly increase the spectral efficiency of the network. The advantages of D2D communications go beyond spectral efficiency; they can potentially improve throughput, energy efficiency, delay, and fairness.

However, D2D and IoT communications face energy consumption constraints due to limitations in the energy storage capacities of the devices' batteries. It is possible to prolong the lifespan of energy-constrained wireless devices by harvesting energy from external sources, such as solar, wind, vibrations, thermoelectric, and radio frequency (RF). A promising approach to EH is simultaneous wireless information and power transfer (SWIPT). A fundamental trade-off has been revealed between power and information in early information theory studies on SWIPT. Through power splitting or time switching schemes, this technique harvests energy and processes information. Receiver antennas change periodically between ID circuits and EH circuits based on TS sequences. Moreover, TS receivers need accurate scheduling and synchronization of energy and information. On the other hand in the PS approach, after the signal is received by the PS receiver, it is divided into two power streams with different power levels and a certain PS ratio. For simultaneous ID and EH, both power streams go through a decoder and an energy harvester.

Since there are more wireless devices competing for limited resources, such as time and frequency bands, modern wireless networks need resource allocation techniques. In D2D and IoT communications, the key objective of resource allocation is to balance performance and energy consumption in order to ensure an adequate supply of energy, maintaining reliable and efficient communications between devices. For D2D networks with numerous aspects to be managed together, designing a joint resource management design is typically not easy and often demands significant computing power.

In order to solve the above problem, we proposed a novel approach, which involves decomposing the original EE optimization problem into three subproblems: Firstly, the subchannel assignment problem and PS factor, secondly, the power allocation problem and finally, the time scheduling problem. The Q-learning technique is used to solve the first subproblem in which discrete variables are associated with it. The size of the problem becomes very large when optimizing all variables using the learning technique, so it is not practical to optimize all variables using the learning technique. In addition, minibatch has a limited size and cannot be used for a large number of variables due to this restriction. Meanwhile, the RL method is highly appropriate for discrete variables, however, it may also be utilized to deal with continuous variables, such as the allocation of power. Although the power allocation vectors can be discretized and solved using RL algorithms, there is a possibility that the outcome will not be as accurate as the original solution.

To solve the continuous variables, i.e., power allocation and time allocation, we applied the majorization–minimization (MM) and Dinkelbach methods and converted the original nonconvex problem into a convex one. Then, we could come up with the locally optimal solution.

The effectiveness of our suggested RL-based resource allocation approach was demonstrated through simulation. We considered different parameters such as distance, number of D2D pairs and IoT users, number of Dinkelbach iterations, transmit power level of BS, and number of episodes. We demonstrate that our proposed solution is superior in terms of the average and sum EE based on the parameters mentioned above. In particular, our proposed algorithm improves EE by about 5.26%, 110.52%, and 143.90% for distance $d = 10$ m in comparison to a recent method, constrained allocation, and random allocation.

4.2 Suggestions for Future Studies

- In order to solve the time allocation problem, we can consider the use of neural networks or deep learning methods. It has been established that an algorithm called Mini Batch Stochastic Gradient Descent (MB-SGD) can help to overcome the large time complexity associated with convex optimization problems. It should be noted that after converting our non-convex problem into a convex one using MM and Dinkelbach, the MB-SGD adjusts its parameters iteratively in order to optimize time allocation function.

- In order to reduce interference, enhanced spectrum efficiency, and reduced latency with high reliability in the proposed system model, we can consider Non-orthogonal multiple access (NOMA) channels. In our system model, we considered identically Rayleigh fading channel. The working principle behind NOMA is Power Domain Multiplexing, which means access to the channel is shared by using different Power levels allocated to the individual users. The use of NOMA has been shown to improve spectrum efficiency in certain scenarios. Such a study should also evaluate the effect of delay in optimizing EE.

- As discussed previously, D2D communication is a kind of short range communication service. Therefore the effective coverage range of D2D communication is very limited. When two D2D users are communicating with each other, if the channel conditions between them is deeply faded (e.g., due to building occlusion) or the distance between them increases instantaneously (e.g., between the high-speed vehicles and stationary vehicles), the D2D communication will be interrupted. Therefore, the multi-hop D2D communication is an efficient method to further boost transmission performance where the relay nodes can assist D2D users. As a result, we can consider relay-assisted D2D network to optimize achievable data rates of users at the edge of the network, and network coverage.

- As our users are resource-limited devices (such as IoT users), we should evaluate the data traffic in our network. It is known that IoT devices have restricted storage capacity. Thus, if they have a large amount of data for processing, it takes a long time to complete. As a result, the network will experience long delays. Additionally, as IoT devices are battery-limited, in practice is not possible to process large amounts of stored data. Task offloading techniques have been developed to minimize power consumption while data are processed. In these techniques, in order to obtain a reduction in latency or energy consumption and provide a better user experience, part of data will be transferred and offloaded to the edge servers for processing. Typically, an initial and critical part of this process is deciding whether to offload, i.e. the offload decision. After determining whether to offload, the next question to consider is how much and what should be offloaded. It is a promising idea to consider a system model with one or multiple edge servers for processing data.