QATAR UNIVERSITY

COLLEGE OF ENGINEERING

DEEP REINFORCEMENT LEARNING FOR EFFICIENT UPLINK NOMA SWIPT TRANSMISSIONS

BY

MOHAMED ABDELHAMID MOHAMED ELSAYED

A Thesis Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Master of Science in Computing

January 2021

COMMITTEE PAGE

The members of the Committee approve the Thesis of

Mohamed Abdelhamid Mohamed Elsayed defended on 01/11/2020.

Dr. Amr Mohamed
Thesis Supervisor

Dr. Ahmed Badawy
Thesis Co-Supervisor

Dr. Mohamed Abdallah
Committee Member

Dr. Mohsen Guizani
Committee Member

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

Name, Mohamed Abdelhamid Mohamed Elsayed, Masters : January: 2021, Master of Science in Computing

Title: Deep Reinforcement Learning for Efficient Uplink NOMA SWIPT Transmissions

Supervisor of Thesis: Dr. Amr Mohamed.

A key rival technology in radio access strategies for next generation cellular communications is non-orthogonal multiple access (NOMA) due to its enhanced performance compared to existing multiple access techniques such as orthogonal frequency division multiple access (OFDMA). The work in this thesis proposes a framework for an energy efficient system geared towards wireless exchange of intensive data collected from distributed Internet of things (IoT) sensor nodes connected to an edge node acting as a cluster head (CH). The IoT nodes utilize an adaptive compression model as an extra degree of freedom to control the transmitted rate going to the CH. The CH is an energy constrained node and may be battery operated. The CH is capable of radio frequency (RF) energy harvesting (EH) using simultaneous wireless power transfer (SWIPT). The proposed framework exploits deep reinforcement learning (DRL) mechanisms to achieve smart and efficient energy constrained up-link NOMA transmissions in IoT applications requiring data compression. In particular, the DRL maximizes the harvested energy at the CH while enforcing the data compression ratio constraints at the transmitting nodes and satisfying the outage probability constraints at the CH. The data compression in this type of sensor networks is vital in order to minimize the power consumption of the different sensors (transmitting nodes), which increases its service lifetime.

DEDICATION

To my family, for their support and faith in me.

ACKNOWLEDGMENTS

This work was made possible by NPRP grant NPRP12S-0305-190231 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

It is my pleasure to express my deep thanks to my Master supervisors Dr. Amr Mohamed, and Dr. Ahmed Badawy for their continuous advise during my Masters studies. I am really grateful for their constant guidance and valuable feedback which I would not be able to complete the Thesis without. Special thanks to my Master referees committee for their valuable time and effort in assessing my work. I would like to thank my wife, Rehab, and all of my family members, for their continuous encouragement, patience and unwavering love during my study. They always motivating and inspiriting me to progress with my studies.

Most of all, I would like to thank ALLAH, for the blessings and sound belief in Him, health and sanity, and for putting me in the right path that allowed me to meet people that have been so kind to me and giving me the opportunity to reciprocate.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

[1] Mohamed Elsayed, Massudi Mahmuddin, Ahmed Badawy, Tarek Elfouly, Amr Mohamed, and Khalid Abualsaud, "Walsh transform with moving average filtering for data compression in wireless sensor networks." 2017 IEEE 13th International Colloquium on Signal Processing and its Applications (CSPA), 2017.

[2] Mohamed Elsayed, Ahmed Badawy, Massudi Mahmuddin, Tarek Elfouly, Amr Mohamed, and Khalid Abualsaud, "FPGA implementation of DWT EEG data compression for wireless body sensor networks." 2016 IEEE Conference on Wireless Sensors (ICWiSE), 2016.

[3] Mohamed Elsayed, Ahmed Badawy, Ahmed El Shafie, Amr Mohamed, and Tamer Khattab, "Deep Reinforcement Learning Algorithm for Smart Data Compression under NOMA-Uplink Protocol" 2020 IEEE Conference on Electrical and computer Engineering.

[4] Mohamed Elsayed, Ahmed Badawy, Ahmed El Shafie, Amr Mohamed, and Tamer Khattab, "A Framework for Data Compression in Uplink NOMA SWIPT Systems using Reinforcement learning" 2020 Under submission in IEEE Transactions on Communications.

CHAPTER 1: INTRODUCTION

Motivation

Maintaining the viability of mobile communications networks over the next era promotes new technology solutions that need to be architected and developed so that it can adapt to future challenges [1]. Introducing and designing such capability of radio access technology in wireless mobile communications is an essential aspect in terms of cost efficiency and system reliability. The multiple access approach is a vital part of the radio access technology. Currently, orthogonal multiple access (OMA) based on orthogonal frequency division multiple access (OFDMA) is approved for the 3.9th and 4th generation (4G) mobile communication systems, including LTE and LTE Advanced [2]. OMA is suitable for the field of packet domain services to achieve good system performance by time and frequency domain scheduling utilizing channel-awareness with fast single user identification at the receiver. However, more improvements to system efficiency and quality of service (QoS) are needed in the future, particularly at the cell edge. NOMA uses a new approach that exploits the power-domain for user multiplexing, which has not been widely used in previous generations. In NOMA, many users are multiplexed on the power domain at the transmitter side, while successive interference cancellation (SIC) is used for DE-multiplexing of the received signal at the receiver side [3].

Wireless network clustering is widely used for dynamic networking in emerging communication environments. Clusters are self-established and self-maintained communication networks used mainly when the base station is far away from the communicating nodes. In cluster communication networks, the central node called the cluster

head (CH) receives the signal from the base station and transmits it to other nodes in the cluster (e.g. in the case internet of things (IoT) sensor nodes). The CH is also used to receive data from the IoT devices and sends the aggregate signal to a base station [4]. Cluster networks are also used to establish communication in rescue operations, remote administration, relief work, and emergency maintenance. A key application of using cluster communication systems is wireless remote healthcare monitoring systems [5]. Wireless networks used for monitoring patients who need continuous surveillance can utilize clustering techniques.

An important example of wireless remote health monitoring is patients with epilepsy who need continuous electroencephalogram (EEG) monitoring. EEG sensing nodes collect intensive data from the patient and need to transmit this data to the CH. Data compression for such sensor networks is vital in order to minimize the power every sensor consumes to maximize its service lifetime and optimize the power consumption. Meanwhile, keeping minimum level of data distortion is vital for this kind of signal due to its importance for medical diagnosis whether by the medical experts or using an autonomous classification system.

In order to investigate the whole system under practical conditions, channel impairments and energy constraints have to be taken into consideration. Fading channel realizations coupled with users' quality of service (QoS) requirements result in outage probability constraints for NOMA up-link which need to be considered [6]. Moreover, battery operated CH units having limited access to power sources and using energy harvesting (EH) mechanisms are key system features to take into consideration [7]. Optimizing the performance of such practical systems is not easy, thus, using state-of-the-art Deep Reinforcement Learning (DRL) mechanism to find this trade-off between

the system parameters will be a viable approach enabling valuable system efficiency enhancements.

Problem Statement

In this work, we focus on building a framework for establishing energy efficient smart data compression and energy harvesting under NOMA up-link protocol. The framework involves multiple NOMA users connected to an edge node (cluster head node) to transmit vital data under NOMA up-link protocol. This system design seeks a trade-off between the involved system parameters to maximize energy harvesting at the cluster node, while maintaining the main system constraints on QoS represented by outage probability and compression ratio represented by data distortion.

Deliverables

1. A proposed framework for providing an energy-efficient system to compress and transmit data, e.g. medical information such as electroencephalogram (EEG) collected from patients suffering from certain chronic diseases for the purpose of continuous monitoring, to an edge node using NOMA up-link techniques.

2. Incorporation of energy harvesting at the edge node mentioned above to cover its energy requirements.

3. A deep reinforcement learning (DRL) algorithm to find the optimal trade-off between energy harvesting and data compression under NOMA up-link protocol.

4. A simulation model of the system to provide benchmarking to check the optimal solution delivered by the proposed DRL approach.

Methodology

The research approach will rely on mathematical modeling and analysis of the data compression, outage probability, energy harvesting, and node power consumption under a realization of the NOMA channel to extract a closed form mathematical model for each. The deduced mathematical models will be utilized to formulate an optimization problem that represents the energy harvesting optimization objective in terms of the different parameters and taking into account the different required constraints. The optimization problem is solved using heuristic techniques. Additionally, a DRL mechanism is constructed and exploited to provide an optimal solution of the optimization problem. The research plan can be divided into work packages (WPs) as follows:

WP1: Literature survey on the main system pillars including NOMA, data compression, energy harvesting and DRL.

WP2: Comprehensive mathematical modeling and analysis of the above pillars deriving clear model for each of them to facilitate the formulation of an optimization framework.

WP3: Design and simulation of algorithms which portrays and explores the trade-off between these pillars.

WP4: Formulate an optimization model for the entire NOMA system.

WP5: Propose novel and efficient heuristic and DRL-based solutions to address the trade-off between energy harvesting and performance constraints such as outage probability and distortion.

WP6: Conducting comparative studies between the proposed techniques and the state-of-the-art to analyze the pros and cons of each solution.

WP7: Thesis write up and publications.

Thesis Objectives and Contribution

In this thesis, we focus on three major Objectives:

1. We aim at developing a mathematical analysis to find closed form expressions for adaptive data compression, outage probability, energy harvesting, and node power consumption under NOMA up-link protocol. These closed form expressions can be used to formulate an optimization problem to describe the whole system. By solving this optimization problem, we can find out the trade-offs between data compression, outage probability, energy harvesting and node power consumption. This objective is tailored towards the problem mentioned in the motivation section about finding the optimal parameters that maximize the harvested energy, while keeping reasonable system performance under the communication constraints.

2. Develop a set of algorithms to realize the solution for the optimization framework in order to demonstrate the relation between various parameters of the system. These algorithms will tackle the problem of realizing such system under practical channel, and measure the system performance under the different constraints.

3. The last objective is to realize the whole system practically by designing a comprehensive DRL method to figure out the trade-off between the system parameters. A final comparison between the different solutions will be conducted.

Therefore, the contributions can be summarised as follows:

1. Propose a comprehensive mathematical analysis to find a closed form expressions of cluster node harvested energy, outage probability, and power consumption of

the sensor node with respect to the data compression and distortion ratios.

2. Design a DRL agent to achieve the trade-off between the system parameters in real-time while satisfying design requirements.

3. Evaluate the proposed system against multiple currently followed heuristics approaches with results to show the performance of these techniques and the DRL performance.

Thesis Overview

The remainder of this thesis is organized as follows: Chapter 2 describes the main concepts, terminologies, and state-of-the-art frameworks. We also evaluate and contrast our work to others' related work. Chapter 3 provides a mathematical framework with simulation and proposes the DRL algorithm. In Chapter 4, we introduce the concept of energy harvesting to the system. Chapter 5 concludes and summarize the work, review key results, and suggests potential future improvements.

CHAPTER 2: BACKGROUND AND RELATED WORK

Background

The proliferation of broadband multimedia applications, such as machine-type networking, video, mobile gaming, HDTV, 3D TV,VoIP, and the booming of a wide range of wireless sensor networks including the Internet of Things (IoT) have motivated the evolution towards 5th generation (5G) networks [8]. Due to this extraordinary technological impact and the current limitations of wireless resources, reliable wireless communication networks need to be developed to fill this crucial performance gap. Historically, wireless communication networks employed several multiple access techniques. First-generation (1G) wireless network employed Frequency Division Multiple Access (FDMA), 2G employed Time Division Multiple Access, 3G employed Code Division Multiple Access, and 4G employed Orthogonal Frequency Division Multiple Access (OFDMA). In other words, today's wireless networks allocate radio resources to users based on Orthogonal Multiple Access (OMA) principles [9].

Non-Orthogonal Multi-Access (NOMA) has recently received massive attention as a promising solution for spectral efficiency, user fairness, better connectivity, enhanced data rate, and reduced latency in 5G networks [10]. The fundamental concept behind NOMA is the suitability of multiple access (MA) for 5G networks to maximize the spectral efficiency through allowing nodes to transmit simultaneously with no constraints on orthogonality of the frequency sub-carriers. The essential reason for embracing NOMA in 5G is its capacity to serve numerous clients utilizing the same time and frequency resource. There exists two main approaches to achieve NOMA; namely code domain and power domain [11]. The work in [11] focuses on power-domain NOMA, which

from this time forward, is referred to as NOMA. NOMA exploits superposition at the transmitter, and successive interference cancellation (SIC) at the receiver.

NOMA accomplishes superior efficiency in terms of utilizing the spectrum by serving multiple users at the same time, which means increasing the connectivity. A vital point in the NOMA approach is that users with better channel conditions have knowledge about the messages of other weaker users as it can decode their messages before applying SIC. Relying on this concept, the strong users can work as relays to improve the performance of the weaker users; this approach is called cooperative NOMA

Typical NOMA DOWN LINK

In down-link NOMA, the base station (BS) or cluster head sends a superimposed signal to all NOMA users, allocating more power to the far and weaker users due to the worst channel conditions they have and less power to the near user with better channel conditions. The near (strong) user first subtracts the signal of the far (weak) user through SIC, and then decodes its signal. The weak user considers the signal of the strong user as noise and detects its signal directly [12]. Without loss of generality, and as an example, Fig. 2.1 describes a typical NOMA down-link scenario for two users.The conventional NOMA down-link uses a power allocation system, where high transmission power allocated for users with bad channel conditions and vice versa. The BS transmits a superimposed signal x to all the NOMA users, utilizing the entire system bandwidth. The received signal at user U_i can be described as $y_i = H_i x + n_i$ where, the superimposed signal x can be described as $x = \sum_{i=1}^{k} \sqrt{p_i} x_i$, x_i is the message signal for user i, p_i represents the power allocated for this user, k is the total number of users in the NOMA cluster, and n_i is the AWGN of user U_i [13]. Thus, for a given user

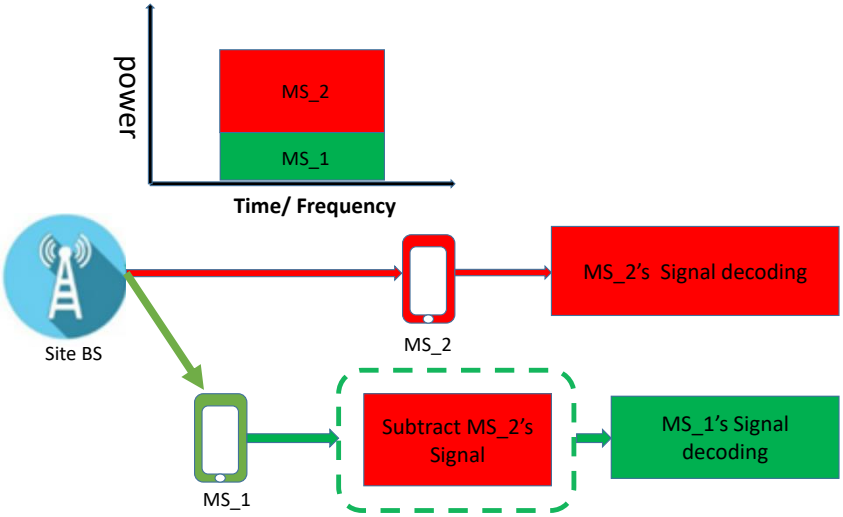

Figure 2.1: Downlink NOMA.

in the NOMA cluster, heavy interfering signals are mainly due to high power message signals from relatively weaker channel users. Meanwhile, every user cancels the strong interference by encoding, re-modulating and deduction of the signal obtained x, and then cancels any intra-cell interference by the other users, whereas, the lowest channel gain user gets the interference from all users within the same cell. Moreover, the strong user has a better channel gain, but that doesn't mean that the signal quality is higher. A strong user is generally assigned a lower transmit power, and a weak user is assigned more power. So the signal of the weak user is the strongest one. Thus, NOMA does not contradict the basic principle of SIC, the first decoding of the strongest signal should be done.

Typical NOMA UP-LINK

In an up-link NOMA system, the BS receives a superimposed signal from mobile users. SIC is deployed at the users (receiving nodes (BS)) [14]. In up-link NOMA,

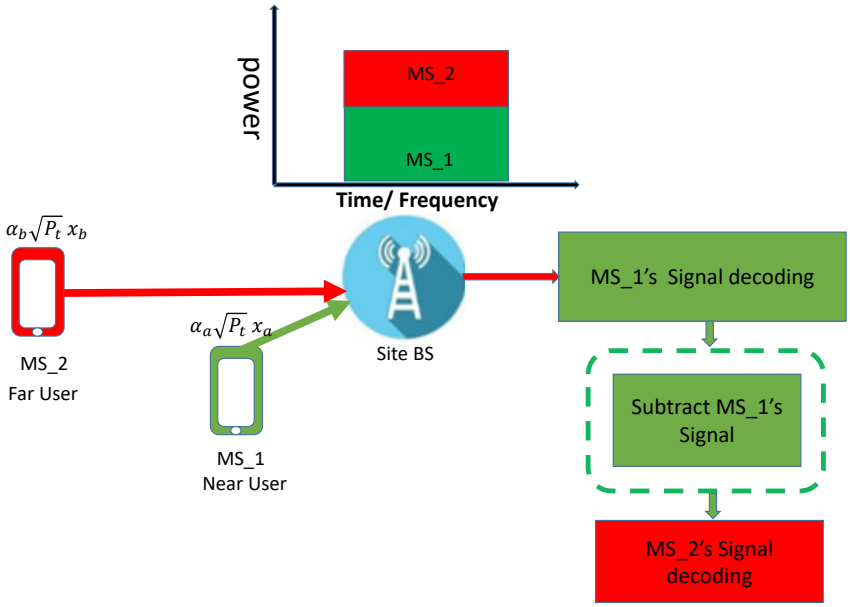

Figure 2.2: Up-Link NOMA.

users are typically transmitting with the same transmission power. Since the channels of various users in the up-link is unique, each signal exhibits different channel gains. As a consequence, the signal power obtained that corresponds to the user with the highest channel is the strongest at the BS. Therefore, this signal is the first one to be decoded at the BS, while treating all the other users in the cluster with relatively weaker channels as interference. The decoded signal is then subtracted from the superimposed received signal and the process repeats among the remaining users in the cluster. Hence, the transmission of the lowest channel gain user will be decoded with almost zero interference. Considering two-user NOMA scenario, the received signal at the BS can be described as $y = \beta_a\sqrt{P_t}h_a x_a + \beta_b\sqrt{P_t}h_b x_b + n$, where P_t is the available transmission power, β_a and β_b are the power split factors for the two users, x_a and x_b are the transmitted signals while n the AWGN of the BS [12]. Figure 2.2 presents a two user up-link scenario.

Data compression

Data compression is the process of altering, encoding, or transforming the data sample structure in such a way that it requires less disk space and less encoding and transmission capacity and consumed power. Such process is motivated by the single-modality and power limitation typically existing in today's IoT devices, making such devices optimized for special type of applications e.g. medical applications. The process of signal reconstruction is done at the receiver side to retrieve the transmitted signal. This data retrieval may be associated with distortion percentage for the case of lossy compression, based on the compression ratio [15]. Many articles have tackled the importance of this aspect from different perspectives; an example is shown in [16].

In IoT applications, data compression shall play a significant role in its efficient design and implementation. As stated earlier, it is very likely that IoT nodes are battery-operated and hence by exploiting an efficient and optimized data compression scheme at the transmitting IoT nodes, battery life will be extended and the frequency of battery replacement will be reduced. Examples of data compression in video streaming for wireless sensor network is presented in [17] and for electroencephalogram (EEG) monitoring for wireless body sensor networks is presented in [16].

Another example is the Wireless Body Sensor Networks (WBSN) that offer essential support to patients who require continuous care and monitoring. One of the critical applications of WBSN is electroencephalogram (EEG) monitoring devices, which can continuously monitor and record vital patient signs. EEG monitoring devices are likely to be portable due to the specificity of the application. EEG data compression is a key to minimize the transmission power and, jointly, increase battery life [18]. Long

monitoring time, a large number of electrodes and a high sampling rate together produce high Electroencephalography (EEG) data size. There is, therefore, a need for more bandwidth and storage space for efficient data transmission. EEG data compression is, hence a fundamental issue to efficiently transmit EEG data with less bandwidth and to store it in less space.

Energy Harvesting

Energy harvesting from wireless signals could redefine mobile connectivity. Recently, researchers have developed a novel way that can convert energy from wireless signals into electricity [19]. Radio frequency (RF) energy is currently being transmitted by a huge number of worldwide radio transmitters, including mobile telephones, portable radios, cellular base stations, and TV/radio stations. With this ubiquitous supply of RF energy, charging devices for RF harvesting is a viable approach. This allows battery-based devices to be charged to reduce new batteries replacement and/or extend the operating life of systems with removable batteries. Battery-free systems may also be designed to use storage capacitors continuously charged instead of batteries [20]. The ability to maintain RF-to-DC performance over a variety of operating conditions, including fluctuations in input and resistance to output loads, is an essential element of RF energy harvesters.

Simultaneous wireless information and power transfer (SWIPT), allows for the transmission of information and energy on the same RF signal at the same time [21]. There are two main categories for SWIPT; power splitting (PS) and time switching (TS). Under PS SWIPT, the received RF signal is divided using a power divider between signal decoding and harvesting energy. Under TS SWIPT, the receiver keeps communication

scheme keeps alternating between power and data transmission. PS SWIPT exhibits a better performance than TS SWIPT in terms of achieved data rate and harvested energy.

Outage Probability

Outage probability is the probability that the receiver power value is below the threshold where the power value is related to the minimal cluster signal - to - noise ratio (SNR). In this case, there is a high probability that an outage will occur at this time period. It is an indicator of the quality of the communication channel, which means, finding the possibility that a specific transmission rate is not supported because of variable channel capacity [12]. the outage probability is a vital metric in order to measures the performance of a system. it measure the capability of the NOMA system to satisfy the the user QoS constraint.

Deep Reinforcement Learning

Artificial intelligence (AI) can be considered as the intelligent software that can solve problems and make decisions by itself. As a branch of AI, Machine Learning (ML) algorithms and techniques can learn from data to take decisions or to classify an input based on some feature understanding. Reinforcement learning is a category of machine learning suitable for optimal control and decision-making process. It is different from supervised and unsupervised learning techniques since there is no input/output mapping like supervised learning. In addition, there is no hidden pattern recognition, as exists in unsupervised learning. In reinforcement learning approaches, we usually deal with a dynamic environment, in which an agent learns how to interact with this environment by taking a sequence of actions (which we call a policy) to maximize a global reward over

time. Therefore, the main components of the reinforcement learning system include an environment, which defines the context of the problem. It is the territory where the agent explores and makes actions. The other component is the agent itself, which has a sequence of actions for the environment. The main point is the agent does not know anything about the internal dynamics of the environment. Nevertheless, the agent tries to figure out how the environment works by doing random actions and observing how the environment responds to these actions. This response comes to the agent in two separate pieces of information, as the agent can observe the change of the state of the environment as well as a reward or punishment signal. The main idea behind reinforcement learning is imitating human learning mechanisms. The training in reinforcement learning is an interactive process, in which, each interaction with the environment trains the agent, and this process can be realized in a dynamic environment where the state of the system may change by each interaction. This may contradict the concept of other supervised and unsupervised learning, as the training is isolated from the real environment. Reinforcement learning differs from supervised learning in the needless to present labeled input / output pairs and the absence of the explicit need to correct sub-optimal actions. In other words, the emphasis is on finding a balance between exploration (of unknown territory) and exploitation of current knowledge .[22].

The actions of the agent decide not only its current (immediate) reward, but also (at least probabilistic) the next state of the environment. When determining the action to take, the agent must take the next state into account as well as the immediate reward. The model of long-run optimally used by the agent determines exactly how the value of the future should be taken into account. Moreover, the agent has to be able to learn from the delayed reward and must be able to learn which of its behaviors are desirable. It may take

a long sequence of actions, may receive insignificant punishment, then finally attains the best policy that allows it to take the best action based on the system state. Therefore, reinforcement learning can usually be regarded as 5 main tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, where $\mathcal{S}$ is the state's space, $\mathcal{A}$ is the actions space, $\mathcal{T}$ is the state transition values as a function of given action a in a given state, $\mathcal{R}$ is the reward and γ is the discount factor that adjust the action.

Related Work

Uplink NOMA has been considered in several research work recently. The authors in [23], [24] investigated the scenario of two uplink NOMA users under statistical quality of service (QoS) delay constraints extracted based on user's effective capacity only. They compared the performance of NOMA users with OMA users. However, the system structure does not include neither energy harvesting nor data compression. An iterative analysis of multi-cell NOMA devices with imperfect SIC has been shown in [25]. The authors aimed at minimizing the power consumption under QoS constraints in particular. But in their model, they assumed that many users have the same impact of interference. They did not tackle the impact of user pair selection. In [26] rate splitting (RS) is being investigated for a NOMA uplink device with a close and far users pairs following cyclic prefixed single carrier transmissions. Frequency-domain equalization is used to support SIC at base-station. This scenario was for two users not located in discs around the base station. They assumed a fixed distance conventional uplink users. However, all the work presented in [23]–[26] does not explore neither concepts of energy harvesting nor data compression and the possible impact of these concepts on the system performance. Moreover, the work presented in [27] investigated the outage

performance of land mobile satellite (LMS) device composed of two terrestrial user nodes executing NOMA in an uplink with either successive interference cancellation or joint decoding at the satellite receiver. A cumulative distribution function (CDF)-based scheduling scheme is investigated in [28] for the uplink NOMA network. Taking into account the SIC and SIC power constraints, they proved that CS-NOMA can achieve better performance than random pairing scheme and OMA scheme.

SWIPT schemes were introduced in [29], [30] under consideration of the EH at the receiving node while the proposed system does not take into account NOMA as the multi-access scheme. On the other hand, RF EH in wireless sensor networks under uplink NOMA scheme in the presence of eavesdroppers considering the possible secrecy outage probability of wireless sensor nodes and base station has been investigated in [31]–[33]. This research focused only on the investigation of the secrecy performance of NOMA for RF EH WSNs without the consideration of data compression at the transmitting nodes. A deferment prospective has been introduced in [34] as they tackle the challenge of leveraging NOMA paradigm for uplink communication from an unmanned aerial vehicle (UAV) to cellular base station, under spectrum sharing with the existing ground users. In [35], the authors suggested a cooperating NOMA, where only the nearest user is an EH node that works as a relay to the far user. The work in [36] optimized energy efficiency in non-cooperative NOMA under power budget and data rate constraints without exploiting the concept of EH to extend the system life time. However, all of the above related work in [31]–[35] does not directly address the impact of data compression on both outage probability and EH under uplink NOMA protocol. An optimization technique for power distribution to boost energy efficient IoT-NOMA network performances has been presented in [37], however the proposed system focuses mainly on the down-link

scenario and disregarded the uplink situation.

Several data compression techniques have been exploited within the context IoT as in [38]–[40]. Data compression can occur using compressive sensing as in [40], which has a low complexity that comes at the cost of poor performance when compared to transform based compression. Discrete wavelet transform (DWT) data compression and reconstruction methods have high construction accuracy. EEG DWT data compression is presented in [18], [41], [42] as a lossless compression technique for EEG signal. However, because of the randomness of the EEG signal, high compression rates cannot be attained with lossless compression. Other transforms such as Walsh transform can also be used for the purpose of data compression as in [43]. Obviusoly, none of the research work presented in [18], [38]–[43] studies the impact of data compression in uplink NOMA EH scenario.

Deep Reinforcement Learning Related work

For optimization in dynamic domains, DRL has shown impressive results. For example, [44] demonstrates the ability of a single DRL agent to adapt to various wireless network circumstances, while still optimally allocating resources (transmission power) to various network nodes. Authors in [45] jointly optimize resource distribution and user association in heterogeneous cellular networks for offloading mobile traffic and it showed that optimal interaction between network efficiency and user quality of service could be achieved. A survey on using DRL in broad applications of IoT has been presented in [46]. Exploiting DRL approaches in uplink NOMA has been studied in [47]–[51]. An investigation of the performance of Federated Learning (FL) update is presented in [47] for mobile edge devices that are connected to the parameter server

wirelessly. The proposed system applies NOMA together with gradient compression in the wireless uplink. The work presented in [48] proposes an uplink NOMA framework for ultra-dense network communications. They study the dynamic energy efficiency problem. The Markov Decision Process model is built through a quantification of resources at Access Points and user equipment to ensure the real-time requirements of user equipment. An investigation of sub-carrier assignment jointly with Power allocation issue in NOMA multi-user uplink system that maximizes energy efficiency thus safeguarding QoS of all-user has been carried out in [50]. They proposed two models using deep q-network (DQN) to figure out the optimum sub-carrier assignment policy and the other model was based on DDPG network to dynamically optimize the transmit power of all users. The work proposed in [49] examined a user clustering based resource allocation under uplink NOMA Multi-cell systems using DRL techniques. It executes user grouping based on Network-traffic to effectively leverage the available resources. Moreover, DRL in the decision making for grant-free NOMA systems has been shown in [51], in order to avoid collisions and improve the system throughput in an unknown network environment. However, the work presented in [47]–[51] neither considers data compression in the proposed uplink NOMA scheme nor application of SWIPT.

CHAPTER 3: DEEP REINFORCEMENT LEARNING ALGORITHM FOR SMART DATA COMPRESSION UNDER NOMA-UPLINK PROTOCOL

In this chapter, we discuss the first contribution of the thesis, which is to design a Deep Reinforcement Learning (DRL) approach that optimizes the quality of the health monitoring data sent to the cloud through a NOMA-Uplink multi-access channel, while meeting the resource constraints of the edge node responsible for delivering the data to the cloud.

System Model

The system model assumes that multiple EEG nodes continuously collect information from patients under surveillance. These nodes are distributed in clusters around mobile edge nodes (ENs) as shown in Figure 3.1. Each edge node is surrounded by a group of EEG nodes named as $u_1, \ldots, u_k$. The whole system is operating under NOMA-up-link protocol. The nodes send the collected data from the patients to the cluster edge node, which will re-transmit this data to the service provider or a cloud.

We assume that there is an agent, which manages the operation of the edge node to optimize the distortion of the EEG data sent to the cloud, while meeting the rate constraints of the edge node. The users are randomly located around the edge node following a homogeneous Poisson Point Process (PPP) so that, each user node is far from the cluster edge node by a certain distance r_i. The nature of the collected data is intense, hence, it necessitates compression prior to transmission to save transmission energy as the nodes are battery-operated and jointly require high bandwidth resources. We assume each node is equipped with an adaptive data compression mechanism based on DWT employing a threshold-based technique shown in [52] . To avoid the data

lengths synchronization problem at each transmission cycle, we assume these sensor nodes are always saturated, which means that these nodes have data to send in each period all the time. As part of our analysis and optimization, Cluster edge nodes will optimize compression ratios based on the distances between itself and the NOMA users and jointly optimize the average distortion among other optimization parameters. Figure 3.1 shows the proposed system architecture.

We assume the wireless channel between cluster edge node and the NOMA users is modeled as a block fading channel, which implies that the channel coefficient remains constant during the transmission block but vary randomly between transmission blocks. We assume that the cluster edge node has full knowledge of the channel state information (CSI) and based on that, the edge node assigns each sensor node the required compression ratio to optimize the overall system performance. The amplitude of the channel gain is assumed to be Rayleigh; therefore the channel distribution follows a complex Gaussian distribution $\mathcal{CN}(0, 1)$. Furthermore, the receiver noise of all nodes is modeled as additive white Gaussian noise (AWGN) with zero mean and variance σ^2. To maintain the system connectivity, the outage probability constraint must be full-filled and should be kept below a minimum threshold enough to allow all nodes to transmit their data without outage.

Proposed Framework and Reinforcement Learning Modeling

The following part shows the derivation of the mathematical model for our proposed system. According to [53], the up-link NOMA system is processed entirely differently from the down-link-NOMA system. In up-link-NOMA protocol, the intended data to be transmitted to the Cluster edge node from the NOMA users are first compressed using

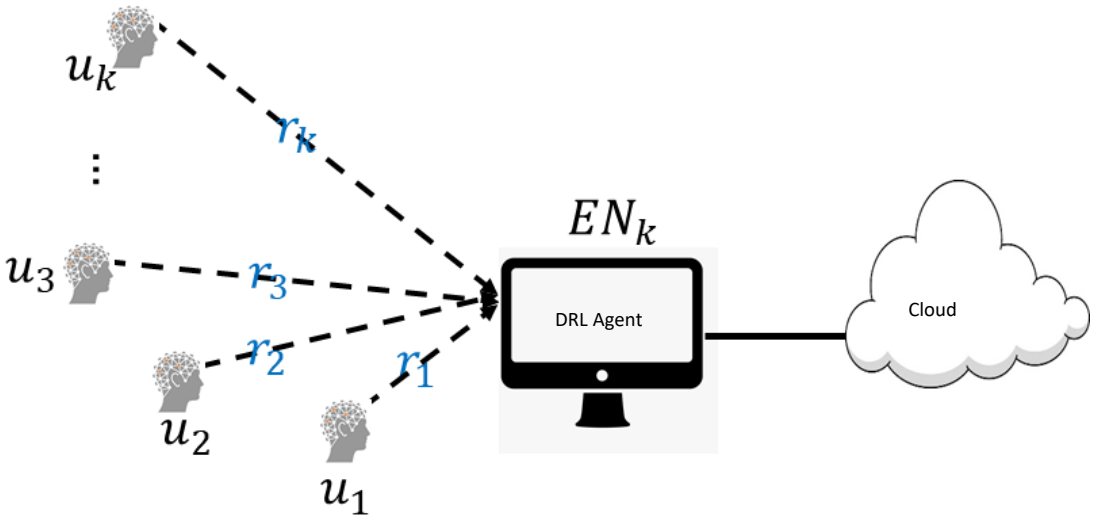

Figure 3.1: System Model for smart Compression Under NOMA Up-Link.

adaptive data compression technique. As per our assumption, all the NOMA users are implementing adaptive data compression using a well known discrete wavelet transform (DWT). Therefore, after compression, the generated data length at a given user node is:

$$N_c = N_a(1 - k), \tag{3.1}$$

where N_a, N_c are the number of samples before and after compression respectively, and $k \in [0, 1]$) represent compression ratio of the raw data at a given node. The decrease in the number of samples of transmitted data due to compression is at the cost of receiver-side distortion after restoration of data. As an example for intensive data transmission, the proposed framework uses the compression paradigm in [15] for EEG. However, it is easy to extend the proposed framework without compromising its generality to take into account various data compression models. By referring to the findings in [15], the amount of the encoding distortion is the Percentage Root-mean - square difference

(PRD) between the recovered EEG signal, and the original signal. Using a real time implementation model in [15] the relationship of encoding distortion to compression ratio, can be described as:

$$D = \frac{d_1 e^{(1-k)} + d_2(1-k)^{-d_3} + d_4 F^{-d_5} - d_6}{100}, \tag{3.2}$$

where d_1, d_2, d_3, d_4, d_5 and d_6 are the parameters estimated statistically from the typical EEG model used in [15] and F is the wavelet filter length of the adopted DWT scheme. The NOMA users simultaneously transmit signals to the cluster edge node over the the available bandwidth applying a control algorithm to allocate the desired compression ratio from the cluster edge node in order to control the sum data rate at the cluster edge node and minimize the outage probability of each user. The cluster edge node receives a superimposed signal from the NOMA users [53]. Moreover, the message of the near user will be decoded first since it has the highest SNR among all users [54]. The cluster edge node applies successive interference cancellation (SIC) to decode the other user's signals after subtracting the near user signal and it will keep doing the same until decoding all users. The objective of this work is to minimize the expected distortion among all users through adapting their compression ratio to meet the total rate of the edge node, and under real block fading channel. The NOMA users adjust the power of their transmitted signal using the NOMA power factors. The NOMA users simultaneously transmit their compressed signals to the cluster head node. The received signal at the edge node per time slot, which is superimposed signal of all users, can be

written as [54]:

$$\boldsymbol{y} = c_1\beta_1\sqrt{P_t}\,\boldsymbol{h}_1\,\boldsymbol{x}_1 + c_2\beta_2\sqrt{P_t}\,\boldsymbol{h}_2\,\boldsymbol{x}_2 + ... + c_k\beta_k\sqrt{P_t}\,\boldsymbol{h}_k\,\boldsymbol{x}_k + \boldsymbol{n},$$

where $\boldsymbol{x}_1 \in \mathbb{C}^{N_c\times 1}$, up to $\boldsymbol{x}_k \in \mathbb{C}^{N_c\times 1}$ are the transmitted signals from each sensor node respectively, $\boldsymbol{y} \in \mathbb{C}^{N_c\times 1}$ is the received signal at the edge node, $\boldsymbol{n}$ is the additive white Gaussian noise (AWGN), which follows $\mathcal{CN}(0,\sigma^2)$ where σ^2 is the noise variance, $\boldsymbol{h}_1$ up to $\boldsymbol{h}_k$ are the complex channel gains of the users $1, 2, .., k$, respectively. $c_1 = \sqrt{\frac{1}{1+r_1^{\alpha}}}$, up to $c_k = \sqrt{\frac{1}{1+r_k^{\alpha}}}$, where r_1 up to r_k are the distances between the edge node and users $1, 2, .., k$, respectively, and $\alpha is the path loss exponent. and \beta_i$ is the transmit power split factor at each user. It is assumed that all users have the same available transmission power, P_t.

Outage performance

As per the NOMA-up-link protocol, the edge node applies successive interference cancellation (SIC) by detecting and decoding the signals sequentially starting from the near user moving to farther ones. Hence, the messages of the farther users are treated as interference. The signal to interference plus noise ratio during the decoding process can be written as:

$$\rho_1 = \frac{\rho\beta_1^2 c_1^2 \mid \boldsymbol{h}_1 \mid^2}{1 + \rho(\beta_2^2 c_2^2 \mid \boldsymbol{h}_2 \mid^2 + ... + \beta_k^2 c_k^2 \mid \boldsymbol{h}_k \mid^2)} \tag{3.3}$$

Where $\rho = \frac{P_t}{\sigma^2}$ is the signal to noise ratio. Therefore, the rate at which the edge node can decode the message sent by the near-user u_1 correctly is

$$R_1 = \log_2(1 + \rho_1). \tag{3.4}$$

The edge node will apply SIC to decode the second nearest signal by canceling out the nearest user signal from the received signal. The signal to interference plus noise ratio for the second near user becomes

$$\rho_2 = \frac{\rho\beta_2^2 c_2^2 \mid \boldsymbol{h}_2 \mid^2}{1 + \rho(\beta_3^2 c_3^2 \mid \boldsymbol{h}_3 \mid^2 + ... + \beta_k^2 c_k^2 \mid \boldsymbol{h}_k \mid^2)}. \tag{3.5}$$

and the rate at which the edge node can decode the message sent by the second nearest user correctly is

$$R_2 = \log_2(1 + \rho_2). \tag{3.6}$$

Finally, the rate at which the edge node can decode the message sent by the farthest k user correctly is

$$\rho_k = \rho\beta_k^2 c_k^2 \mid \boldsymbol{h}_k \mid^2 \tag{3.7}$$

$$R_k = \log_2(1 + \rho_k). \tag{3.8}$$

Using the arithmetic-geometric inequality, we can have the upper bound of the sum rate at the cluster edge as:

$$R_s = \log_2(1 + \rho(\beta_1^2 c_1^2 \mid \boldsymbol{h}_1 \mid^2 + \beta_2^2 c_2^2 \mid \boldsymbol{h}_2 \mid^2 + ... + \beta_k^2 c_k^2 \mid \boldsymbol{h}_k \mid^2)).$$

Therefore, we can define the data rate region to avoid outage considering the target data rate of each user (assuming target data rate thresholds for the NOMA users as $\mathcal{R}_1$ up to $\mathcal{R}_k$, respectively)

$$\begin{aligned} \mathcal{R}_1 &\leq R_1 \\ \mathcal{R}_2 &\leq R_2 \\ &\vdots \\ \mathcal{R}_k &\leq R_k \\ R_1 + R_2 + ... + R_k &\leq R_s, \end{aligned} \tag{3.9}$$

where $R_1, R_2, ...$ and R_k are shown in equations (3.4), (3.6), ..., (3.8), respectively and $\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_k$ representing the threshold rates to avoid outage of NOMA users respectively. Therefore, in order to avoid outage of users.

$$u = \beta_1^2 c_1^2 \mid \boldsymbol{h}_1 \mid^2 + \beta_2^2 c_2^2 \mid \boldsymbol{h}_2 \mid^2 + ... + \beta_k^2 c_k^2 \mid \boldsymbol{h}_k \mid^2 \leq \tau \tag{3.10}$$

where

$$\tau = \frac{2^{\sum_{i=1}^{k} \mathcal{R}_i} - 1}{\rho}. \tag{3.11}$$

The main objective of data compression in our proposed system is to save scarce energy at the NOMA users, which could be IoT nodes with limited access to power source and/or battery operated. The energy consumed by user $i \in [1 : k]$ to transmit their compressed and encoded bits can be given by

$$E_i = E_i^e + E_i^t, \tag{3.12}$$

where E_i^e is the energy consumed for encoding at users i and E_i^t is the energy consumed during transmission of the encoded samples at users i. The encoding energy is negligible compared to transmission energy. Therefore, E_i^t can be written in terms of the rates, R_i, already defined above as:

$$E_i^t = \beta_i^2 \frac{P_t \, \ell_i}{R_i}, \tag{3.13}$$

Where β_i is the power split factor for user i, P_t is the transmission power, R_i is data rate, and ℓ_i is the length of data to be sent from user i. Assuming x_i is the channel gain defined as $x_i = \frac{k\varphi}{N_o}|h_i|^2$, where $k = \frac{-1.5}{log(5BER)}$ as in[15], and N_o is the noise spectral density. so, the required transmission energy to send a data of length ℓ_i with rate R_i is

$$E_i^t = \frac{\beta_i^2 \ell_i}{R_i x_i}(2^{R_i} - 1), \tag{3.14}$$

Optimization Problem

Defining a new parameter denoted by D', which is the complement value of the distortion $D' = 1 - D$ to indicate the user's signal quality, we can formulate the optimization problem as

$$\mathbf{P_1}: \max_{\beta_1,\ldots,\beta_k,k1,\ldots k_k} \mathbb{E}\{D'\}, \tag{3.15}$$

$$\text{s.t.} \quad u \leq \beta$$

$$\delta_i \leq E_i \leq B_{r_i}^t,$$

$$0 \leq k_i, \beta_i \leq 1,$$

where k_i is the compression ratio of user i and $\delta_i = min(\beta_i, P_\epsilon)$ where P_ϵ is the minimum required energy to transmit data without outage. β_i is the power split factor for NOMA users. B_r^t is the current battery level of node i, $B_{ri}^{t+1} = \Gamma_i - E_i^t$ and Γ_i is the available power budget for this node. One thing to notice here is that the problem 3.15 is generally NP hard [15], in addition to the fact that the solution of this problem is greedy with respect to time. In other words, the solution generated will be optimal only at one time step, and hence the optimization has to be performed at each time step, making this solution inefficient for a long time horizon. Therefore, we propose to use DRL as a method to devise a policy of how to set the decision variables to optimize the system performance over a long time horizon.

DRL Agent Design

The environment should be designed to describe the main parameters of the system that will interact with the agent per each time step t. According to the problem description, the environment $\mathbb{E}$ will have a continuous behavior as the episode keeps running with no break state. The agent's behavior will be described by a policy π, which maps states $S_1, S_2, ..S_n \in \mathcal{S}$ into a given actions $a_1, a_2, ..a_n \in \mathcal{A}$ at each time step $t \in \mathcal{T}$, where $\mathcal{S}$ is the state space and $\mathcal{A}$ is the action space. During the experiment at each time step, the environment state S_t will correspond to an action a_t from the agent based on the policy π and then generates the next state $(s_t, a_t) \rightarrow S_{t+1}$. As a result, an immediate delayed reward $\mathbb{R}(S_t, a_t) \rightarrow r_t$ will be attained by the agent. The total cumulative reward across the experiment starting from time $t' = t$ can then be calculated as:

$$\Gamma_t = \sum_{t'=t}^{T} r_t \gamma^{t'-t}, \tag{3.16}$$

where T is cumulative time for the experiment and $\gamma \in [0, 1]$ represents the discount factor. The interactive state/action value function, which is also called the Q-function of a policy π, can be written as:

$$^{\pi}(s, a) = \mathbb{E}_{\pi}[\Gamma_t | s_t = s, a_t = a]. \tag{3.17}$$

The Q-function describes how efficient it is for the agent to perform a specific action in a state as part of a policy π. Replacing Γ_t with its value, the Q-function becomes

$$Q^{\pi}(s, a) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{T} r_t \gamma^{t'-t} | S_t = s, a_t = a\right]. \tag{3.18}$$

the optimal action $a'(s)$ that can be taken at any given state can be calculated according to:

$$a'(s) = \max_{a} Q'(s, a), \tag{3.19}$$

which represents the policy learned. The state space $\mathcal{S}$ in our model includes every possible state at every time step t. It consists of average distortion, node energy consumption as well as the channel state at this time t. All the values are normalized between [0,1], therefore, the normalized state will be $s_t = (\hat{E_i^t}, \hat{D_i^t}, \hat{h_i^t}), \forall_s \in \mathcal{S}$. The agent's action at a given time t,is the NOMA power split factor β_i^t for the NOMA node and the data compression ratio k_i^t. The agent's actions are also normalized between [0,1]. Hence, the action space at a given time step t will be $a_t = (\hat{k_i^t}, \hat{k_i^t}), \forall_s \in \mathcal{A}$. The state transitions in our model are deterministic in the system, since the above values are analytically based on the derived equations mentioned above.

DDPG-based Approach for Minimizing Distortion

Deep Deterministic Policy Gradient (DDPG) is a model-free off-policy algorithm for continuous action learning [55]. It incorporates Deterministic Policy Gradient (DPG) and Deep Q-Network (DQN), uses the DQN experience replay and slow-learning target networks, and is based on DPG, which can operate across continuous spaces of actions and states. The "deterministic" in DDPG refers to the fact that the actor explicitly evaluates the action rather than the distribution of the likelihood over actions. Q-learning is the most common off-policy reinforcement learning algorithm because the Q-function updates the Q-values based on action that are outside the current policy and initializes the Q table randomly. Using optimization techniques directly in continuous space of actions is difficult to apply because it is necessary to optimize greedy policies at every time step. Such kind of optimization is too slow to be realistic with large or unconstrained problems and nontrivial action spaces. Alternatively, DDPG approach, which is based on actor-critic networks, can be used to devise an efficient suboptimal policy for continuous action space. The actor neural network maintains the parametric actor function $\mu(s|\theta^{\mu})$ that specifies the current policy by mapping states to a specific action deterministically. In DQN the optimal action is taken by taking argmax over all actions of Q-values. In DDPG the actor is a network of policies that does precisely that. It explicitly evaluates the action bypassing the argmax as.

$$Q^{'}(s,a) = arg\max_{a} Q^{'}(s,a), \tag{3.20}$$

However, to measure a state's Q-value, the actor output is fed into the Q-network for calculation of the Q-value. For calculating the Q values, we use the target critic network

and pass the action calculated by the target actor network. Each network has a time-delayed copy of itself. These target networks have been used to stabilize the learning process and these target networks will be updated in a soft manner based on the main networks. The critic network $\mathbb{Q}(S, a)$ learns using the Bellman equation which describes the optimal action value function as shown in [56] which is:

$$Q^{\mu}(s_t, a_t) = \mathop{\mathbb{E}}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma \max_{a_{t+1}} Q^{\mu}(s_{t+1}, a_{t+1})], \tag{3.21}$$

To encourage exploration, a certain Gaussian noise is applied to the policy-determined action. The optimal approximation of the action value function will be given by $Q(s, a; \phi)$, Where ϕ is the parameter set for the Q-value neural network, meanwhile, the approximation of the action value function , shall be described by $\mu(s; \theta)$, where θ is the set of parameters for the policy of the Neural network. The function approximator $Q(s_t, a_t; \phi)$ is assumed to be differential with respect to the moving action statement, which means for one policy $\mu(s; \theta)$ we can create a gradient based learning rule that minimizes the expensive computation of $\max_{a_t} Q(s_t, a_t)$ over the continuous action space to be $Q(s_t, \mu(s_t; \theta))$. The mean square Bellman error (MSBE) represents the the error function, which indicates how far the approximation from satisfying Bellman equation. The loss function for a given state sampled from the environment can be defined for both networks as:

$$\mathcal{L}(\phi_t, D) = \mathop{\mathbb{E}}_{(s,a,r,s^{'},d) \sim D}[\Big(Q_{\phi_t}(s, a) - y(r, s^{'}, d)\Big)^2], \tag{3.22}$$

and the temporal difference (TD) target error is

$$y(r, s^{'}, d) = r + \gamma(1 - d)\left(Q_{\phi_t}(s^{'}, \mu(s^{'}; \theta))\right), \tag{3.23}$$

where D is a set of transitions sampled from the environment and d is the terminal. Therefore, the main objective is to learn a deterministic policy $\mu(s, \theta)$ that allows the actions to maximize $Q(s_t, a_t; \phi)$. The complete architecture of the DDPG system is shown in Figure 3.2. The main purpose of the parameter ϕ is policy evaluation. The training process will utilize the replay buffer that represents the previous experience of learning. This can improve the data and stabilize the training process of the neural networks, accordingly. Each network has a time-delayed copy of itself in order to stabilize the minimization process of MSBE during the training. In order for the algorithm to achieve a stable behavior, the replay buffer must be large enough to contain a wide range of experiences. The DDPG Training and testing algorithm in pseudo code is shown in Algorithm 1.

As shown in Figure 3.2, which describes the DDPG model architecture based on our objective function in (3.15). The main goal is to minimize the expected distortion in (3.15) by incorporating its impact in the reward function in order to obtain the trade-off between minimum distortion, satisfying the main constraints in the problem including the compression ratios and transmission rates of the NOMA users. The Markov Decision Processes (MDPs) shall be considered while designing the environment E. The MDPs shall model the relations between the agent and the environment, while the environment changes continuously. The environment is episodic, which represents the lifetime of the nodes batteries, and the episodes represent the dynamics of the environment

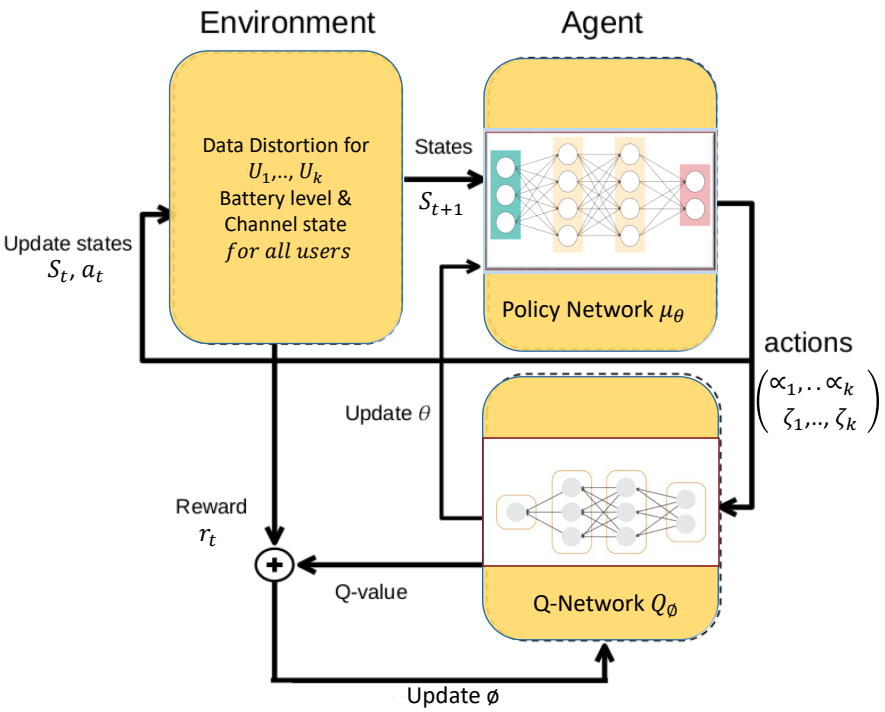

Figure 3.2: DDPG system environment/agent architecture model 1.

changes. The environment is fully observable by the agent, and the state, s_t, is represented by $[E_1, .., E_k, D_1, .., D_k, h_1, .., h_k]$, where $[E_1, .., E_k, D_1, .., D_k, h_1, .., h_k] \in \mathcal{S}$ and $h_1, .., h_k$ represents the channels between users and the edge node. All of the parameters have to be normalized to train the network's nodes; therefore, the state will be $s_t = [\hat{E}_1, .., \hat{E}_k, \hat{D}1, .., \hat{D}_k, \hat{h}_1, .., \hat{h}_k], \forall s \in \mathcal{S}.$, where $a_{t+1} \rightarrow \mu(s_t|\theta^\mu)$ and s_{t+1} implies that the next state is sampled by the environments following the distribution $P(\cdots|s, a)$. We can notice that the state transitions are not deterministic in the system due to the fact that all of these parameters depend on the randomness of the channel estimation. Nevertheless, the state transition can still be calculated by invoking our optimization parameters $[\beta_1, .., \beta_k, k_1, .., k_k]$ where $[a_t = \beta_1, .., \beta_k, k_1, .., k_k] \in \mathcal{A}$ represents the action space of the system. The agent will generate these actions and invoke them with the environment to calculate the next state and get the discounted reward based on the

previous state as shown in Algorithm 1.

Reward Function

The reward function at each time step, t, is the most important one that describes the main parameters of the optimization problem and it is a function of the current state s_t and current action a_t. The aim of the problem is to maximize D'_{i_t}, while satisfying the main constraints. Therefore, D'_{i_t} must be involved in the reward function and jointly we need to optimize the energy consumption at the nodes. Hence, we assume a battery indicator χ_i to be involved into the state parameters and $\chi_i^{t+1} = \chi_i^t - E_i^t$. The reward function is given by

$$r_t = \begin{cases} \lambda_1(1 - D'_t) + \lambda_2 \sum_i \frac{(\chi_i^t - E_i^t)}{\chi_0} & \text{if remaining constraints in (3.15) hold} \\ -1 & \text{otherwise} \end{cases}, \quad (3.24)$$

where λ_1, λ_2, are the weights for each term and $\lambda_1 + \lambda_2 = 1$. The condition for the reward is the constraints listed in the problem above in order to ensure the parameters do not go below certain lower bounds, otherwise the reward will be penalized by -1. The weights above play an important role in the system performance to obtain the trade-off between these parameters. These weights could be adjusted based on the system requirements. χ_0 represents the total battery capacity.

Performance Evaluation

A simulation using Matlab has been conducted with numerical results for data compression under our proposed up-link NOMA approach for multiple users located around the edge node with different distances. User 1 is the closest to the edge node,

Algorithm 1 Deep Deterministic Policy Gradient algorithm [57]

Input:Initializing Q network Q_ϕ and policy network μ_θ with weights ϕ and θ,

Initializing the target networks parameters with $\phi_{targ} \leftarrow \phi$, $\theta_{targ} \leftarrow \theta$

Initialize Replay buffer $\mathcal{D}$

for episode i = 1 to M **do**

Receive initial state s_0

for t = 1:I **do**

Select action $a_t =$ Clip $(\mu_\theta(s) + \varepsilon, a_{low}, a_{high})$ where $\varepsilon \sim \mathcal{N}(0, 0.1)$

Execute action a_t in environment,observe next state s', reward r and done signal d.

Store (s, a, r, s', d) in replay buffer $\mathcal{D}$

if it's time to update and there are enough samples in $\mathcal{D}$ **then**

Randomly sample a batch $\mathcal{B} = (s, a, r, s', d)$ of transitions from $\mathcal{D}$

Compute targets y:$y(s', r, d) = \Big(r + \gamma(1-d) Q_{\phi_{targ}}\big(s', \mu_{\theta_{targ}}(s')\big) \Big)$

Compute loss function $\mathcal{L}$: $\mathcal{L}(\phi, \mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',d) \in \mathcal{B}} \Big(Q_\phi(s,a) - y(s',r,d) \Big)^2$

Update Q network parameters by one step of gradient descent:

$\phi \leftarrow \phi - \eta_\phi \nabla_\phi \mathcal{L}(\phi, \mathcal{D})$

Update policy network parameters by gradient ascent:

$\theta \leftarrow \theta + \eta_\theta \nabla_\theta \frac{1}{|\mathcal{B}|} \sum_{s \in \mathcal{B}} Q_\phi\big(s, \mu_\theta(s)\big)$

Update target networks parameters:

$\phi_{targ} \leftarrow (1-\rho)\phi_{targ} + \rho\phi$

$\theta_{targ} \leftarrow (1-\rho)\theta_{targ} + \rho\theta$

Testing

next User 2, while User 3 is the farthest with the lowest channel gain.

DDPG Model Convergence

Figure 3.3 shows the convergence behavior of our DDPG-based algorithm. The algorithm had an exploration decay rate of $\phi = 0$ during the first 100 episodes, which implies that the entire experiment was observed during those episodes. This means that the algorithm analyses the whole continuous action space to figure out the most recompensed actions necessary to develop the optimal policy, thereby maximizing the benefits. Since we have continuous action spaces, exploration is done via adding noise to the action itself. Afterwards, the entire exploration term decreases to almost 0, so that the full exploitation is accomplished and hence searches for actions that only yield the greatest possible rewards. The efficiency has stabilized for all the channel gain values, converging roughly after 2000 episodes, i.e., the algorithm finds an optimal policy to achieve the highest reward. In Figure 3.3, we present the average reward of the proposed DDPG algorithm, which takes into account immediate changes in the environment as a function of channel gain and battery level. Hence, it reduces multiple needs for re-optimization when the environment changes. Since the distortion has direct relation with the compression ratio as it increases while compression ratio increases in nonlinear relationship, the other factor that has impact is the NOMA power split factor. This impact arise due to the increasing the transmission power reduces the effect of the error channel on the transmitted data. The DRL shows the minimum average distortion is occurred when $\beta_1 = 1$,$\beta_2 = 0.1$ and $\beta_3 = 0.21$ the compression ratios was $k_1 = 0.68$,$k_2 = 0.12$ and $k_3 = 0.12$. However, In order to ensure training fairness and maximize efficiency between energy consumption of the nodes and the compression ratio, we set $\lambda_1 = 0.7$

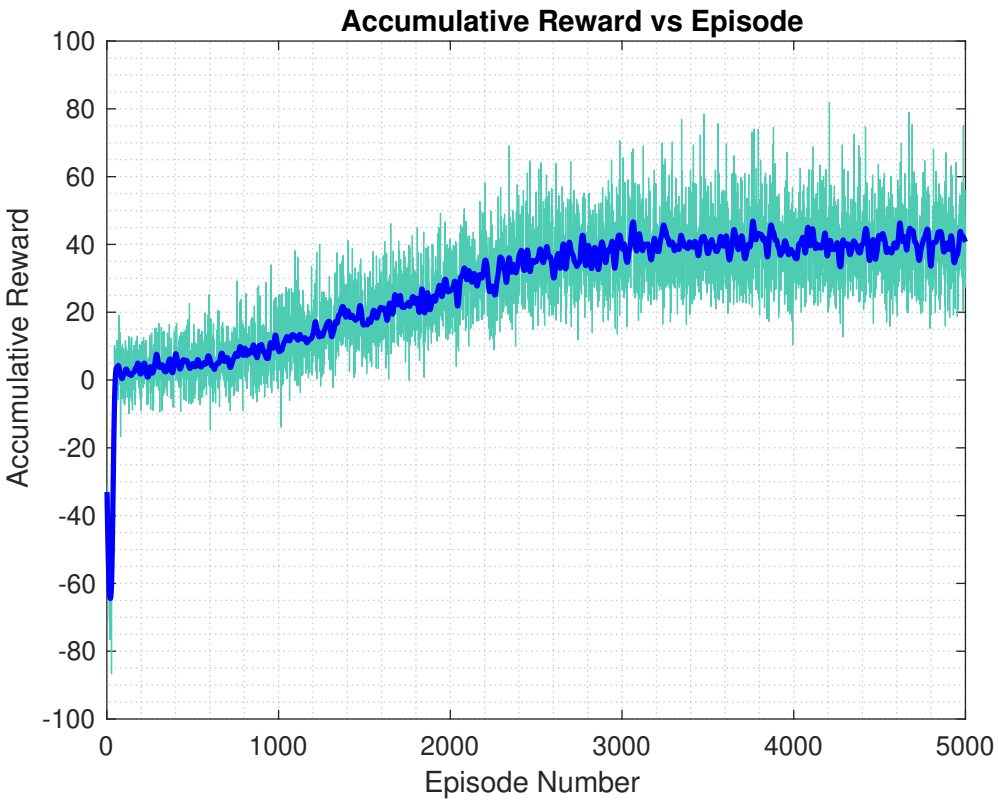

Figure 3.3: DDPG model reward versus episodes.

and $\lambda_2 = 0.3$. The results reflects the effect of NOMA power split factors as well as the compression ratio on the system performance.

Adaptive compression and distortion

Figure 3.4 shows the average distortion versus the remaining battery level for different users. As the average distortion increases (this mean higher compression ratio) the remaining battery increases. Figure 3.5, presents the remaining battery level against the compression ratios for the different users. As shown, the remaining battery level increases with the compression ratio. The results have been calculated with respect to the channel gain for these users.

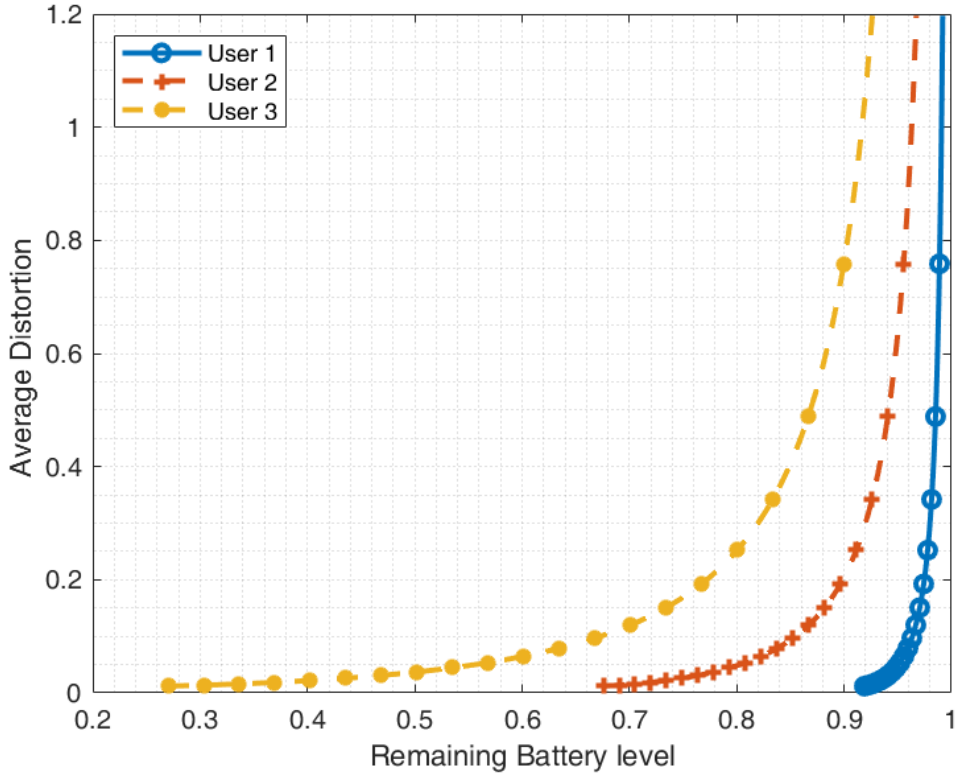

Figure 3.4: Average distortion versus remaining battery level for 3 different users.

Energy consumption

Finally, Figure 3.7 reflects the exponential relationship between the node energy consumption and the NOMA power split factor β, as β raises up the energy consumption raises too.In Figure 3.6, we show an example of the energy consumption versus the compression ratio for the nearest node to cluster head. the graph depicts the decreasing of power consumption with the increase of the compression ratio. We can clearly notice the effect of the NOMA allocated power transmission ratio $\beta_1, \beta_2, \beta_3$ on the energy consumption from the graph. Figure 3.8 shows the outage probability versus the compression ratio at different power split factors. the graph shows most of outage event can be controlled by adjusting the targeted data rate threshold.

To benchmark between the performance of the greedy solution (where the greedy

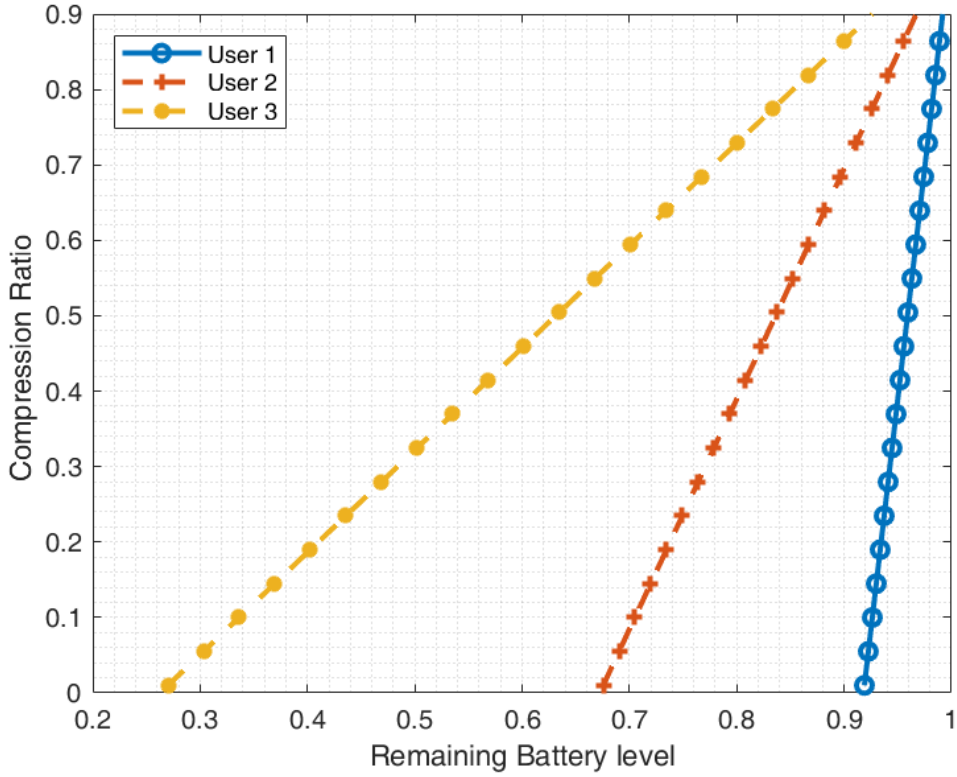

Figure 3.5: Compression ratio versus remaining battery level for 3 different users.

solution is greedy with respect to time) and the DRL solution, figure 3.9 shows the average reward versus time at a certain time period. The reward decaying reflects the effect of the battery level on the average reward as the battery level decay with time. Figure 3.10 presents the average distortion with time of the DRL solution with average level around 25%. Figure 3.11 reflects the battery level changes with time from the DRL.

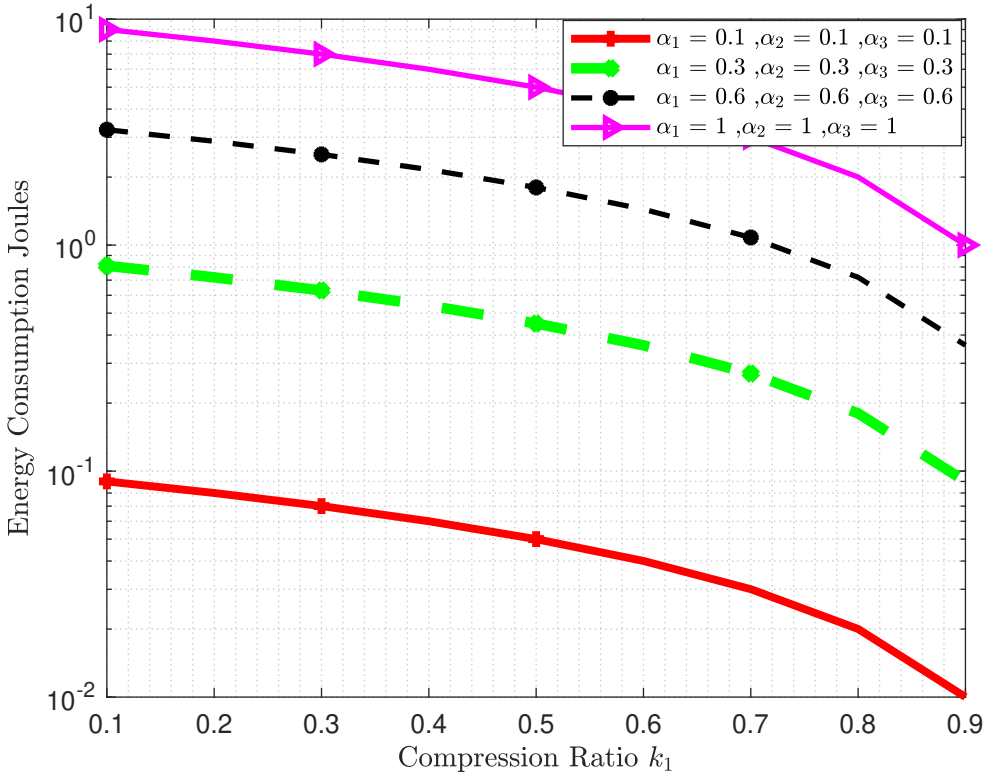

Figure 3.6: Energy consumption vs compression ratio for user 1.

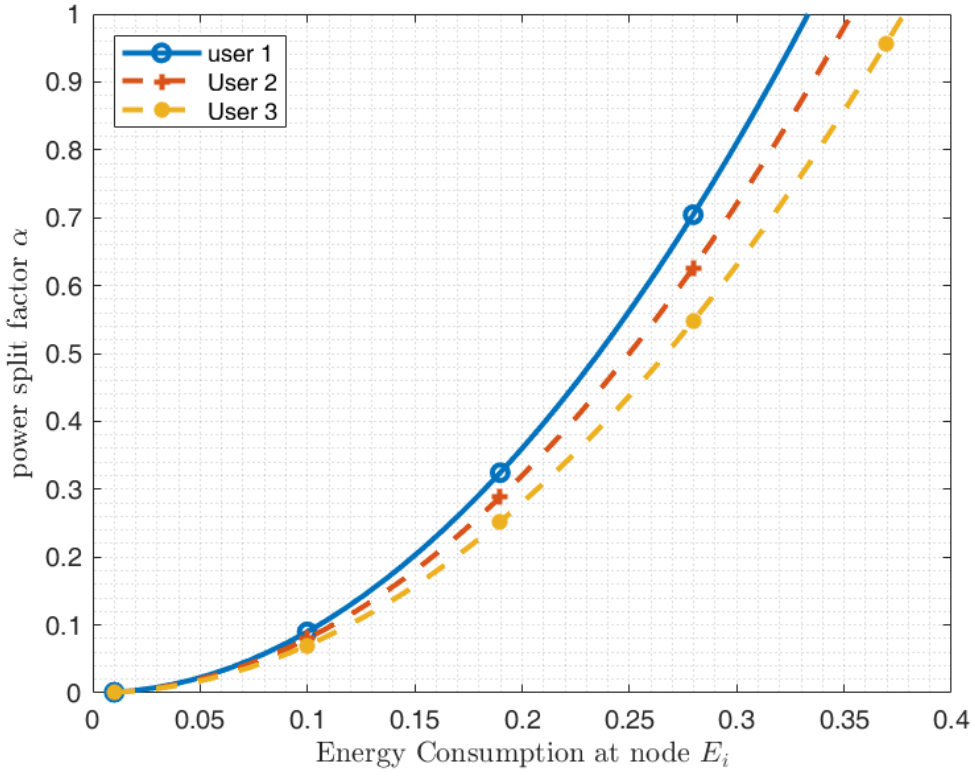

Figure 3.7: Energy consumption vs NOMA power split factor for 3 different users.

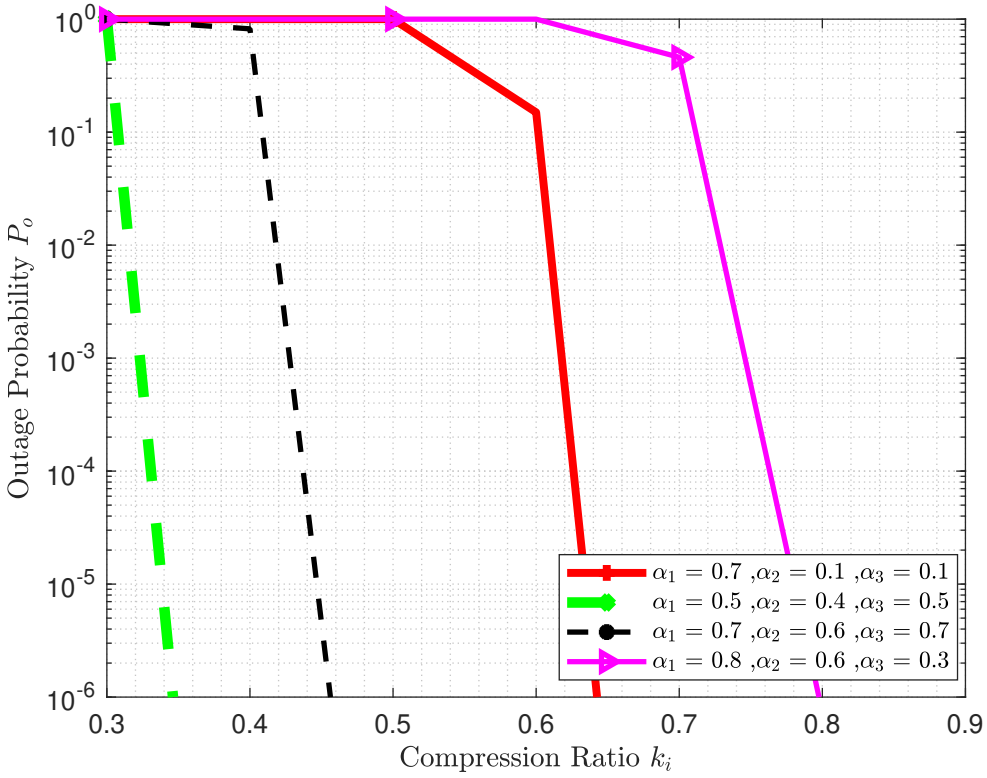

Figure 3.8: Outage probability versus compression ratio for 3 different users.

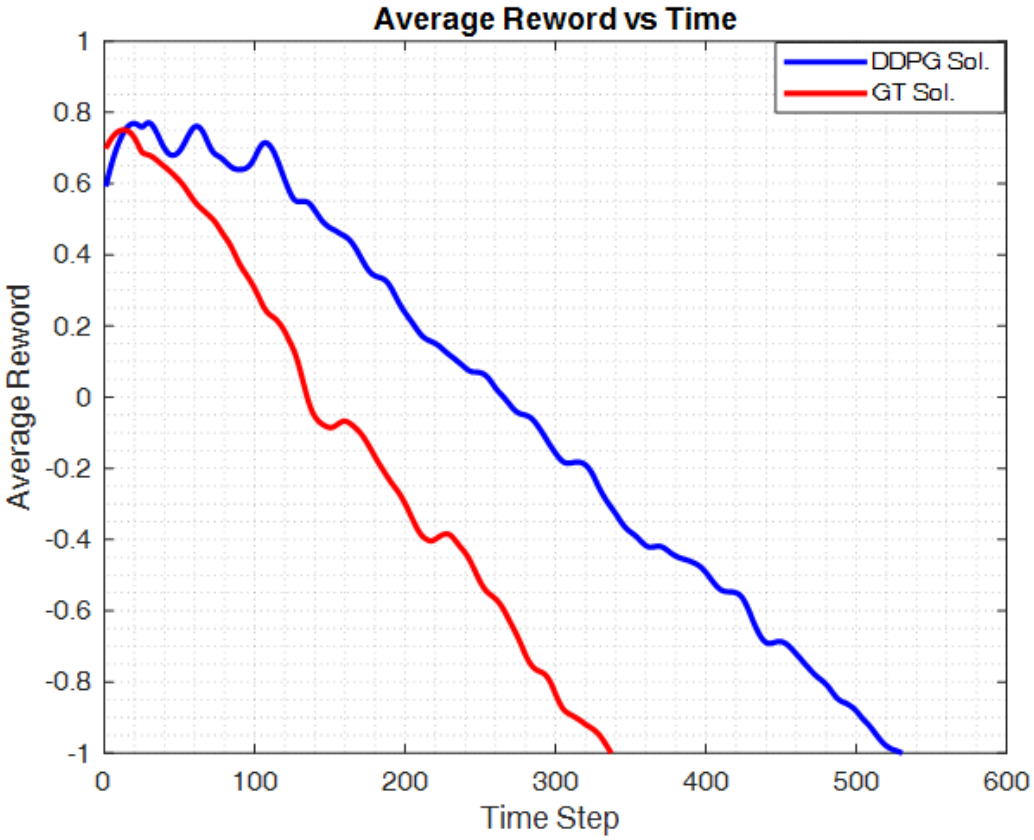

Figure 3.9: Average reward vs Time.

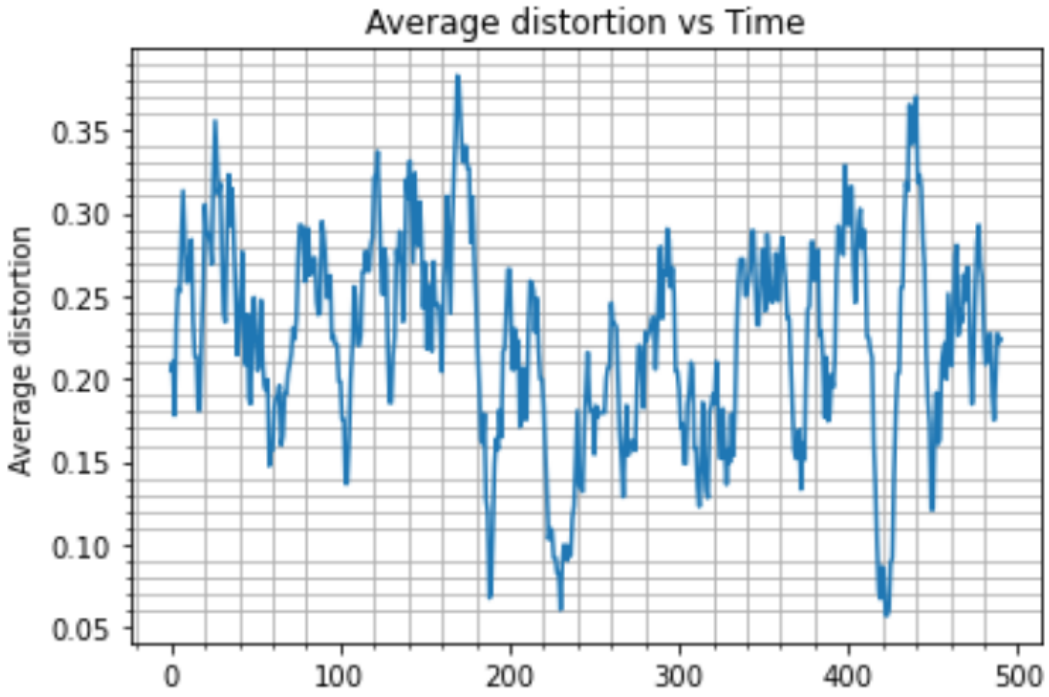

Figure 3.10: Average distortion with Time from DRL.

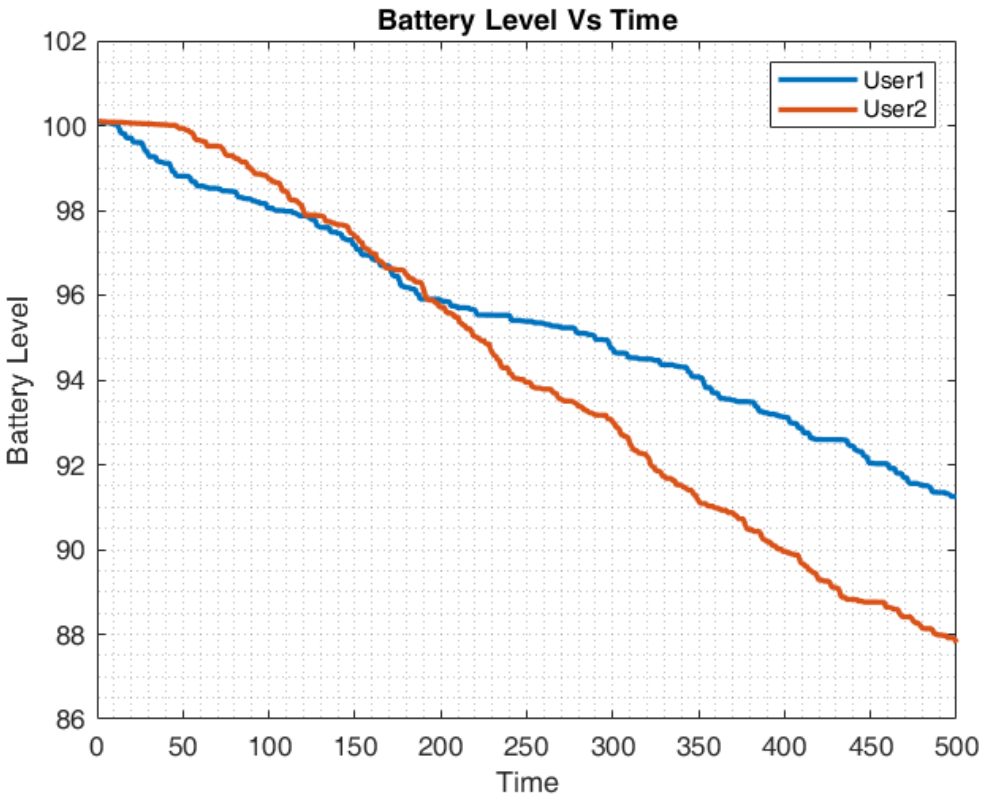

Figure 3.11: Battery level with Time from DRL.

CHAPTER 4: DEEP REINFORCEMENT LEARNING FOR EFFICIENT DATA TRANSMISSION AND ENERGY HARVESTING UNDER NOMA-UP-LINK PROTOCOL

System Model

The system model that we introduced in chapter 3 has been expanded to tackle the Energy harvesting at the cluster head node using the SWIPT paradigm. We consider a network operating under a NOMA scenario with users grouped in discs surrounding the cluster-head (CH) node. The discs are categorized into two sets: near user discs denoted by A_j, where $j \in [1 : J]$ and J is the total number of near users discs, and far users discs denoted B_m, where $m \in [1 : M]$, where M is the total number of far user discs. Users located in disc A_j that has inner radius, r_{a_1} and outer radius r_{a_2} are referred to as near users and denoted as $a_1, \ldots, a_i, \ldots, a_n$. Users located in disc B_m that has inner radius, r_{b_1} and outer radius r_{b_2} are referred to as far users and denoted as $b_1, \ldots, b_i, \ldots, b_n$. Note that $r_b > r_a$. In each disc, the users are randomly located according to the homogeneous Poisson Point Process (PPP). The NOMA users are assumed to be connected to IoT sensors that collect data from surrounding environment such as images, videos or EEG data. These types of data are intense and require compression prior to transmission to save transmission energy and increase bandwidth efficiency.

We assume that wireless channel between CH and the NOMA nodes are modeled as block fading channels, which implies that the channel coefficient remain constant during the transmission block but vary randomly between transmission blocks. The channel is modeled as Rayleigh channel that follows a complex Gaussian distribution $\mathcal{CN}(0, 1)$. Furthermore, the receiver noise is modeled as additive white Gaussian noise (AWGN)

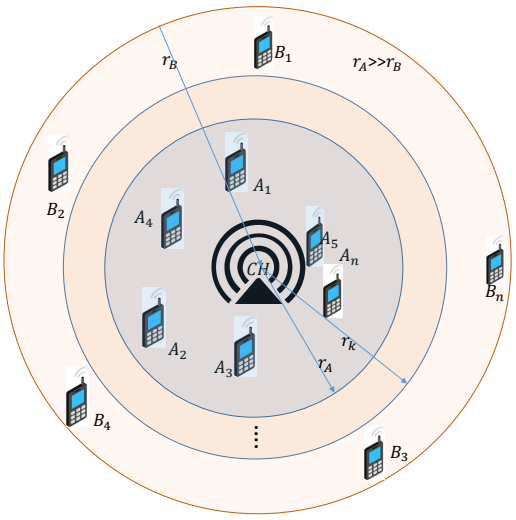

Figure 4.1: Network Topology

with zero mean and variance σ^2.

Transmission Protocol description and Analysis

In the following, we form the basis for our system analysis and derivations.

Framework Description

In our unlink (UL) NOMA framework, the intended data to be transmitted to the CH from the two paired users are first compressed using adaptive data compression technique. According to our proposed model, we assume that user a_i from disc A_j is paired with user b_i from disc B_m. Selection of the users within the discs is not in the scope of this work and hence moving forward, we will drop the subscripts i and j and will denote the selected NOMA pair as a and b. The paired users simultaneously transmit signals to the CH over the same frequency resource using a power control algorithm imposed by the CH. The CH receives a superimposed signal from the two users. Then the message of the near user a will be decoded first due to the possibility

Table 4.1: Summary of the notations.

Parameter Name	Discretion
$\mathbf{z}_a$ and $\mathbf{z}_b$	Data to be sent by users a and b
$\mathbf{\Psi}_{j,k}$	The wavelet functions
k_a, k_b	The compression ratios
d_a, d_b	Distance from the Cluster head for user a and b
$\mathbf{H}_k^t \in \mathbb{C}^{N_c \times N_c}$	The Teoplitz channel matrix with the channel coefficients between CH and users k as its first column
β_a, β_b	NOMA power factors for user a and b
θ	Power split factor for the cluster head node
P_t, σ^2	Available transmission power at the nodes and the noise variance
β_a, β_b	The distortion ratios of the received signal
Tr$\{\mathbf{X}\}$	denotes the trace of matrix $\mathbf{X}$.
k_a, k_b	the compression ratios for user a and b
${}_{[z]2}F_1\left(a, b; c; d\right)$	The incomplete version of the Gauss hypergeomtric function
$\epsilon, \delta_a, \delta_b, \mu_a, \mu_b$	Threshold levels for outage probability,node energy, and distortion ratios respectively
$\mathcal{R}_a, \mathcal{R}_b$	Target data rates for users a and b

of having stronger SNR than the other user b. The CH then applies SIC to remove the decoded signal that belongs to a from the received signal prior to decoding the message sent by b.

We assume the CH is capable of energy harvesting (EH) using SWIPT. The CH applies a power splitting (PS) scheme, where a fraction of the received signal is used for EH.

As per the information available at the CH and based on key criteria including required harvested energy, outage probability, compression ratio and distortion ratio, the CH selects the near user disc from the set of the available near user discs (set A) and the far user disc from the available far user discs (set B). In addition, the CH assigns to the user the desired compression ratios and the NOMA power factors that adjust the transmitted power for the near and far users selected from the selected discs. The objective is then to maximize the harvested energy by the CH, while minimizing the distortion ratio and outage probability and maximizing the compression ratio at the NOMA users.

Adaptive Data Compression

The data to be transmitted by the two NOMA users are first compressed. Adaptive data compression is implemented using a famed, discrete wavelet transformation (DWT) compression approach presented in [15]. There exists many types of the wavelet function including Daubechies, Haar and Morlet[1].The users employ a thresholding-based technique for data compression that uses discrete wavelet series expansion. Data to be sent by users a and b are denoted as $\mathbf{z}_a$ and $\mathbf{z}_b$, respectively. A DWT operation is

[1]The reader is referred to **mallat_book**, [58] and the references therein for more details on how the wavelet functions are designed.

applied on $\mathbf{z_a}$ and $\mathbf{z}_b$, which yield to a compressed signal donated as $\mathbf{x_a}$ and $\mathbf{x_b}$. The compression ratio for each of user $a$and b can then calculated by

$$k_a = (1 - \frac{N_a}{L_a}) \times 100, \tag{4.1}$$

$$k_b = (1 - \frac{N_b}{L_b}) \times 100, \tag{4.2}$$

where L_a and L_b are the lengths of the original signal and N_a and N_b are the number of non-zero samples generated after the thresholding at users a and b, respectively. It is assumed that the IoT sensor nodes, i.e., NOMA users are identical and hence these steps are same at the two nodes. Those steps are then reversed at the receiver side to yield the transmitted signals $\mathbf{x}_a$ and $\mathbf{x_b}$. Since these steps are common and for simplicity, it is assumed that the number of transmitted samples after passing the signals through those blocks are mapped one to one. In other words, the number of the lengths of the transmitted signals by users a and b are N_a and N_b, respectively.

NOMA Uplink Transmission

The two NOMA users split the number of samples to be transmitted into a fixed time slots with length N_c. Hence, we have $N_a = \ell_a N_c$ and $N_b = \ell_b N_c$, where ℓ_a and ℓ_b are integer values that represent the number of needed time slots for users a and b, respectively. When $\mod(N_a, N_c) \neq 0$ or $\mod(N_b, N_c) \neq 0$, the last time slot in user's a or b transmission is appended by a number of zeros $N_z^a = N_c - \mod(N_a, N_c)$ and $N_z^b = N_c - \mod(N_b, N_c)$, respectively, where $\mod(.,.)$ is the modulus operation. Hence, we will have $\ell_a = \frac{N_a + N_z^a}{N_c}$ and $\ell_b = \frac{N_b + N_z^b}{N_c}$. It is worth noting that since $N_a \gg N_c$ or $N_b \gg N_c$, appending the last time slot by zeros should

have a negligible impact on the transmission efficiency.

The two NOMA users adjust the power of their transmitted signal using the NOMA power factors. The two NOMA users simultaneously transmit their compressed signals to the CH. The received signal at CH per time slot can be written as

$$\mathbf{y} = \beta_a c_a \sqrt{P_t}\, \mathbf{H}_a^t\, \mathbf{x}_a + \beta_b c_b \sqrt{P_t}\, \mathbf{H}_b^t\, \mathbf{x}_b + \mathbf{n}, \tag{4.3}$$

where $\mathbf{x}_a \in \mathbb{C}^{N_c \times 1}$, $\mathbf{x}_b \in \mathbb{C}^{N_c \times 1}$, $\mathbf{y} \in \mathbb{C}^{N_c \times 1}$, $\mathbf{n} \in \mathbb{C}^{N_c \times 1}$, β_k, $k \in \{a, b\}$, is the allocated power factors to user k, and $\mathbf{H}_k^t \in \mathbb{C}^{N_c \times N_c}$ is the Teoplitz channel matrix with the channel coefficients between CH and users k as its first column. The channel taps are assumed to be complex Gaussian random variables with zero mean and variance $\sigma_{k,n}^2$ for the n_{th} tap of the channel between CH and user k. $c_a = \sqrt{\frac{1}{1+d_a^\alpha}}$, $c_b = \sqrt{\frac{1}{1+d_b^\alpha}}$, where d_a and d_b are the distances between CH and a, b, respectively, α is the path loss exponent. $\mathbf{n}$ is the AWGN, which follows $\mathcal{CN}(0, \sigma^2)$, where σ^2 is the noise variance and P_t is the available transmit power at users a and b, which we assume to be the same.

The CH splits its received signal between EH and signal decoding using a PS factor θ. The signal that will be then decoded can be written as

$$\mathbf{y}_d = \theta \left(\beta_a c_a \sqrt{P_t}\, \mathbf{H}_a^t\, \mathbf{x}_a + \beta_b c_b \sqrt{P_t}\, \mathbf{H}_b^t\, \mathbf{x}_b \right) + \mathbf{n}. \tag{4.4}$$

The CH applies SIC by decoding the message of the near user a first. The CH then removes this signal from the received signal before decoding the far user's signal. After reversing the transmission steps stated earlier, $\mathbf{y}_d$ yields $\hat{\mathbf{x}}_a$ and $\hat{\mathbf{x}}_b$. The original data is retrieved at the CH through an inverse DWT operation that is applied on $\hat{\mathbf{x}}_a$ and $\hat{\mathbf{x}}_b$. The distortion ratios calculated through the root mean square difference between the original

and reconstructed data is given by:

$$D_a = \frac{||\mathbf{x}_a - \hat{\mathbf{x}}_a||}{||\mathbf{x}_a||} \times 100, \tag{4.5}$$

$$D_b = \frac{||\mathbf{x}_b - \hat{\mathbf{x}}_b||}{||\mathbf{x}_b||} \times 100, \tag{4.6}$$

where $\hat{\mathbf{x}}_a$ and $\hat{\mathbf{x}}_b$ are the retrieved data at users a and b, respectively.

Outage Probability

As per uplink NOMA protocol, the CH applies SIC by detecting and decoding the signal of the near user a first. Hence, the signal of the far user is treated as interference. The rate at which CH can correctly decode the message sent by the near user a is

$$R_a = \log_2 \det \Bigg(\mathbf{I}_{N_c} + \rho\theta^2 \beta_a^2 c_a^2 \mathbf{H}_a^t \mathbf{H}_a^{t\,*} \Bigg(\mathbf{I}_{N_c} + \rho\theta^2 \beta_b^2 c_b^2 \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \Bigg)^{-1} \Bigg), \tag{4.7}$$

where $\mathbf{I}_{N_c}$ is the identity matrix with size $\mathbf{N_c} \times \mathbf{N_c}$ and $\rho = \frac{P_t}{\sigma^2}$. The CH then cancels out this signal from the received signal prior to decoding the far user's signal. The rate at which CH can correctly decode the message sent by the near user b is

$$R_b = \log_2 \det \left(\mathbf{I}_{N_c} + \rho\theta^2 \beta_b^2 c_b^2 \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right). \tag{4.8}$$

The sum rate is then

$$R_s = \log_2 \det \left(\mathbf{I}_{N_c} + \rho\theta^2 \beta_a^2 c_a^2 \mathbf{H}_a^t \mathbf{H}_a^{t\,*} + \rho\theta^2 \beta_b^2 c_b^2 \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right). \tag{4.9}$$

Using the arithmetic-geometric inequality, we can have

$$\begin{aligned} R_s &= \log_2 \det \left(\mathbf{I}_{N_c} + \rho\theta^2 \beta_a^2 c_a^2 \mathbf{H}_a^t \mathbf{H}_a^{t\,*} + \rho\theta^2 \beta_b^2 c_b^2 \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right) \\ &\leq \log_2 \left(1 + \rho\theta^2 \beta_a^2 c_a^2 \text{Tr} \left\{ \mathbf{H}_a^t \mathbf{H}_a^{t\,*} \right\} + \rho\theta^2 \beta_a^2 c_a^2 \text{Tr} \left\{ \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right\} \right), \end{aligned} \tag{4.10}$$

where Tr$\{\mathbf{X}\}$ denotes the trace of matrix $\mathbf{X}$.

The data rate region is defined as

$$\begin{aligned} \mathcal{R}_a &\leq R_a \\ \mathcal{R}_b &\leq R_b \\ R_a + R_b &\leq R_s \end{aligned} \tag{4.11}$$

where R_a, R_b, and R_s are defined in Equations (4.7), (4.8), and (4.10), respectively. Assuming target data rates for users a and b are $\mathcal{R}_a$ and $\mathcal{R}_b$ respectively, the outage probability

$$\begin{aligned} P_o &= \Pr \left\{ R_s < \mathcal{R}_a + \mathcal{R}_b \right\} \\ &= \Pr \left\{ \log_2 \det \left(\mathbf{I}_{N_c} + \rho\theta^2 \beta_a^2 c_a^2 \mathbf{H}_a^t \mathbf{H}_a^{t\,*} + \rho\theta^2 \beta_b^2 c_b^2 \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right) \right\} < \mathcal{R}_a + \mathcal{R}_b \\ &= \Pr \left\{ \beta_a^2 c_a^2 \text{Tr} \left\{ \mathbf{H}_a^t \mathbf{H}_a^{t\,*} \right\} + \beta_b^2 c_b^2 \text{Tr} \left\{ \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right\} < \frac{2^{(\mathcal{R}_a + \mathcal{R}_b)} - 1}{\rho\theta^2} \right\}. \end{aligned} \tag{4.12}$$

Letting $u = \beta_a^2 c_a^2 \text{Tr} \left\{ \mathbf{H}_a^t \mathbf{H}_a^{t\,*} \right\} + \beta_b^2 c_b^2 \text{Tr} \left\{ \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right\}$, we have

$$P_o = \Pr \left\{ u < \frac{2^{(\mathcal{R}_a + \mathcal{R}_b)} - 1}{\rho\theta^2} \right\}. \tag{4.13}$$

Note that $\text{Tr}\left\{\mathbf{H}_k^t \mathbf{H}_k^{t\,*}\right\} = \sum_{i=1}^{N_c}\sum_{j=1}^{N_c}|h_{i,j}^k|^2$. Since $h_{i,j}^k$ follows $\mathcal{CN}(0,1)$, $|h_{i,j}|$ follows a Rayleigh distribution with a scale parameter $\frac{1}{\sqrt{2}}$. Therefore, $|h_{i,j}|^2$ follows exponential distribution with parameter 1. Since $\mathbf{H}_k^t$ is a lower triangular matrix, the number of non-zeros elements in $\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}|h_{i,j}|^2$ is $N_0 = \frac{N_c(N_c+1)}{2}$. Therefore, $\text{Tr}\left\{\mathbf{H}_k^t \mathbf{H}_k^{t\,*}\right\}$ follows a Gamma distribution with shape N_0 and scale 1, i.e., $\text{Tr}\left\{\mathbf{H}_a^t \mathbf{H}_a^{t\,*}\right\} \sim Gamma(N_0, 1)$. Hence, the random variable u is defined as the weighted sum of Gamma random variables, hence the exact CDF of u can be given by [59]

$$P_o = \left(\frac{\beta_b^2 c_b^2}{\beta_a^2 c_a^2}\right)^{N_0^2} \times \left[\frac{2^{(\mathcal{R}_a+\mathcal{R}_b)}-1}{\beta_b^2 c_b^2 \rho\theta^2}\right] {}_2F_1\left(2N_0, N_0; 2N_0; \left(1-\frac{\beta_b^2 c_b^2}{\beta_a^2 c_a^2}\right)\right), \tag{4.14}$$

where ${}_{[z]2}F_1\left(a, b; c; d\right)$ is the incomplete version of the Gauss hypergeomtric function. We would like to note that since $\mathbf{H}_k^t$ is Toeplitz lower triangular matrix, we can have

$$\begin{aligned}\text{Tr}\left\{\mathbf{H}_k^t \mathbf{H}_k^{t\,*}\right\} &= N_c|h_0^k|^2 + (N_c-1)|h_1^k|^2 + (N_c-2)|h_2^k|^2 + \cdots + (N_c-\nu_k)|h_{\nu_k}^k|^2 \\ &\overset{(e)}{\approx} N_c \sum_{\ell=1}^{\nu_k}|h_\ell^k|^2 \overset{(f)}{\approx} N_c\end{aligned} \tag{4.15}$$

where ν_k is the delay spread of the channel between CH and user k with number of channel taps equal to $\nu_k + 1$. When N_c is much greater than ν_k, we can approximate $N_c - \nu_k$ as N_c. Hence, the approximation (e) holds. On the other hand, the approximation (f) holds from law of large numbers, assuming each channel tap has variance equal to $\sigma_{b,n}^2 = 1/\nu_b$. This leads to the nice fact that $\beta_a^2 c_a^2 \text{Tr}\left\{\mathbf{H}_a^t \mathbf{H}_a^{t\,*}\right\} + \beta_b^2 c_b^2 \text{Tr}\left\{\mathbf{H}_b^t \mathbf{H}_b^{t\,*}\right\}$ is almost deterministic with approximated value $\beta_a^2 c_a^2 \text{Tr}\left\{\mathbf{H}_a^t \mathbf{H}_a^{t\,*}\right\} + \beta_b^2 c_b^2 \text{Tr}\left\{\mathbf{H}_b^t \mathbf{H}_b^{t\,*}\right\} = N_c[\beta_a^2 c_a^2 + \beta_b^2 c_b^2]$. Hence, we can mitigate any outage event by adjusting $\mathcal{R}_a + \mathcal{R}_b$ such that $N_c[\beta_a^2 c_a^2 + \beta_b^2 c_b^2] \geq \frac{2^{(\mathcal{R}_a+\mathcal{R}_b)}-1}{\rho\theta^2}$. Or the other parameters based on what is fixed and

what is variable.

Consumed Energy at NOMA users

The main objective of data compression is to save scarce energy at the NOMA users, which could be IoT nodes with limited access to power source and/or battery operated. The energy consumed by users a and b to transmit their compressed and encoded bits can be given by

$$E_a = E_a^e + E_a^t, \tag{4.16}$$

$$E_b = E_b^e + E_b^t, \tag{4.17}$$

where E_a^e and E_b^e are the energies consumed for encoding at users a and b, respectively and E_a^t and E_b^t and the energies consumed during transmission of the encoded samples at users a and b, respectively. E_a^t and E_b^t can be written in terms of the rates R_a and R_b already defined in (4.7) and (4.8), respectively, as

$$E_a^t = \ell_a \beta_a^2 \frac{P_t\ \kappa_a}{R_a}, \tag{4.18}$$

$$E_b^t = \ell_b \beta_b^2 \frac{P_t\ \kappa_b}{R_b}, \tag{4.19}$$

where κ_a and κ_b are lengths of the transmitted bits per hertz within a time slot at users a and b, respectively. The encoding energy comprises the energy consumed during DWT and quantization and encoding steps. The encoding energies at a and b can be written

as [15]

$$E_a^e = E_a^d + E_a^q, \tag{4.20}$$

$$E_b^e = E_b^d + E_b^q, \tag{4.21}$$

where E_a^d and E_b^d are the energies consumed during DWT operation at users a and b, respectively and E_a^q and E_a^q are the energies consumed during quantization at users a and b, respectively. E_a^d and E_b^d can be written as [15]

$$E_a^e = F_a\ N_a \left(\sum_{n=0}^{D_a} \frac{1}{2^n} \right) E_{comp} + N_a(1-k_a)E_{cs}, \tag{4.22}$$

$$E_b^e = F_b\ N_b \left(\sum_{n=0}^{D_b} \frac{1}{2^n} \right) E_{comp} + N_b(1-k_b)E_{cs}, \tag{4.23}$$

where F_a and F_b are the lengths of the filters used to implement DWT at users a and b, respectively. D_a and D_b are the number of DWT decomposition levels at users a and b, respectively. E_{comp} is consumed energy per computation and E_{cs} is the energy consumed at each analog to digital conversion step as in [60].

Harvested Energy

The CH harvests energy in the analog domain before signal decoding. The harvested energy can be calculated as

$$E_h = \frac{\eta \bar{\theta}^2 P_t}{N_c} \Big[\ell_a T_a \beta_a^2 c_a^2 \text{Tr}\left\{ \mathbf{H}_a^t \mathbf{H}_a^{t\,*} \right\} + \ell_b T_b \beta_b^2 c_b^2 \text{Tr}\left\{ \mathbf{H}_b^t \mathbf{H}_b^{t\,*} \right\} \Big], \tag{4.24}$$

where $\bar{\theta} = 1 - \theta$, T_a and T_b are the harvesting energy times at users a and b, respectively, and η is efficiency of the RF conversion process. Assuming normalized power profile with equal-power taps, for channel between CH and user $k \in \{a, b\}$, we have $E\{|h_i^k|^2\} = 1/(\nu_k + 1)$. Thus, we have

$$\begin{aligned} E\left\{\text{Tr}\left\{\mathbf{H}_k^t \mathbf{H}_k^{t*}\right\}\right\} &= \frac{1}{\nu_k + 1}\Bigg(N_c + (N_c - 1) + (N_c - 2) \\ &\quad + \cdots + (N_c - \nu_k)\Bigg) \\ &= \frac{1}{\nu_k + 1}(\nu N_c - \sum_{n=1}^{\nu_k} n) \\ &= \frac{1}{\nu_k + 1}(\nu_k N_c - \frac{\nu_k(\nu_k + 1)}{2}) \\ &= \frac{\nu_k}{\nu_k + 1}(N_c - \frac{(\nu_k + 1)}{2}) \\ &\approx \frac{\nu_k N_c}{\nu_k + 1} \end{aligned} \tag{4.25}$$

where the last approximation occurs when N_c is significantly larger than the delay spread, which is the typical case. In addition, this could be approximated to N_c. Substituting into the EH expression, we get the expectation as

$$\begin{aligned} E\{E_h\} &= \frac{\eta\bar{\theta}^2 P_t}{N_c}\Bigg[\ell_a T_a \beta_a^2 c_a^2 \frac{\nu_a}{\nu_a + 1}(N_c - \frac{(\nu_a + 1)}{2}) \\ &\quad + \ell_b T_b \beta_b^2 c_b^2 \frac{\nu_b}{\nu_b + 1}(N_c - \frac{(\nu_b + 1)}{2})\Bigg], \\ &\approx \frac{\eta\bar{\theta}^2 P_t}{N_c}\Bigg[\ell_a T_a \beta_a^2 c_a^2 (N_c - \frac{(\nu_a + 1)}{2}) \\ &\quad + \ell_b T_b \beta_b^2 c_b^2 (N_c - \frac{(\nu_b + 1)}{2})\Bigg] \\ &\approx \frac{\eta\bar{\theta}^2 P_t}{N_c}\Bigg[\ell_a T_a \beta_a^2 c_a^2 N_c + \ell_b T_b \beta_b^2 c_b^2 N_c\Bigg] \\ &\approx \eta\bar{\theta}^2 P_t\Bigg[\ell_a T_a \beta_a^2 c_a^2 + \ell_b T_b \beta_b^2 c_b^2\Bigg] \end{aligned} \tag{4.26}$$

where in the first approximation we assumed that $\frac{\nu_k}{\nu_k+1} \approx 1$, $k \in \{a, b\}$. On the other hand, the second approximation (last approximation) is based on the assumption that $N_c \gg \nu_k/2$.

Optimization Problem

We optimize

$$\begin{aligned}
\mathbf{P_2}: \max_{\beta_a,\beta_b,\theta,k_a,k_b} &: \mathbb{E}\left\{E_h\right\}, \\
\text{s.t.} \quad & P_o \leq \epsilon, \\
& E_a \leq \delta_a, E_b \leq \delta_b, \\
& D_a \leq \mu_a, D_b \leq \mu_b \\
& 0 \leq \beta_a, \beta_b, \theta \leq 1, \\
& d_a \in \mathbf{d}_A,\ d_b \in \mathbf{d}_B, \\
& d_a < d_b.
\end{aligned} \tag{4.27}$$

where $\mathbf{d}_A$ and $\mathbf{d}_B$ are the vectors that contain the distance values for the sets of near and far user discs, respectively. Some Remarks on Eqn. (4.27): There are some trade-offs in (4.27) as follows

- $\mathbf{d}_A$ and $\mathbf{d}_B$ contain discrete values, which implies that this constraint is non-convex, and yields that the optimization problem above is non-convex.
- The higher the compression ratio, the lower the harvested energy at the CH.
- The higher the compression ratio, the lower the consumed energy at the transmitting IoT nodes.

- The CH can adjust the compression ratio for each user based on each desired value.
- The higher the PS factor, the higher the harvested energy, but the lower the SNR and hence the higher the outage probability and the distortion ratio.
- The higher the NOMA power split factor, the higher the consumed energy at the transmitting nodes, and the higher the harvested energy at the CH.

Regarding the solution of (4.27), the optimization problem is non-convex due to the non-convexity of the objective function and the constraints. Hence, we cannot exploit existing efficient methods used to solve convex problems.

Using exhaustive search (also known as brute force) is not the most efficient way to solve the problem. However, exhaustive search is typically used to benchmark other solving methods, is easy to implement and will always find a solution if it exists. Please note that some of our optimization variables are discrete values with small set, and even the range of the continuous ones could be divided into O points and the optimization problem could be evaluated accordingly. Increasing the value of O increases both accuracy of the solution as well as the complexity. In addition, we optimize the average performance, the optimization problem will be resolved only when the average/statistical parameters are changed, which does not occur frequently.

Due to the high degree of complexity of the optimization problem in (4.27), we propose to use DRL approach to find the optimized values as described the following Section.

Reinforcement learning peers the input data with the delayed reward value making the agent take actions that lead to higher rewards. When a new data point is fit for training, the agent can see the reward from the previous data point. Therefore, reinforcement learning can usually be regarded as 5 tuple $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma\}$, where $\mathcal{S}$ is the state's space, $\mathcal{A}$ is the actions space, $\mathcal{T}$ is the state transition values as a function of given action a in a given state, $\mathcal{R}$ is the reward and γ is the discount factor that adjust the action.

Agent and Environment

The environment should be designed to describe the main parameters of the system that will interact with the agent per each time step t. According to the problem description, the environment $\mathbb{E}$ will have a continuous attitude as the episode keeps running with no break state. The agent's behavior will be described by a policy π, which maps states $S_1, S_2, \ldots, S_n \in \mathcal{S}$ into a given actions $a_1, a_2, \ldots, a_n \in \mathcal{A}$ at each time step $t \in T$. During the experiment, the environment state S_t will take an action a_t from the agent according to the policy π and then generates based on the state transition dynamics, the next state $(s_t, a_t) \rightarrow S_{t+1}$ and an immediate delayed reward $R(S_t, a_t) \rightarrow r_t$ at each time step t. The total cumulative discounted future reward across the experiment starting from time $t' = t$ can be calculated by $R_t = \sum_{t'=t}^{T} r_t \gamma^{t'-t}$, where T is cumulative time for the experiment and $\gamma \in [0, 1]$ represents the discount factor. The interactive state/action value function which also called Q function of a policy π can be written as $Q^{\pi}(s, a) = \mathop{\mathbb{E}}_{r_t, s_t \sim E, a_t \sim \pi}[R_t | s_t = s, a_t = a]$, it describes how efficient it is for the agent to perform a specific action in a state with the policy π and by replacing R_t with its

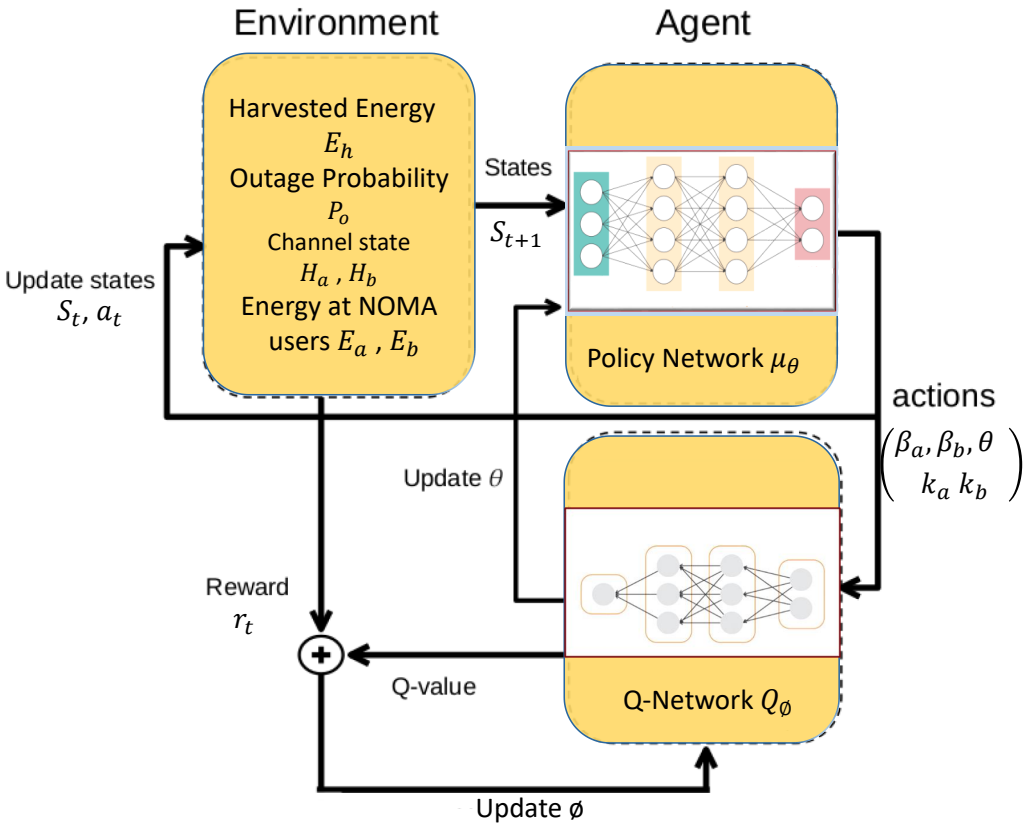

Figure 4.2: DDPG System environment/agent Architecture model 2

value the Q-function will be $Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{t'=t}^{T} r_t \gamma^{t'-t} | S_t = s, a_t = a\right]$, the main objective of the agent is to learn an optimal policy π' that harvest the optimal Q-function $Q'(s,a) = \max_{\pi} Q^{\pi}(s,a)$ and this will be the policy of learning.

Again, using Q–learning directly in continuous spaces of action is difficult to apply because it is necessary to optimize greedy policies at every step of the time. Such kind of optimization is too slow to be realistic with large or unconstrained problems and non-trivial action spaces.

Proposed DDPG-based approach to maximize harvested energy

Based on our objective function in (4.27), we propose the deep deterministic policy gradient (DDPG) algorithm as a practical optimizer to imply that DRL can be used in

such optimization problem and to benchmark its results with our Matlab simulation.

Architecture of the proposed DDPG-based approach

The complete architecture of the DDPG system is shown in Figure 4.2. The main objective is to learn a deterministic policy $\mu(s, \theta)$ that allows the actions to maximize $\mathbb{Q}(s_t, a_t; \phi)$. The main purpose of parameter ϕ is for policy evaluation. The training process will utilize the replay buffer that represents the previous experience of learning which also, can improve the data and stabilize the training process of the neural networks according to [61]. The mean square Bellman error (MSBE) represents the the error function which indicate how far the approximation from satisfying Bellman equation. Each network will have a time-delayed copy of itself in order to stabilize the minimization process of MSBE during the training.

The actor neural network would maintain the parametric actor function $\mu(s|\theta^{\mu})$ that specifies the current policy by mapping states to a specific action deterministically. However, the critic network $\mathbb{Q}(S, a)$ will learn using the Bellman equation which describes the optimal action value function [56], [57] :

$$\mathbb{Q}^{\mu}(s_t, a_t) = \mathop{\mathbb{E}}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma \max_{a_{t+1}} \mathbb{Q}^{\mu}(s_{t+1}, a_{t+1})],$$

where $a_{t+1} \rightarrow \mu(s_t|\theta^{\mu})$ and s_{t+1} implies that the next state is sampled by the environments following the distribution $P(\cdots|s, a)$. Since the function approximator $\mathbb{Q}(s_t, a_t; \phi)$ is assumed to differential with respect to the moving action statement, which means for one policy $\mu(s; \theta)$ we can create a gradient based learning rules that minimizes the expensive computation of $\max_{a_t} \mathbb{Q}(s_t, a_t))$ over the action space to be

$\mathbb{Q}(s_t, \mu(s_t; \theta)$.

Agent and environment of our optimization problem

The main goal is to maximize the objective function in (4.27) by maximizing its reward function in order to obtain the trade off between maximum harvested energy satisfying the main constrains listed in the problem including the compression and distortion ratios, outage probability and consumed energy by the NOMA users. The Markov Decision Processes (MDPs) shall be considered while designing the environment E. The MDPs shall model the relations between the agent and the environment while the environment changes continuously. The environment will be episodic, where the episodes represents the dynamicity of the environment changes. The environment will be fully observed by the agent and the state s_t will be $[E_h, P_o, E_a, E_b, \mathbf{H_a}, \mathbf{H_b}] \in \mathcal{S}$. All of these parameters have to be normalized to train the network's nodes, therefore, the state will be $s_t = [\hat{E}_h, \hat{P}_o, \hat{E}_a, \hat{E}_b, \hat{H}_a, \hat{H}_b], \forall s \in \mathcal{S}$. We can notice that, the state transitions are not deterministic in the system due to the fact that all of these parameters depends on the randomness of the channel estimation. Nevertheless, the state transition can still be calculated by invoking our optimization parameters $[\theta, \beta_\mathbf{a}, \beta_\mathbf{b}, \mathbf{k_a}, \mathbf{k_b}]$ where $[a_t = \theta, \beta_\mathbf{a}, \beta_\mathbf{b}, \mathbf{k_a}, \mathbf{k_b}] \in \mathcal{A}$ represents the action space of the system. The agent will generate these actions and invoke them with the environment to calculate the next state and get the discounted reward based on the previous state.

Reward function

The reward function at each time step t describes the main parameters of the optimization problem and it is a function of the current state s_t and current action a_t.

The aim of the problem is to maximize the harvested energy E_h while satisfying the main constrains. Therefor, E_h must be involved in the reward function and jointly we need to optimize the compression ratio while the distortion of the data is kept below the threshold. The reward function has to include all the system constrains. From that the reward function is

$$r_t = \begin{cases} \lambda_1 E_h + \lambda_2(\delta_a - E_a) + \\ \quad \lambda_3(\delta_b - E_b) + \\ \quad \lambda_4(\mu_a - D_a) + \lambda_5(\mu_b - D_b) & \text{if remaining constraints in (4.27) hold} \\ -1, & \text{otherwise} \end{cases} \tag{4.28}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$, are the weights for each term and $\lambda_1+\lambda_2+\lambda_3+\lambda_4+\lambda_5 = 1$. The condition for the reward is the constraints listed in the problem above in order to ensure the parameters do not go below a certain lower bounds. Otherwise, the reward will be be penalized by -1. The weights above play important role in the system performance to obtain the trade-off between the optimization parameters. These weights could be adjusted based on the system requirements.

DRL Conversion

During the first 100 episodes, the algorithm had an exploration decay rate of $\phi = 0$, which means that during those episodes, the entire experiment was observed. This implies that the algorithm analyzes the entire continuous space of actions to determine the most rewarded actions required to establish the optimal strategy, thus optimizing the total reward. Exploration is achieved by adding noise to the action itself, as we have continuous action spaces. Subsequently, the entire exploration term decreases to almost 0, such that maximum exploitation is achieved and thus seeks actions that only yield the greatest possible rewards. For all the channel randomness, the efficiency has stabilized, converging approximately after 1500 episodes, i.e. the algorithm achieved an optimal policy to obtain the highest reward. We present the average reward of the proposed D-DDPG algorithm in Figure 4.11, which takes immediate changes in the environment into account as a function of channel gain and node energy consumption. Therefore, as the environment changes, it eliminates numerous needs for re-optimization.The main simulation parameters set to be similar in both Matlab simulation and the DRL model. After reaching the conversion point, the model start to fine tune the action parameters to maximize the total reward in a slow learning rate fashion. We can see this from the slit increase of the reward values with time. The optimal parameters to attain highest harvested energy were when β_a and β_b were at maximum which is 0.9 each and θ was low to 0.1. The compression ratio was zero at node a and 0.75 at node b.

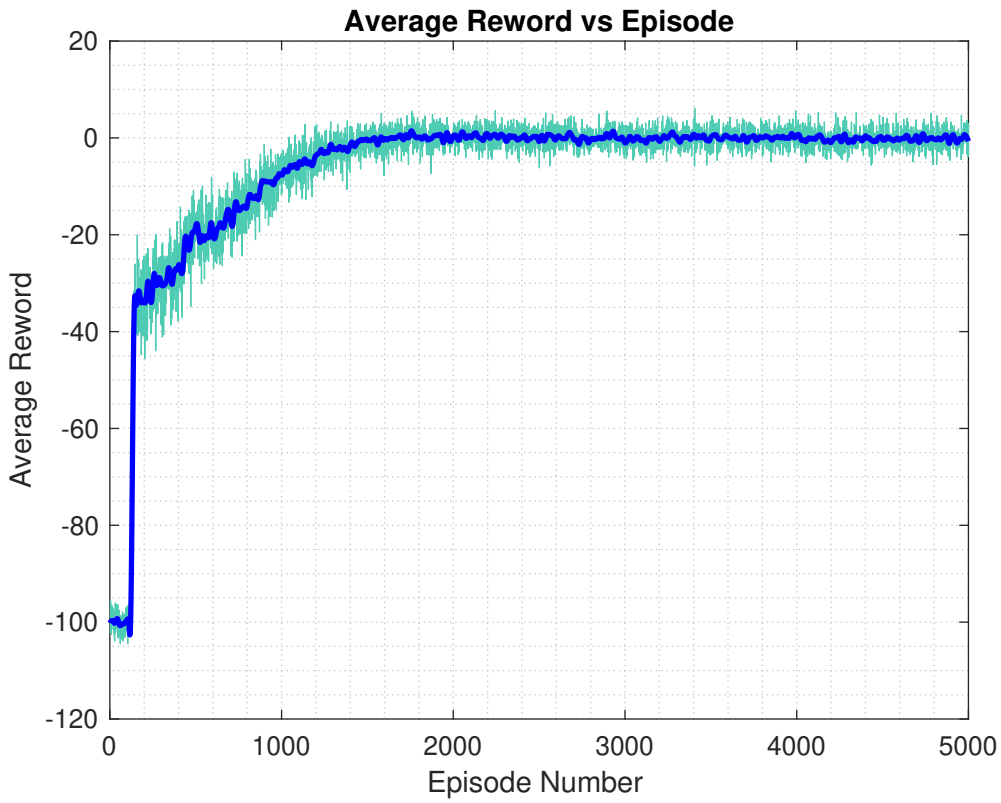

Figure 4.3: DRL Result .

Results Analysis

Comprehensive simulations have been conducted to figure out the effect of each single parameter on the performance of the system. Simulation of the NOMA SWIPT Method Using fading channel realizations based on Rayleigh channel model that obey $\mathcal{CN}(0,1)$ for each user. The main simulation parameters set to be $P_t = 1$ watt and $\sigma^2 = 1 * e - 8$,$N_c = 64, \alpha = 2$. The NOMA users are organized in rings around the cluster head node with different distances disc $A_j \in [5 : 10]$ and disc $B_j \in [11 : 20]$ meters respectively.the objective parameters have been set to be $0 \leq \beta_a, \beta_b, \theta, k_a, k_b \leq 1$ as shown in equation 4.27. The numerical simulation results have been presented in Figures 4.4 - 4.10.

Some Remarks on the simulated results:

- In order to tackle the trade-off between the total energy harvesting (EH) and θ and the effect of β_a, β_b on the harvested energy, we plotted the harvested energy E_h versus θ for different values of β_a, β_b, meanwhile we kept the compression ratios constant at this point. In Figure 4.4, we show the relation between the EH E_h versus θ, where θ is always split between signal decoding and harvesting energy at the cluster head node. As we showed in equation [4.28], the amount of harvested energy decreases when the allocated decoding power θ from the CH increases as this will decrease the amount of power allocated to harvest energy. The second important set of parameters in this curve are β_a and β_b representing the allocated power ratio for transmission for node a and b respectively. As the transmitting nodes spend more power in each transmission the CH will be able to harvest more energy. The highest harvest power was achieved when the node transmission power was at its maximum. Conversely, the energy harvested decreases β_a, β_b decreases. Moreover, The NOMA user pairing is important factor and always have effect on the energy harvesting performance. The best performance was always achieved when we pair the node at the edges of the discs. For example when user $a \in$ disc A_j which have the nearest distance from the CH node is paired with user $b \in$ discB_j which have the nearest distance from the CH node in disc B_j, this gives the highest harvested energy in most of the case and depends on the values of β_a, β_b. This implies that, it important to pair the between the nodes on the edges closest to the CH node. And on the contrary, Pairing the meddle nodes is will not preserve the highest harvested energy.

- On the other hand, to study the effect of the compression ration on the harvested energy, we presented in Figure 4.5, the relation between the harvested energy versus θ and we kept β_a, β_b fixed. As presented in the graph, the harvested energy always depends on the transmitted number of samples. The highest results shows that without data compression we get more energy. This behavior holds for the two right sub figures in the graph as the total number of transmitted sampled when user a transmit 90 percent of its samples and user b transmit only 25 percent or vice versa, this means the total transmitted samples is above 50 percent. However, in the left down corner sub figure both users transmit only 50 percents of the total number of samples and therefore, lowest harvested energy.

- In Figure 4.6, we present the relation between the outage probability and the CH power split factor θ with respect to the variation in the values of β_a, β_b. A threshold level ϵ has been fixed to measure any outage event of the users. And as we presented in equation [4.19] that we can mitigate any outage event by adjusting $\mathcal{R}_a + \mathcal{R}_b$ such that $N_c[\beta_a^2 c_a^2 + \beta_b^2 c_b^2] \geq \frac{2^{(\mathcal{R}_a+\mathcal{R}_b)}-1}{\rho\theta^2}$. Or the other parameters based on what is fixed and what is variable. Therefore, as β_a, β_b varies we presented the outage events in different case. when β_a, β_b are high grantee no outage event will occurred and as both of them goes low there will be a possibility of outage for some users especially when pairing the meddle nodes. the factor that controls this behavior is the targeted date rate $\mathcal{R}_a + \mathcal{R}_b$. The target data rate can be adjusted based on the application requirement. We notice from the graph that at higher average transmission power of the NOMA users, we satisfy the outage condition for the majority of user pairing, except pairing the middle users in disc A_j with the farthest of disc B_j.

- The node's consumed energy is always function of the transmission power and the compression ration. The transmission power per each transmission is dominated by the compression ratio. The compression ratio reduces the power consumption as it goes high. As per the system model we assumes that the transmission power ratio is always controlled by the CH node. however, we can notice the effect of the compression ratio on the node consumption for user a in Figure 4.7. As the compression ration goes high, we notice a considerable decrease of the consumed energy based on the values of β_a and distance from the CH. Since we fixed the transmission with each time slot to N_c samples so the compression ratio will determine the total number of slots that required to transmit the current available data. Both of θ and the distance from the CH have minimum impact on the node's energy consumption.

- At Node b, the energy consumption is shown in Figure 4.8. Because node b is the far user we can notice the effect of the value of distance from the CH. Compression ratio and β_b values have the same impact as for node a on the consumed energy. The lowest consumption was when β_a and β_b was low for both nodes.

- Figure 4.9 represents the relation between compression ratio versus distortion for user a, while θ is constant with varying β_a. Since the signal is transmitted over an error channel, the SNR between the transmitter and receiver is the main factor of distorting the signal. Since user a (the near user) is always suffering interference from the far user b, the signal received from user a will face much distortion than the signal coming from user b. The value of β_a could minimize the level of distortion as β_a have a higher value. In Figure 4.10, we can see the

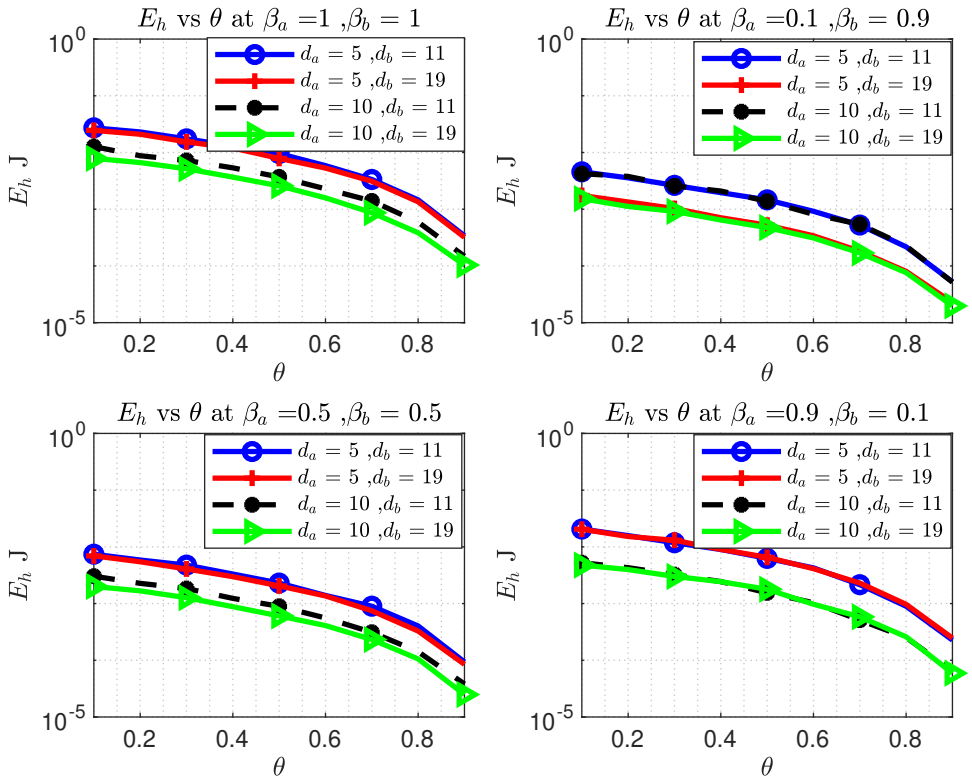

Figure 4.4: Energy Harvesting vs θ with constant compression ratio

level of distortion is below the threshold level in all cases due to the higher SNR between this node and the receiver. The DRL solution shows better performance then the greedy solution where the greedy solution is greedy with respect to time. Figure 4.11 shows the average reward versus time for both of the greedy solution and the DRL solution. The average reward that achieved from the DRL is higher significantly the the average reward achieved from the greedy solution. Because of the target of the DRL is to maximize the expected harvested energy, we can in figure 4.12 the average harvested energy gained from the DRL is higher significantly than what we get from the greedy solution.

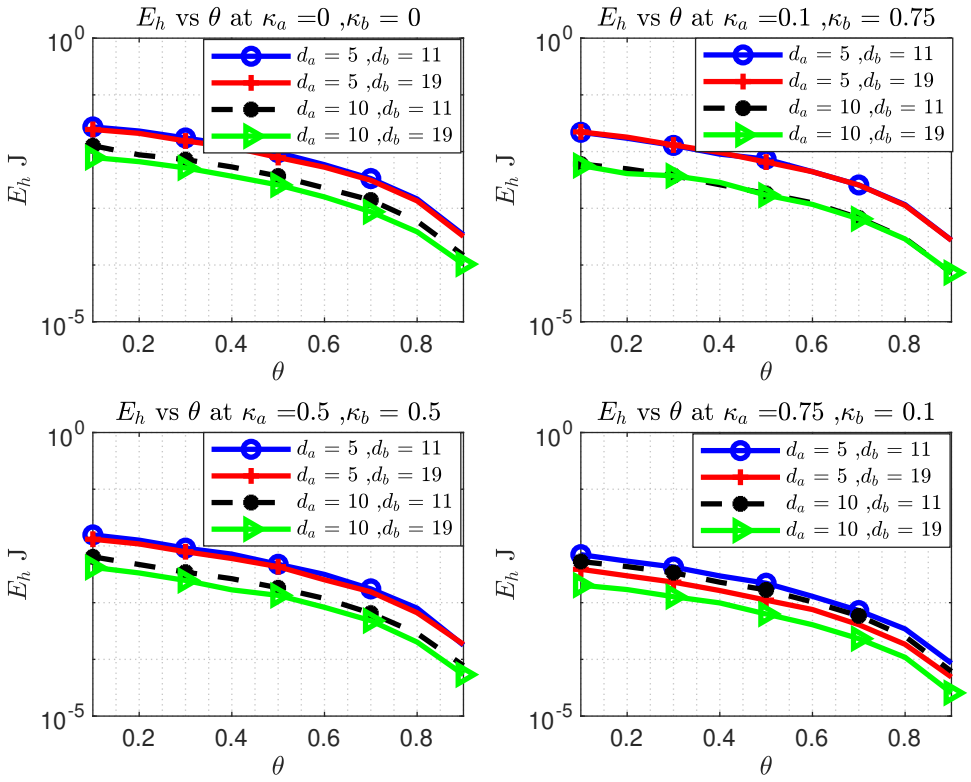

Figure 4.5: Energy harvesting vs θ with constant β

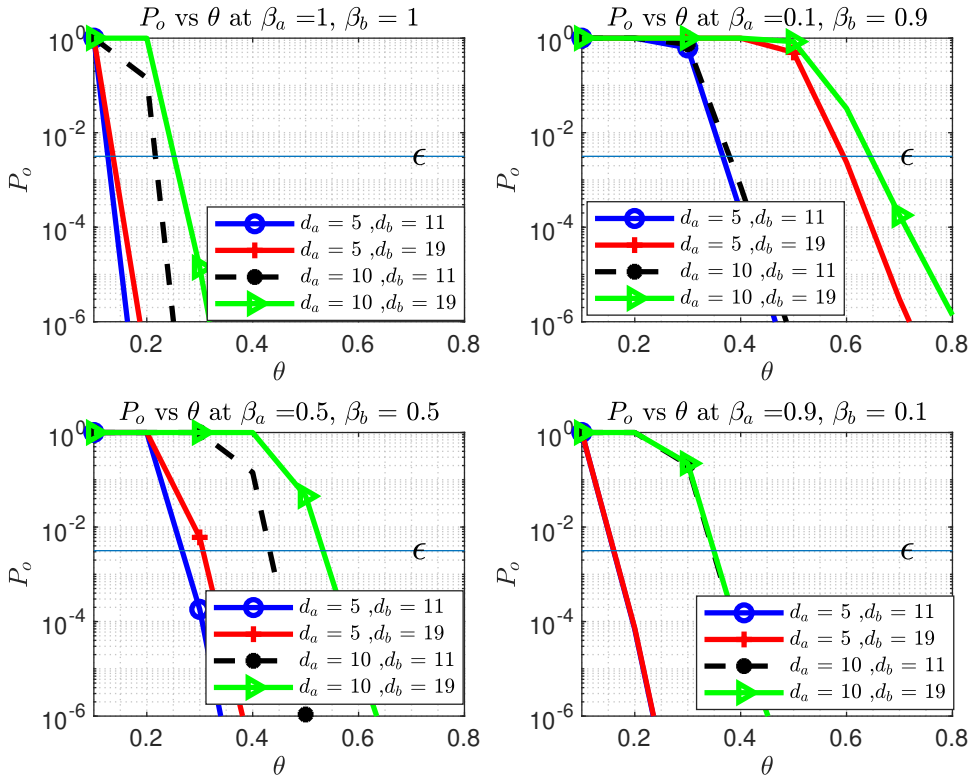

Figure 4.6: Outage probability vs θ.

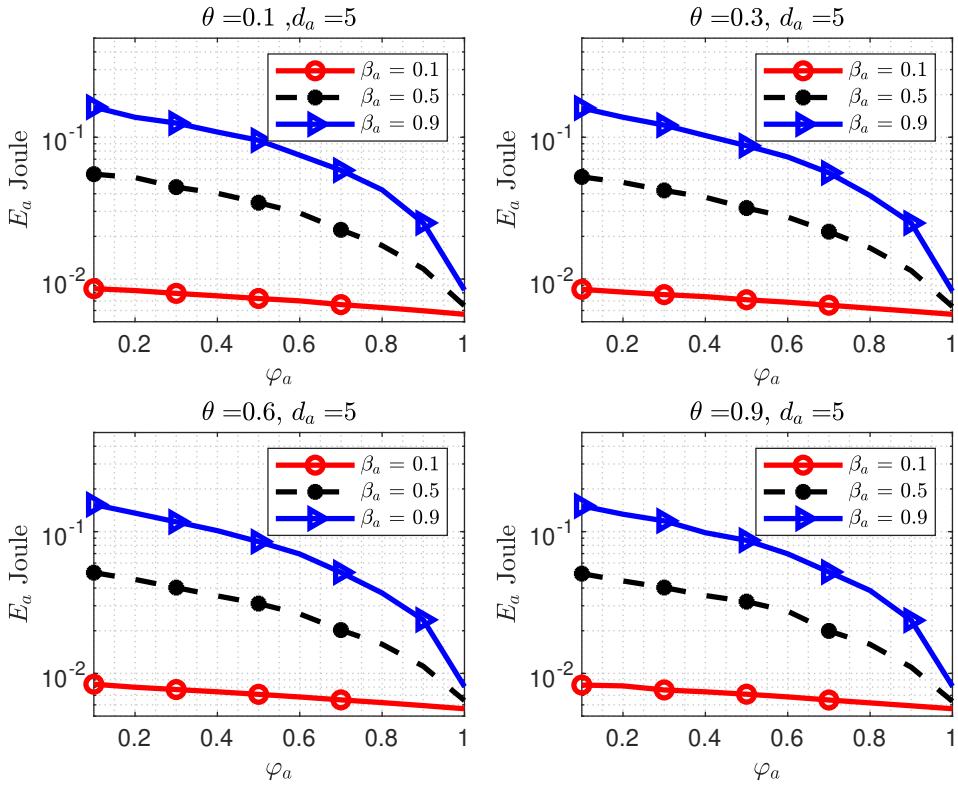

Figure 4.7: Node a energy consumption vs Compression Ratio .

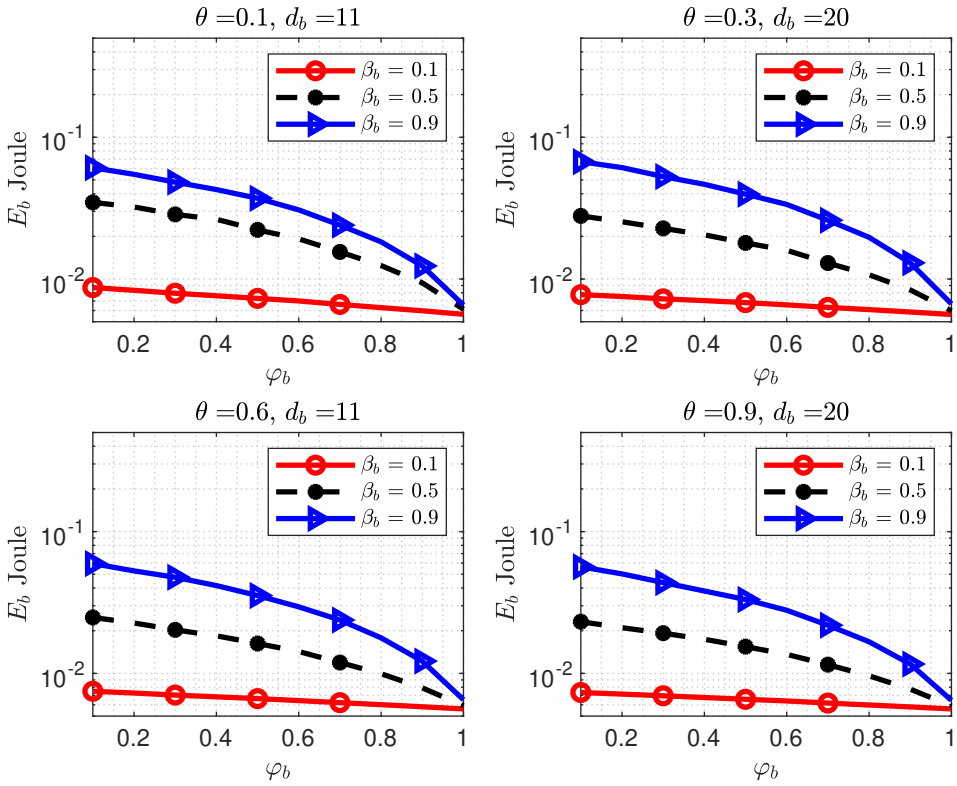

Figure 4.8: Node b energy consumption vs Compression Ratio.

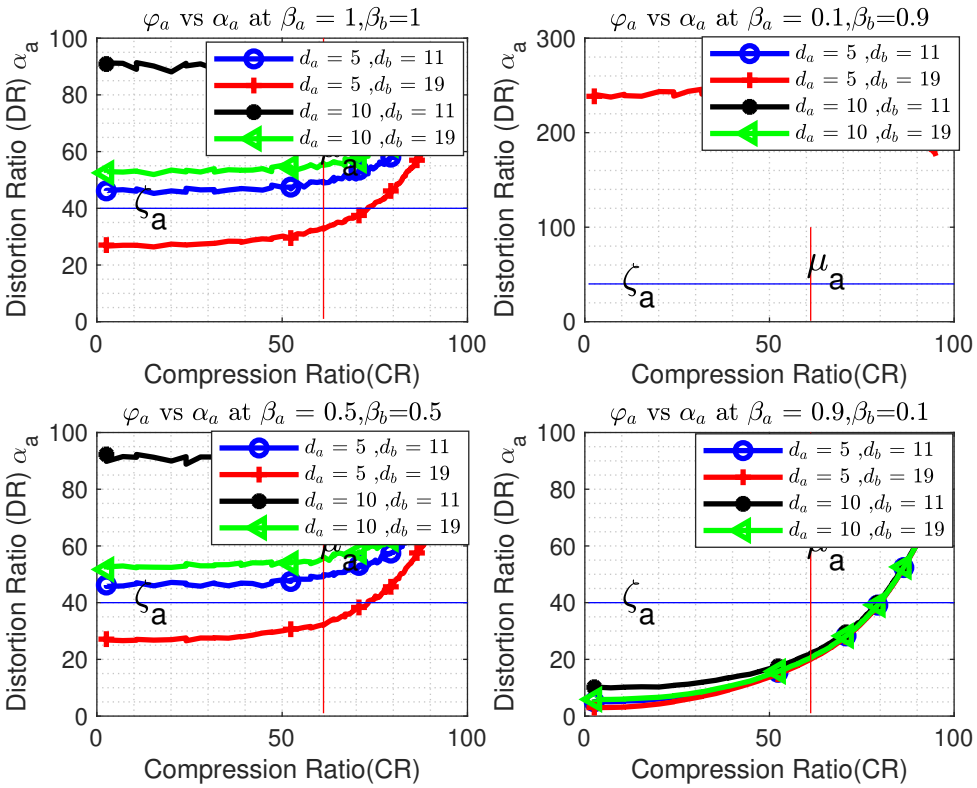

Figure 4.9: Distortion Ratio vs Compression Ratio for user a .

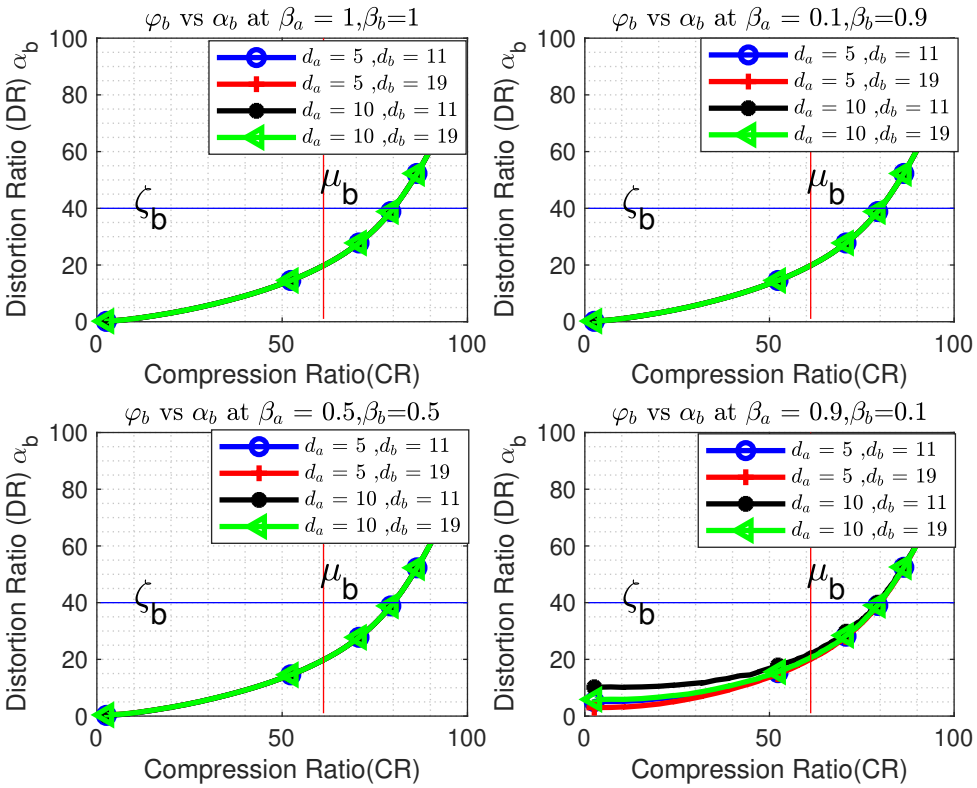

Figure 4.10: Distortion Ratio vs Compression Ratio for b .

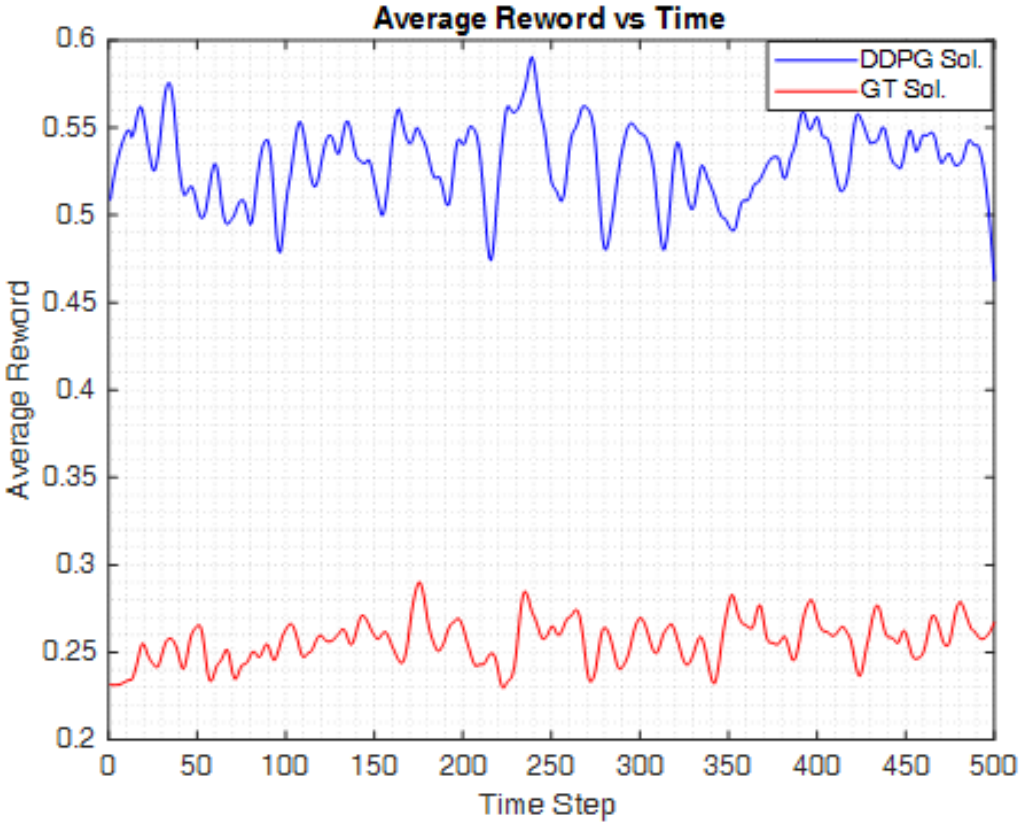

Figure 4.11: Average reward Vs. Time.

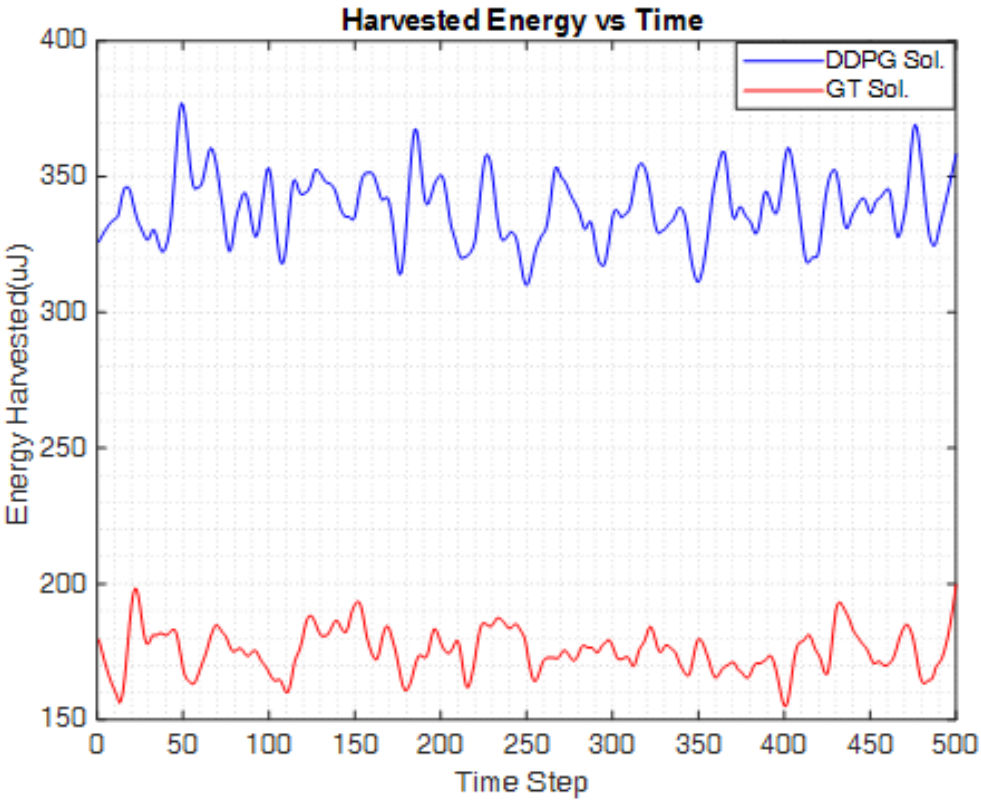

Figure 4.12: Average harvested energy Vs. Time.

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

In this thesis work, we investigated the scope of integrating multiple IoT nodes capable of smart data compression in a system coordinated through a cluster head node capable of harvesting energy from the wireless signal under NOMA up-link protocol for smart and efficient data transmission. The main objective was to identify the desired trade-off between the system parameters in order to archive higher performance in terms of data compression with acceptable level of distortion, meanwhile gain the highest possible harvested energy to cover the power expenses of the cluster head node. The system architecture is designed to satisfy the applied constraints in terms of outage probability of the transmitting node and target rate to minimize out probability. In this regards, we introduced two different system models described in chapter 3 and chapter 4 respectively.

Firstly, we considered wireless remote monitoring sensor networks, with an energy-limited nodes, transmitting a collected data from the surrounding environment, such as medical EEG data to an edge node under NOMA-up-link protocol. We presented a smart model for data compression with minimal expected distortion among NOMA users using the DRL approach. The problem was formulated in the form of satisfying the constraint of node's outage probability, pre-transmission signal compression, and considering the power budget constraint of the user nodes. The proposed DDPG agent was able to learns the optimal policy that leads to a trade-off between the users' minimum expected distortion and the probability of outage by achieving the highest possible reward. the agent was designed to adjust the node's transmission power and the desired compression ratio.

The results in this part show the effect of both NOMA power split factor and compression ratio on the total expected distortion. The total distortion due to data encoding and quantization reached its maximum level when compression ratio was at the highest value. In addition, the higher the compression ratio, the lower the energy consumption of the node. Also, the NOMA power split factors could minimize the effect of the noisy channel on the distortion, such that the higher NOMA power split factor the higher energy consumption. Finally, the outage events can be controlled by the targeted data rate threshold.

Secondly, we expanded the system model to investigate the energy harvesting requirements to grant the cluster head node longer life time. The system was designed to include multiple nodes arranged in discs surrounding the cluster head node. We studied the visibility of pairing users in order to gain the maximum harvested energy from the RF signal. we investigated the effect of each single system parameter on the performance of the harvested energy and outage probability. Moreover, we investigated the energy consumption at each node in order to increase the battery life time of these nodes. A second dynamic DRL model was designed to determine the optimal system parameters that allow the system to operate under the highest possible performance.

Compiling the generated results from DRL and the optimization-based solution would lead to understanding the effect of each parameter on the performance. The harvested energy is higher when the transmitted number of samples is higher. However, NOMA power split factors β_a and β_b will have the main impact on the harvested energy as the highest harvested energy was achieved when both of them were to the maximum. The outage probability could be mitigated by controlling the threshold of the targeted data rate. Both of NOMA power split factors and compression ratio as well as the targeted

data rate have impact on the node energy consumption especially when the values of β_a and β_b are high. Finally, the level of distortion depends on both compression ratio and SNR between the transmitter and the receiver, such that the higher the SNR the lower the distortion.

Future Work

- The system model tackled in this thesis work was based on a group of IoT devices connecting to one cluster head (CH). As a future work, we will look into large systems utilizing multiple cluster heads or multi-base station systems. The optimization problem in this case will be highly complex, and can be mapped to multi-agent reinforcement learning, where the agents at the base stations can compete or cooperate to achieve the best policy that optimizes the global system reward.

- In our model we assumed that the nodes are saturated and always have data to send during the transmission time. The model can be expanded to study the performance of probabilistic effect of the data arrival distribution on the performance of the NOMA system, because this will have impact on the interference with the other nodes.

- We assumed in our model that the transmission rate is fixed to N_c value during each transmission. Therefore, the model can investigate the transmission of various data rate per time slot in the future, and the effect of that on the harvested energy.

- We can also consider certain energy constraint per slot on the source, then compute processing time due to enabling DWT (or any other filters). This will allow us

to use the remaining energy for data transmission. Since data transmission power impact both data rates and EH rates, the compression ratio would impact both.

- We can expand the model to include mutable CH nodes and use DRL to study the feasibility of increasing the life time of these nodes based on priority of data transmission.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (noma) for cellular future radio access," in *2013 IEEE 77th vehicular technology conference (VTC Spring)*, IEEE, 2013, pp. 1–5.

[2] T. RAN, "Requirements for further advancements for e-utra (lte-advanced),"," *June 200S*, 2008.

[3] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[4] E. Nadolski, *Virtualized cluster communication system*, US Patent 8,707,083, Apr. 2014.

[5] A. B. Noel, A. Abdaoui, T. Elfouly, M. H. Ahmed, A. Badawy, and M. S. Shehata, "Structural health monitoring using wireless sensor networks: A comprehensive survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 3, pp. 1403–1423, 2017.

[6] X. Chen, A. Benjebbour, A. Li, and A. Harada, "Multi-user proportional fair scheduling for uplink non-orthogonal multiple access (noma)," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*, IEEE, 2014, pp. 1–5.

[7] S. Kim, R. Vyas, J. Bito, K. Niotaki, A. Collado, A. Georgiadis, and M. M. Tentzeris, "Ambient rf energy-harvesting technologies for self-sustainable standalone wireless sensor platforms," *Proceedings of the IEEE*, vol. 102, no. 11, pp. 1649–1666, 2014.

[8] M. H. Alsharif and R. Nordin, "Evolution towards fifth generation (5g) wireless networks: Current trends and challenges in the deployment of millimetre wave, massive mimo, and small cells," *Telecommunication Systems*, vol. 64, no. 4, pp. 617–637, 2017.

[9] S. K. Sharma, M. Patwary, and S. Chatzinotas, "Multiple access techniques for next generation wireless: Recent advances and future perspectives," *EAI Endorsed Transactions on Wireless Spectrum*, vol. 2, no. 7, 2016.

[10] M. Liaqat, K. A. Noordin, T. A. Latef, and K. Dimyati, "Power-domain non orthogonal multiple access (pd-noma) in cooperative networks: An overview," *Wireless Networks*, pp. 1–23, 2018.

[11] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5g networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, 2017.

[12] S. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2016.

[13] H. Tabassum, M. S. Ali, E. Hossain, M. J. Hossain, and D. I. Kim, "Uplink vs. downlink noma in cellular networks: Challenges and research directions," in *2017 IEEE 85th vehicular technology conference (VTC Spring)*, IEEE, 2017, pp. 1–7.

[14] M. Al-Imari, P. Xiao, M. A. Imran, and R. Tafazolli, "Uplink non-orthogonal multiple access for 5g wireless networks," in *2014 11th international symposium on wireless communications systems (ISWCS)*, IEEE, 2014, pp. 781–785.

[15] A. Awad, M. Hamdy, A. Mohamed, and H. Alnuweiri, "Real-time implementation and evaluation of an adaptive energy-aware data compression for wireless eeg monitoring systems," in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, IEEE, 2014, pp. 108–114.

[16] A. A. Abdellatif, M. G. Khafagy, A. Mohamed, and C.-F. Chiasserini, "Eeg-based transceiver design with data decomposition for healthcare iot applications," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3569–3579, 2018.

[17] A. Arar, A. Mohamed, A. A. El-Sherif, and V. C. M. Leung, "Optimal resource allocation for green and clustered video sensor networks," *IEEE Systems Journal*, vol. 12, no. 3, pp. 2117–2128, 2018.

[18] M. Elsayed, A. Badawy, M. Mahmuddin, T. Elfouly, A. Mohamed, and K. Abualsaud, "Fpga implementation of dwt eeg data compression for wireless body sensor networks," in *2016 IEEE Conference on Wireless Sensors (ICWiSE)*, IEEE, 2016, pp. 21–25.

[19] X. Zhang, J. Grajal, J. L. Vazquez-Roy, U. Radhakrishna, X. Wang, W. Chern, L. Zhou, Y. Lin, P.-C. Shen, X. Ji, *et al.*, "Two-dimensional mos 2-enabled flexible rectenna for wi-fi-band wireless energy harvesting," *Nature*, vol. 566, no. 7744, pp. 368–372, 2019.

[20] U. Olgun, C.-C. Chen, and J. L. Volakis, "Efficient ambient wifi energy harvesting technology and its applications," in *Proceedings of the 2012 IEEE International Symposium on Antennas and Propagation*, IEEE, 2012, pp. 1–2.

[21] K. W. Choi, S. I. Hwang, A. A. Aziz, H. H. Jang, J. S. Kim, D. S. Kang, and D. I. Kim, "Simultaneous wireless information and power transfer (swipt) for internet of things: Novel receiver design and experimental validation," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 2996–3012, 2020.

[22] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[23] M. Bello, W. Yu, A. Chorti, and L. Musavian, "Performance analysis of noma uplink networks under statistical qos delay constraints," *arXiv preprint arXiv:2003.04758*, 2020.

[24] M. Pischella, A. Chorti, and I. Fijalkow, "Performance analysis of uplink noma-relevant strategy under statistical delay qos constraints," *IEEE Wireless Communications Letters*, vol. 9, no. 8, pp. 1323–1326, 2020.

[25] M. Zeng, W. Hao, O. A. Dobre, Z. Ding, and H. V. Poor, "Power minimization for multi-cell uplink noma with imperfect sic," *IEEE Wireless Communications Letters*, 2020.

[26] H. Liu, T. A. Tsiftsis, K. J. Kim, K. S. Kwak, and H. V. Poor, "Rate splitting for uplink noma with enhanced fairness and outage performance," *IEEE Transactions on Wireless Communications*, 2020.

[27] S. A. Tegos, P. D. Diamantoulakis, J. Xia, L. Fan, and G. K. Karagiannidis, "Outage performance of uplink noma in land mobile satellite communications," *IEEE Wireless Communications Letters*, 2020.

[28] H. Lu, X. Xie, Z. Shi, M. Kadoch, M. Cheriet, and J. Cai, "Outage probability of cdf-based scheduling for uplink noma with practical sic considerations," in

2020 International Wireless Communications and Mobile Computing (IWCMC), IEEE, 2020, pp. 1031–1036.

[29] A. El Shafie, K. Tourki, and N. Al-Dhahir, "An artificial-noise-aided hybrid ts/ps scheme for ofdm-based swipt systems," *IEEE Communications Letters*, vol. 21, no. 3, pp. 632–635, 2016.

[30] T. M. Hoang, A. El Shafie, D. B. da Costa, T. Q. Duong, H. D. Tuan, and A. Marshall, "Security and energy harvesting for mimo-ofdm networks," *IEEE Transactions on Communications*, vol. 68, no. 4, pp. 2593–2606, 2019.

[31] T. G. Nguyen, C. So-In, H. Tran, *et al.*, "Outage performance analysis of energy harvesting wireless sensor networks for noma transmissions," *Mobile Networks and Applications*, vol. 25, no. 1, pp. 23–41, 2020.

[32] A. Badawy and A. ElShafie, "Securing ofdm-based noma swipt systems," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2020.

[33] A. Salem, L. Musavian, E. Jorswieck, and S. Aissa, "Secrecy outage probability of energy-harvesting cooperative noma transmissions with relay selection," *IEEE Transactions on Green Communications and Networking*, 2020.

[34] N. Senadhira, S. Durrani, X. Zhou, N. Yang, and M. Ding, "Uplink noma for cellular-connected uav: Impact of uav trajectories and altitude," *IEEE Transactions on Communications*, 2020.

[35] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 938–953, 2016.

[36] J. Tang, J. Luo, M. Liu, D. K. C. So, E. Alsusa, G. Chen, K. Wong, and J. A. Chambers, "Energy efficiency optimization for noma with swipt," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 3, pp. 452–466, 2019.

[37] J. A. Ansere, J. H. Anajemba, S. H. Sackey, C. Iwendi, and M. Kamal, "Optimal power distribution algorithm for energy efficient iot-noma enabled networks," in *2019 15th International Conference on Emerging Technologies (ICET)*, IEEE, 2019, pp. 1–5.

[38] J. Azar, A. Makhoul, M. Barhamgi, and R. Couturier, "An energy efficient iot data compression approach for edge machine learning," *Future Generation Computer Systems*, vol. 96, pp. 168–175, 2019, ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2019.02.005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X18331716.

[39] A. Ukil, S. Bandyopadhyay, and A. Pal, "Iot data compression: Sensor-agnostic approach," in *2015 Data Compression Conference*, 2015, pp. 303–312.

[40] F. Wu, K. Yang, and Z. Yang, "Compressed acquisition and denoising recovery of emgdi signal in wsns and iot," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 5, pp. 2210–2219, 2018.

[41] G. Xu, J. Han, Y. Zou, and X. Zeng, "A 1.5-d multi-channel eeg compression algorithm based on nlspiht," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1118–1122, Aug. 2015, ISSN: 1070-9908. DOI: 10.1109/LSP.2015.2389856.

[42] A. Awad, A. Mohamed, A. A. El-Sherif, and O. A. Nasr, "Interference-aware energy-efficient cross-layer design for healthcare monitoring applications," *Computer Networks*, vol. 74, pp. 64–77, 2014, ISSN: 1389-1286. DOI: https://doi.

org/10.1016/j.comnet.2014.09.003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128614003119.

[43] M. Elsayed, M. Mahmuddin, A. Badawy, T. Elfouly, A. Mohamed, and K. Abualsaud, "Walsh transform with moving average filtering for data compression in wireless sensor networks," in *2017 IEEE 13th International Colloquium on Signal Processing its Applications (CSPA)*, 2017, pp. 270–274.

[44] H.-S. Lee, J.-Y. Kim, and J.-W. Lee, "Resource allocation in wireless networks with deep reinforcement learning: A circumstance-independent approach," *IEEE Systems Journal*, vol. 14, no. 2, pp. 2589–2592, 2019.

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

[46] F. Bu and X. Wang, "A smart agriculture iot system based on deep reinforcement learning," *Future Generation Computer Systems*, vol. 99, pp. 500–507, 2019.

[47] H. Sun, X. Ma, and R. Q. Hu, "Adaptive federated learning with gradient compression in uplink noma," *arXiv preprint arXiv:2003.01344*, 2020.

[48] X. Liu, X. Chen, Y. Chen, and Z. Li, "Deep learning based dynamic uplink power control for noma ultra-dense network system," in *International Conference on Blockchain and Trustworthy Systems*, Springer, 2019, pp. 774–786.

[49] W. Ahsan, W. Yi, Y. Liu, Z. Qin, and A. Nallanathan, "Reinforcement learning for user clustering in noma-enabled uplink iot," in *2020 IEEE International*

Conference on Communications Workshops (ICC Workshops), IEEE, 2020, pp. 1–6.

[50] Y. Zhang, X. Wang, and Y. Xu, "Energy-efficient resource allocation in uplink noma systems with deep reinforcement learning," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, IEEE, 2019, pp. 1–6.

[51] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of uplink grant-free noma system," *IEEE Internet of Things Journal*, 2020.

[52] A. Awad, R. Hussein, A. Mohamed, and A. A. El-Sherif, "Energy-aware cross-layer optimization for eeg-based wireless monitoring applications," in *38th Annual IEEE Conference on Local Computer Networks*, IEEE, 2013, pp. 356–363.

[53] F. W. Murti and S. Y. Shin, "User pairing schemes based on channel quality indicator for uplink non-orthogonal multiple access," in *2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, 2017, pp. 225–230.

[54] B. Kim, W. Chung, S. Lim, S. Suh, J. Kwun, S. Choi, and D. Hong, "Uplink noma with multi-antenna," in *2015 IEEE 81st vehicular technology conference (VTC Spring)*, IEEE, 2015, pp. 1–5.

[55] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," 2014.

[56] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[57] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[58] J. T. Olkkonen, *Discrete Wavelet Transform, Theory and Applications*. inTechOpen, 2011, ISBN: 978-953-307-185-5. DOI: 10.5772/649.

[59] F. Di Salvo, "The exact distribution of a weighted convolution of two gamma distributions," Italian, 2006, pp. 511–514.

[60] B. Murmann, "A/d converter trends: Power dissipation, scaling and digitally assisted architectures," in *2008 IEEE Custom Integrated Circuits Conference*, 2008, pp. 105–112.

[61] S. Zhang and R. S. Sutton, "A deeper look at experience replay," *arXiv preprint arXiv:1712.01275*, 2017.