Article

D2D Mobile Relaying Meets NOMA—Part II: A Reinforcement Learning Perspective

Safaa Driouech ^{1,2} , Essaid Sabir ^{1,3,*} , Mounir Ghogho ⁴ and El-Mehdi Amhoud ⁵ ¹ NEST Research Group, LRI Lab., ENSEM, Hassan II University of Casablanca, Casablanca 20000, Morocco; driouech.safaa@hotmail.fr² Laboratoire de Recherche en Informatique, Sorbonne Université, CNRS, LIP6, F-75005 Paris, France³ Department of Computer Science, University of Quebec at Montreal, Montreal, QC H2L 2C4, Canada⁴ TICLab, International University of Rabat, Rabat 11100, Morocco; mounir.ghogho@uir.ac.ma⁵ School of Computer Science, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco; elmehdi.amhoud@um6p.ma

* Correspondence: e.sabir@ensem.ac.ma

Abstract: Structureless communications such as Device-to-Device (D2D) relaying are undeniably of paramount importance to improving the performance of today's mobile networks. Such a communication paradigm requires a certain level of intelligence at the device level, thereby allowing it to interact with the environment and make proper decisions. However, decentralizing decision-making may induce paradoxical outcomes, resulting in a drop in performance, which sustains the design of self-organizing yet efficient systems. We propose that each device decides either to directly connect to the eNodeB or get access via another device through a D2D link. In the first part of this article, we describe a biform game framework to analyze the proposed self-organized system's performance, under pure and mixed strategies. We use two reinforcement learning (RL) algorithms, enabling devices to self-organize and learn their pure/mixed equilibrium strategies in a fully distributed fashion. Decentralized RL algorithms are shown to play an important role in allowing devices to be self-organized and reach satisfactory performance with incomplete information or even under uncertainties. We point out through a simulation the importance of D2D relaying and assess how our learning schemes perform under slow/fast channel fading.

Keywords: D2D relaying; 5G/B5G/6G; biform game; self-organized devices; Nash equilibrium; distributed reinforcement learning; NOMA/OMA

Citation: Driouech, S.; Sabir, E.; Ghogho, M.; Amhoud, E.-M. D2D Mobile Relaying Meets NOMA—Part II: A Reinforcement Learning Perspective. *Sensors* **2021**, *21*, 1755. <https://doi.org/10.3390/s21051755>

Academic Editors: Xianbin Wang and Peter Chong

Received: 13 January 2021

Accepted: 15 February 2021

Published: 4 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Motivations and New Trends

The fifth-generation (5G) of mobile networks is being massively deployed worldwide, promising a never-before-seen user experience. 5G is expected to provide the basis for intelligent networks with some artificial intelligence (AI) applications and capabilities. However, a zero-touch network with fully intelligent orchestration and management remains elusive and will only be achieved in future [1]. In order to overcome the limitations of 5G deployment and fulfill the stringent requirements of future applications and services, the research community has started to explore and shape the next generation of wireless communications under different labels, such as B5G, 5GC, and 6G. In this context, some articles discussing the vision and open challenges for 6G have recently appeared in the literature, for example, refs. [2–4]. Indeed, 6G is foreseen to support novel data-hungry applications, a plethora of autonomous services and new communication scenarios by around 2030. These technologies will encompass holographic videos, extended reality (XR), flying networks and vehicles, massive unmanned mobility in three dimensions, autonomous driving, telemedicine, haptics, tactile Internet, brain-computing, large intelligent surfaces, connected autonomous systems, high-definition video streaming and cell-free

communications, to name a few. Thus, the amount of wireless data traffic and the number of connected objects are expected to increase hundreds of fold per cubic-meter volume [5]. 6G will connect millions of users and billions of machines everywhere. Thus, many extreme requirements (low energy consumption, long battery life, high intelligence level, high speeds, far larger bandwidth than 5G, high reliability, low latency, high data rates, etc. [2,4,6,7]) need to be addressed by design.

It is a reality that the current centralized systems suffer from large overheads and heavy signaling that might greedily misuse the network's scarce resources. Thus, a distributed system is recommended for dense and ultra-dense environments, as it offloads traffic from the network and minimizes the dependency on its connectivity. Decentralized decision-making exhibits important scalability features and can efficiently handle server breakdowns due to denial of service. One of the most important requirements and trends of the next generation of mobile networks is the self-X paradigm. This promises future networks will be able to perform the required functionalities, such as self-learning, self-configuration, self-optimization, self-healing, self-organization, self-aggregation, and self-protection, with no human intervention [8]. Such a paradigm shift is likely to lead to a more flexible and robust network. In this regard, artificial intelligence (AI) and machine learning (ML) can play significant roles in ensuring a fully connected digital society, supporting 6G autonomy, capturing insights and gaining proper comprehension of the environment. Namely, AI/ML will bring human-like intelligence into every aspect of networking operations (e.g., resource management; network planning and optimization; and failure detection and analysis). AI can be deployed and trained at the device and various network levels (base stations, core network, cloud, etc.). Of course, reaching this goal relies on the availability of data timely streamed to/from devices, in order to enhance operational reliability, real-time predictions and security.

Artificial intelligence has taken us by storm, helping us to make decisions in almost every sector and area of life. On the one hand, the most recent mobile handsets are already offering many exciting AI-based services, such as AI-based cameras, AI-based earphones and intelligent assistants (Siri, Cortana, Google Now, Bixby, Alexa, etc.) [8,9]. On the other hand, current assessment approaches closely rely on mathematical models that capture the structures and dynamics of the communication systems. However, efficient mathematical frameworks often do not offer high accuracy or use heavy mathematical tools with unsuitable computational time/complexity. Additionally, such an analytic model does not exist for some of the building blocks of wireless networks. Therefore, AI/ML will play a crucial role in 6G wireless networks, as it is capable of handling systems that cannot be modeled by standard mathematical frameworks. Now, ML-empowered machines can learn, adapt and improve their operations by exploiting the operational knowledge and experience gained in the form of data or experience. Based on the nature of available data and the explicitness of the learning objectives, an ML scheme can use supervised, unsupervised or reinforcement learning (RL).

Today, we live in a hyper-connected world where huge amounts of data are generated and can be efficiently used to fine-tune every single device's decisions. Decentralized RL is required. By using that scheme, updating information on devices only requires local actions and observed payoffs. Such an algorithm is very important in stochastic/dynamic environments where many parameters are unavailable, unobservable or simply unknown. Centralized ML algorithms require a massive number of geographically distributed smart mobile devices to transmit their collected data to the cloud for training purposes. Due to privacy issues and communication evidence, it is impractical for all wireless mobile devices to transmit their local data to train the ML model. To cope with this and avoid costly training, it is important to introduce distributed, decentralized and edge-deployed learning ML algorithms, such as federated learning (FL) [10,11] and decentralized RL algorithms, that enable mobile devices to collaboratively learn a shared ML model without any explicit data exchange.

1.2. Our Contributions

In this article, we revisit the potential of game theory combined with machine learning to efficiently intelligentize network entities. More precisely, we devote the first part of this article [12] to predicting and analyzing the performance of a wireless network with self-organizing devices, using a biform game perspective. Of course, the system's designer needs to have some information about the network, such as the game's rules, number of users, available strategies, utility functions, game sequence, and external perturbing factors (random channel, mobility, etc.). In this paper, we implement two fully distributed reinforcement learning algorithms at the device level in order to reach stable equilibria.

We invite the reader to see part I [12] of this work for details on the game-theoretic aspects of our D2D relaying scheme. The main contributions of this work are fivefold:

Part I's contributions are related to performance analysis of a self-organizing D2D relaying scheme:

1. We consider a hybrid two-tier scheme where cellular links use Non-Orthogonal Multiple Access (NOMA), whilst D2D links use Orthogonal Multiple Access (OMA). This scheme is suitable for both inband and outband D2D schemes.
2. We fully characterize the Rayleigh channel model and derive closed forms for the outage probabilities of both OMA and NOMA links, and then compute the average throughput perceived by each device in the network.
3. To the best of our knowledge, this work is the first to propose a biform game to capture the devices' behaviors while deciding which radio access network (RAN) to connect (i.e., either cellular or D2D).

Part II's contributions are related to implementing self-organized mode selection using RL. Each device has to strategically learn when to select cellular mode and then act as a relay, and when it has to get access via D2D mode:

4. We propose to empower devices with a self-organization capability, allowing them to reach pure Nash equilibria (linear reward-inaction) and mixed Nash equilibria (Boltzmann–Gibbs dynamics), in a fully distributed manner;
5. We performed extensive simulations to analyze the effects of different parameters on the learning schemes. Insights on accuracy and convergence are also provided.

This article is organized as follows: A biform game and equilibrium analysis are provided in part I [12] of this research. Section 2 presents a comprehensive literature review. The problem is formulated in Section 3. The decentralized RL algorithms are presented in Section 4. Next, we discuss the scalability, convergence and insights of our algorithms in Section 5. Extensive learning simulations are provided in Section 6. Finally, we draw some concluding remarks and list future directions in Section 7.

2. Related Work

Reinforcement learning (RL) is built upon a feedback reward mechanism: At every time step of the learning process, the agent interacts with the environment and observes its state, takes action from the set of available actions that maximizes the long-term cumulative reward and eventually moves to the next state. Some of the well-known techniques developed for network management are the multi-arm bandit, temporal-difference learning, the Markov decision process, SARSA learning, Q-learning and Monte Carlo methods [13–15]. These techniques can be applied in the application layer for proactive caching, data offloading, error prediction, data rate allocation and security. In the network layer, RL can be used for multi-objective routing, packet scheduling, traffic prediction and classification, etc. In the physical layer, RL can be employed for link preservation,

channel tracking, radio access association, on-demand beamforming, energy harvesting, modulation selection, radio identification, etc. [8,13,16–18].

RL consists of a pre-programmed process run by an agent seeking to maximize its cumulative reward over time. It has received substantial attention and has been applied to wireless networks at numerous levels [13,19,20]. The authors of [18] came up with a distributed RL algorithm that dynamically changes the transmitted data and power control at each sensor node according to observed state information such that the data of all sensor nodes are received while minimizing the delay. The authors in [21] came up with a distributed RL-based resource allocation scheme to mitigate interference between D2D users and cellular users. A distributed spectrum allocation framework based on multi-agent deep reinforcement learning was also proposed in [22]. Additionally, the authors in [23] proposed learning schemes that enable cognitive users to jointly learn their optimal payoffs and strategies for both continuous and discrete actions. The authors in [24] proposed an actor–critic reinforcement learning scheme for downlink transmission based on radio resource scheduling policy for long term evolution—advanced (LTE-A), to accomplish resource scheduling efficiently by maintaining user fairness and high QoS capabilities. In [25], the authors proposed a reinforcement learning scheme to optimize routing strategy without human participation. In [26], an RL algorithm was used for distributed packet routing. The latter resulted in less congestion and shorter delivery times.

Most of the related literature on reinforcement learning algorithms focuses on optimizing network performance. We aimed to study and analyze the system to predict the network’s outcome from a biform game-theoretic perspective. Then, we used decentralized learning schemes to learn the options each device should pursue to earn the “best” long-term average profit. On the one hand, game theory allows capturing the behavior of devices and finding stable solutions assuming complete information. On the other hand, RL schemes are used to reach these solutions, only using local information and some observable, possibly perturbed/random metrics.

3. The System

Consider the uplink case of a single 4G/5G/6G cell, where a finite number of devices $\mathcal{N} = \{1, 2, \dots, n\}$ are randomly distributed around the serving Base Station (BS). The devices communicate using NOMA in cellular links combined with conventional OMA for D2D links, as shown in Figure 1. We use a separate band for D2D users (i.e., D2D overlay mode). We use OMA for D2D links to (1) study a hybrid access system, and (2) eliminate interference between cellular and D2D UE. Each device i transmits using power vector $\mathbf{P}_i = (P_i^c, P_i^d)$. In this article, we do not consider multi-homing. When the mobile i is connected to the BS, it transmits with power P_i^c , which implies $P_i^d = 0$. The mobile transmits with power P_i^d (i.e., $P_i^c = 0$) if it is connected through a D2D link. Of course, the transmit power using a D2D link is (much) lower than in cellular communication. Each device is d_i away from the BS and experiences a channel gain h_i . For better readability, the main notation and symbols used in this article are listed in Table 1.

For the sake of simplicity and without loss of generality, the device numbered 1 is assumed to be the closest device to the serving BS, with distance d_1 . It transmits with the lowest power P_1 and experiences the strongest channel gain h_1 . It follows that device n is the farthest with distance d_n from the BS; it uses the highest transmit power P_n and uses the weakest channel h_n . Namely, we have $|h_1|^2 \geq |h_2|^2 \geq \dots \geq |h_{n-1}|^2 \geq |h_n|^2$. Let $w(t)$ be the received noise at the BS and assume each device i transmits its individual signal $s_i(t)$. Then, the aggregate received signal at the BS is:

$$S(t) = \sum_{i=1}^n \sqrt{P_i^c} h_i s_i(t) + w(t), \quad (1)$$

The BS decodes the signals by applying the successive interference cancellation (SIC) technique. The received signal power corresponding to the user of the strongest channel

is likely the strongest at the BS, so the signal is the first to be decoded at the BS and experiences interference from all the remaining weaker channels in the cluster [27,28]. Thus, the transmission of device 1 experiences interference from users with weaker channels in the cluster, whereas the transmission of device n experiences zero interference. In contrast to downlink NOMA, each user in uplink NOMA can independently utilize its battery power to the maximum since the channel gains of all the users are sufficiently distinct [29].

Figure 1. Cellular offloading using device-to-device (D2D) cooperative relaying.

Table 1. Main symbols and their meanings.

Symbol	Meaning
n	Number of active devices in the cell
$\mathbf{P}_i$	Transmission power vector of device i
P_i^c	Transmission power of device i while transmitting to the BS
P_i^d	Transmission power of device i while transmitting over the D2D link
d_i	Distance between device i and the BS
h_i	Channel gain of device i
γ_i	SINR of device i
$\gamma_{i,th}$	SINR-threshold
$P_i^{out}(\gamma_i)$	Outage probability of device i
$\frac{1}{\lambda}$	Mean of the channel gain
R	Transmission rate
$f_{j,k}$	Orthogonality factor between the carriers allocated to active devices j and k
α_c, α_d	Path-loss exponent in cellular and D2D, respectively
x_i	Fraction of throughput relay i gives to D2D devices
$U_i^t(a_i)$	Utility of device i at time t when choosing the action a_i

3.1. Channel Model

In this article, the radio signal experiences attenuation due to the path-loss with exponent α and Rayleigh fading. We denote by γ_i the instantaneous signal to interference and noise ratio (SINR) of device i , and it is given by:

$$\gamma_i = \frac{P_i^c |h_i|^2 d_i^{-\alpha_c}}{\sum_{j=i+1}^n P_j^c |h_j|^2 d_j^{-\alpha_c} + \sigma_N^2}, \quad (2)$$

It is worth noting that the SINR of the weakest device n experiences no interference according to NOMA operation; i.e., $\gamma_n = \frac{P_n^c |h_n|^2 d_n^{-\alpha_c}}{\sigma_N^2}$. σ_N^2 denotes the variance of the thermal

additive white Gaussian noise. When devices communicate through a D2D link, the SINR of a device j relaying on device i is given by:

$$\gamma_j = \frac{P_j^d |h_j^d|^2 d_{j,i}^{-\alpha_d}}{\sum_{k \in \mathcal{N} \setminus \{i,j\}} f_{j,k} P_k^d |h_k^d|^2 d_{k,i}^{-\alpha_d} + \sigma_N^2}, \quad (3)$$

where $f_{j,k}$ is the orthogonality factor between the carriers allocated to active devices j and k . k refers to the other devices communicating through the D2D link and relaying on device i . Each device aims at guaranteeing an instantaneous SINR above a certain threshold $\gamma_{i,th}$ to have successful communication. The outage probability denotes the probability that the SINR is less or equal than a given SINR threshold ($\gamma_{i,th}$). The outage probability $P_i^{out}(\gamma_i)$ of device i is given by [12]:

$$P_i^{out}(\gamma_i) = Pr(\gamma_i \leq \gamma_{i,th}) = 1 - \frac{\prod_{j=i+1}^n \lambda_j \cdot e^{-\frac{\gamma_{i,th}^2 \lambda_i}{P_i^c d_i^{-\alpha_c}}}}{\prod_{j=i+1}^n \left(\lambda_j + \frac{\gamma_{i,th} P_j^c d_j^{-\alpha_c}}{P_i^c d_i^{-\alpha_c}} \lambda_i \right)}, \quad (4)$$

where $\frac{1}{\lambda} \geq 0$ is the mean and scale parameter of the distribution, often taken as equal to 1. The calculus of the outage probability in cellular and D2D modes was detailed in the first part [12] of this work.

3.2. Average Throughput

In general, device i transmits data with a rate R_i with every use of the channel (i.e., in every packet or frame transmission), with the condition that R_i must not exceed its channel capacity; i.e., $R_i \leq \log(1 + \gamma_i)$.

We define the throughput of the transmission as the rate of successful data bits that are transmitted to the destination over a communication channel, expressed as follows:

$$\Theta_i(\gamma_i) = \frac{M}{L} R_i (1 - P_i^{out}(\gamma_i)) = \rho_i (1 - P_i^{out}(\gamma_i)), \quad (5)$$

where $\rho_i = \frac{M}{L} R_i$. M is the data length (in bits). L denotes the total number of bits in a frame.

4. Decentralized Learning Algorithms

In this section, we propose an algorithmic implementation of the biform game framework we developed in [12], by means of decentralized learning algorithms. Each device that is a player in the game operates in a self-organized manner using a decentralized RL algorithm to learn its locally optimal (at Nash equilibrium) strategy over time. Based on that, it converges to the action (communication mode) that improves its throughput. The optimal action is defined as the one with the highest expected environmental reaction value. Decentralized RL learning algorithms are designed to naturally converge to Nash equilibria in games with incomplete information.

A learning scheme is a process within which an agent interacts with the environment and updates its strategy toward a Nash equilibrium, based on its experience built from past rounds. Each device can learn the optimal action from a finite set of actions by iterating its update pattern, even under bounded uncertainty; see Figure 2. Such a decentralized scheme only requires local information and some metrics observed while interacting with the environment [15].

Figure 2. A decentralized reinforcement learning process.

Let us consider n agents representing the n active devices in the network, where $A^t = \{a_1^t, a_2^t, \dots, a_n^t\}$ is the set of all joint agent actions at time t . At the beginning of every time slot/step, each device i , $i \in \mathcal{N}$, chooses either to be a relay ($a_i^t = 0$) or to communicate through D2D ($a_i^t = 1$) while expecting other devices will serve as relays. U_i^t is the average throughput observed/measured by device i at the end of time slot t ; it represents its payoff (reward) at time t . Let $p_i(t)$ denote the probability distribution of the i -th agent over its pure strategies at instant t ; $p_i(0)$ is the initial mixed strategy of player i . Each round of the game consists of each of the players choosing an action randomly and independently based on their respective current strategies and their perceived rewards at the previous step.

The update process of the distributed learning algorithm is depicted in Figure 2. Each device/player constructs and updates its strategy based on its set of own-actions, received reward (i.e., observed/measured throughput) and potentially some belief or estimate of future reward. There is no explicit communication or information exchange between active devices. Here, all the players update their strategies within the same time window in a dynamic interactive fashion. The wireless environment is perturbed by the channel state, devices' mobility patterns, devices' locations, interference, devices' actions, etc.

4.1. Learning Pure Nash Equilibrium (PNE): The Linear Reward-Inaction Algorithm (LRI)

In order to reach a pure equilibrium in a distributed manner under unknown and slow fading channel assumptions, with $|h_1|^2 \geq |h_2|^2 \geq \dots \geq |h_{n-1}|^2 \geq |h_n|^2$, we propose to use the linear reward-inaction (LRI) algorithm for the devices to learn their respective best response strategies. Recall that LRI is a completely decentralized learning algorithm, since each learner ignores information from other agents and only considers its own actions and payoffs from the environment. Every device, independently of its competitors, updates a probability distribution over time to gradually converge to the best strategy. Initially, the player chooses randomly a strategy according to the initial probability distribution; after each time step, the algorithm increases the probability of picking strategies with great utility and decreases (reinforcement) the probability of choosing the other strategies. Let $\mathbf{p}_t = (p_t, 1 - p_t)$ be the probability distribution at instant t , where p_t denotes the probability of choosing the action "relay" at instant t ; thus, $1 - p_t$ is the probability of choosing the action "D2D."

After each transmission, the device receives an acknowledgment (ACK) if the BS successfully receives the message sent in that slot. Otherwise, the receiver feeds back an explicit/implicit NACK. Thus, the main question is: what is the action that can allow a device to receive the maximum possible number of ACKs with every use of channel?

To receive an ACK from the BS, the SINR $\gamma_i(t)$ experienced at time t of device i needs to be higher than the SINR threshold γ_{th} . To model this, we define the reward of agent i for a given action profile $(a_i(t), a_{-i}(t))$ at time slot t as:

$$U_i^t(a_i(t)) = Thp_i(t) * \mathbb{1}_{\{\gamma_i(t) \geq \gamma_{th}\}} = \begin{cases} Thp_i(t) & \text{if } \gamma_i(t) \geq \gamma_{th} \quad (\text{ACK}) \\ 0 & \text{else} \quad (\text{NACK}) \end{cases} \quad (6)$$

where $Thp_i(t)$ is the instantaneous throughput perceived by device i at time instance t , and $\mathbb{1}$ is the indicator function. Of course, the instantaneous throughput depends on the device's strategy, other devices' strategy vectors, environment perturbations, etc. $\mathcal{M}$ denotes the set of devices directly connected to the base station at instant t , with cardinality $m(t)$. Thus, the remaining $n - m(t)$ devices are getting access via D2D.

$$Thp_i(t) = \begin{cases} \rho_i, & \text{if all the devices choose cellular,} \\ x_i \rho_i, & \text{if device } i \text{ is a relay,} \\ \sum_{j \in \mathcal{M}} \frac{1-x_j}{n-m(t)} \rho_j, & \text{if device } i \text{ get access through D2D links,} \\ 0, & \text{No device is connected to the base station.} \end{cases} \quad (7)$$

In case there are many devices communicating through D2D mode, the fraction of throughput given from the relays will be equally divided between the D2D-transmitters. If there is at least one device in the D2D group, then the relay device allocates a fraction of its throughput $1 - x_i$ to that group. x_i allows one to define the mode selection and the cooperation level of device i . For instance, $x_i = 1$ means device i fully opts for cellular mode. Meanwhile, $x_i = 0$ means device i chooses to communicate through D2D link. When $x_i \in]0, 1[$ the device i plays the role of a relay. If x_i is high then device i uses a high fraction for itself and does not relay much; otherwise, it offers more throughput to D2D devices.

According to LRI pattern, the probability of device i serving as a relay at instant $t + 1$ is:

$$p_i(t+1) = p_i(t) + \epsilon_t \frac{U_i^t(a_i(t))}{R} \left(\mathbb{1}_{\{i \text{ is a relay at instant } t\}} - p_i(t) \right), i \in \mathcal{N}. \quad (8)$$

- $p_i(t)$ is the probability of relaying for device i at time t .
- ϵ_t denotes the learning rate parameter at instant t (i.e., the step size).
- $\frac{U_i^t}{R}$ is the environmental response, which is represented here as the fraction between the device i 's instantaneous throughput and the transmission rate.
- $\mathbb{1}_t$ denotes the indicator function that equals 1 when the action chosen by a device at time t is $a_i = 0$; otherwise it equals 0.

The learning algorithm described in Algorithm 1 used by each of the players is explained below:

1. At every instant, each player chooses an action based on its action probability vector. Thus, player i chooses action a_i at instant t , according to the probability distribution $\mathbf{p}_i(t)$;
2. Each player obtains a reward based on its action and the set of all other devices' actions. The reward of player i is $U_i^t(a_i(t))$;
3. Each player updates his action probability based on the rule in Equation (8).

Algorithm 1: Linear reward-inaction algorithm (LRI).**Parameter Initialization**

Each device peaks randomly an initial probability of relaying p_i^0 .

In parallel: at each device i **do**

Repeat

if $a_i(t) = 0$ **then**

$$p_i^{t+1} \leftarrow p_i^t + \epsilon \frac{U_i^t}{R} (1 - p_i^t)$$

else

$$p_i^{t+1} \leftarrow p_i^t - \epsilon \frac{U_i^t}{R} p_i^t$$

Until The stopping criterion

4.2. Learning Mixed Nash Equilibrium (MNE): Boltzmann–Gibbs-Based Payoff-RL

Under the channel randomness assumption (i.e., fast fading channel), each device decides to be indifferent between the actions chosen, and look for the probability of relaying that guarantees the same payoff of either being connected to cellular or D2D. After that every device has chosen an action according to its probabilities, it receives the corresponding payoff and then updates its strategy. The update of strategies occurs following Boltzmann–Gibbs-based Payoff-RL [13]. The players use only information regarding their own choices and perceived payoffs, and do not explicitly require external information concerning the channel gains, the transmit powers, the payoffs, and choices of the other players, i.e., a fully distributed learning algorithm.

The Boltzmann–Gibbs-based Payoff-RL algorithm consists of updating the probability distribution over the possible actions as follows:

$$\begin{cases} p_i(t+1) = \frac{e^{\frac{\hat{U}_i^t(a_i)}{r_i}}}{\sum_{a_i' \in A_i} e^{\frac{\hat{U}_i^t(a_i')}{r_i}}}, a_i \in A_i, i \in \mathcal{N} \\ \hat{U}_i^{t+1}(a_i) = \hat{U}_i^t(a_i) + v_{i,t} \mathbb{1}_{\{a_{i,t}=a_i\}} (U_i^t(a_i) - \hat{U}_i^t(a_i)) \end{cases} \quad (9)$$

where r_i is player i 's temperature. This quantity measures the rationality level of i . Under small values of r_i , the player is strongly rational and the algorithm converges to the Nash equilibrium. Whereas, the player becomes irrational under large values of r_i , which results in selecting available strategies following a uniform distribution. Thus, the algorithm fails to explore the equilibrium strategies. $v_{i,t}$ denotes the learning rate: it counts how many times the actions a_i have been chosen.

The Boltzmann–Gibbs dynamics are summarized in Algorithm 2. In addition to the observed real reward, this learning scheme allows the players to learn mixed strategies by estimating/predicting the payoffs at next time steps. The reward predictor is adjusted using actual real reward, which improves the estimation accuracy.

Algorithm 2: Boltzmann–Gibbs-based payoff-RL.**Parameter Initialization**

Each device peaks randomly an initial probability of relaying p_i^0 .

In parallel: at each device i **do**

Repeat

$$p_i^{t+1} \leftarrow \frac{e^{\frac{\hat{U}_i^t(a_i)}{r_i}}}{e^{\frac{\hat{U}_i^t(a_i)}{r_i}} + e^{\frac{\hat{U}_i^t(a_i)}{r_i}}}$$

$$\hat{U}_i^{t+1}(a_i) \leftarrow \hat{U}_i^t(a_i) + v_{i,t} \mathbb{1}_{\{a_{i,t}=a_i\}} (U_i^t(a_i) - \hat{U}_i^t(a_i))$$

Until The stopping criterion

5. Some Insights on RL Convergence, Fairness and Scalability

While building an iterative scheme, the system designer must deal with many important issues in order to yield an efficient learning algorithm. The efficiency of such a scheme is measured through its accuracy, speed, fairness and scalability features.

5.1. Convergence

The learning algorithm converges when as the iterations proceed, the output gets closer to some specific stationary value. After that, the algorithm settles down on this value and does not improve anymore even when additional training is applied. From a theoretic perspective, most learning algorithms found in the literature fail to converge to the value computed analytically. Indeed, these iterative schemes provide an output that always deviates from theoretic values by some amount. However, in numerous cases a near-optimal solution is acceptable. Here, relative error can be a suitable stopping criterion, i.e., when $\left| \frac{p_i^{t+1} - p_i^t}{p_i^t} \right| \leq \epsilon$, where ϵ is the convergence tolerance that sets the desired accuracy level. It is understood that high accuracy might result in a long convergence time. In some specific cases, the RL algorithm might not converge or simply can diverge: (1) when the speed of convergence is arbitrary slow; or (2) when no attractor point (e.g., Nash equilibrium) exists. In [13], the convergence issue has been extensively discussed. In general, many parameters can affect the speed and accuracy of convergence of the learning algorithms, namely, the number of players, the mobility, the randomness degree of the environment, etc. In this work, we present some parameters that impact the time of convergence of the proposed algorithms, namely, fast/slow fading, power allocation, SINR threshold and the learning rate parameter.

5.2. Fairness

In this article, we fully characterize the pure and mixed Nash equilibria. Nash Equilibrium is defined as a solution where no player can improve its reward by deviating unilaterally, given the actions of the other players. At the PNE, each device must select to exclusively connect to the serving eNodeB or to exclusively connect via D2D. Such a solution presents evident fairness issues. In fact, devices communicating through D2D may transmit at low power due to short links. Meanwhile, the relaying devices consume much higher power, as they have to reach the eNodeB from longer distances and are expected to share, via D2D links, their throughput with unconnected devices. Of course, lack of fairness may induce harmful stability issues and then may result in a breakdown of the whole system (e.g., no device accepts to serve as a relay). Fortunately, mixed Nash equilibrium exhibits better fairness properties, as it allows each device to freely use its available strategies with different probabilities to possibly ensure a stable situation and satisfactory performance.

5.3. Scalability

Scalable algorithms are designed to handle huge amounts of data and perform heavy computations in a cost-effective and fast manner. The tremendous quantity of data that is being generated by IoT, 5G and the future 6G, is openly challenging vendors, operators and wireless research community. Hence, scalability is becoming a prominent topic of focus to conduct exciting researches. Efficient proposals able to deal with massive number of devices and huge data within a reasonable time window are of paramount importance. It is clear that enabling decision-making at wireless devices is beneficial in terms of spectrum utilization, reduced signaling and computational/energy costs [13]. Our schemes are designed to be fully decentralized, so independently of the number of devices. Since the device only uses local information and some observed reward/state, then our RL schemes exhibit nice elasticity property. Our RL-based algorithms are expected to continue performing well at the cost of longer convergence time as the population size increases. Still, thorough scalability analysis has to be conducted to analyze how does our schemes scale up in ultra-dense environments. Prior to designing an RL scheme, one needs to first predict and analyze the agents behavior [12].

6. Performance Analysis

In this section, we evaluate the performances of the RL algorithms using Mathworks Matlab R2020a. For illustrative purposes, we performed simulations for the three-device case. Unless counter-indication, figures were produced using the following setup: $P_1^c = 10$ mW, $P_2^c = 30$ mW, $P_3^c = 50$ mW, $P_1^d = P_2^d = P_3^d = 5$ mW, $R_1 = R_2 = R_3 = 1$ Mbit/s, $L = M = 1024$ bits, $\gamma_{th} = 40$ dB, $\alpha_c = \alpha_d = 3$ and $\sigma_N^2 = -116$ dBm, $d_1 = 100$ m, $d_2 = 300$ m, $d_3 = 500$ m, $f = 10^{-5}$, $x_1 = x_2 = x_3 = 0.5$. As for the channel gains, we consider Rayleigh channels such that $E[h_1] \geq E[h_2] \geq E[h_3]$.

6.1. Reaching a Pure Nash Equilibrium Using the Linear Reward-Inaction Algorithm

Figure 3 depicts the use of LRI learning algorithm to reach the PNE. It is clearly seen that device 1 converges to $a_1^* = 1$ while device 2 and 3 converge to $a_2^* = a_3^* = 0$. So the action that verifies a PNE is (0,1,1), which means that device 1 serves as a relay while device 2 and device 3 connect to the D2D link. It also converges, sometimes, to the other pure equilibria where device 2 communicates through cellular and serves as relay to the other devices. We shall mention that our LRI-based scheme converges to the same pure equilibria found using the framework developed in part I [12]. It converges more frequently to the dominant strategy when exists, independently of the initial probabilities of learning.

Figure 3. Convergence to pure Nash equilibrium using the linear reward-inaction algorithm.

Figure 4 reports the change in the devices' strategies with respect to the SINR threshold. To attain the equilibrium, device 2 decides to serve as a relay in the first figure when γ_{th} decreases, while in the second figure device 1 and 3 serve both as relays and device 2 remains in D2D. Decreasing the SINR threshold increases the probability of the devices to have successful communication and to receive an ACK. For $\gamma_{th} = 40$ dB, the equilibrium is reached when device 1 is the relay, because, as the strongest device, it can attain such threshold and can be a relay to device 2 and device 3. For $\gamma_{th} = 30$ dB, device 2 can also

have a successful communication if device 3 is in D2D, in this way device 1 gets affected by device 2's interference and decides to switch to D2D for better throughput. For $\gamma_{th} = 20$ dB, all the devices can have successful transmissions, but to reach the equilibrium, device 1 and device 3 serve as relays while device 2 is in D2D mode. In this way, device 2 gets rid of all interference and does not interfere with device 1 signal. Additionally, device 3 transmits with no interference.

Figure 4. Convergence to pure Nash equilibrium using LRI, for different SINR threshold values (γ_{th}).

Figure 5 shows the effect of the devices' distances and the transmit power to the equilibrium strategies. In this case, device 2 becomes closer to the BS, so interference applied from device 2 to device 1 increases. This explains device 2 taking the relay role as it is the second strongest device. Additionally, the transmit power plays an important role in defining the equilibrium strategies. However, when P_2 decreases, interference effect decreases and allows device 1 to take over again the relay role.

Figure 5. Convergence to pure Nash equilibrium using LRI, for different transmit powers of device 2.

In Figure 6, we see that device 3 decides to serve as a relay when its distance from the BS decreases. By doing so, interference applied from device 3 to device 2 increases, which decreases γ_2 that is why device 3 serves as a relay. Note that device 3 is the weakest but at the same time it is the only device that doesn't interfere with the other devices' signals, once on the cellular side. It prefers, then, to serve as a relay than to be in D2D and get affected by other devices' interference. However, minimizing the fraction of throughput device 3 gets for itself, if it serves as a relay, minimizes its throughput and ruins the equilibrium. That is why device 2 takes over the relaying mission in the second figure when x_3 decreases. For the same reason, device 1 serves as a relay when x_2 and x_3 decreases. This observation shows the importance and effectiveness of the decentralized RL algorithm to allow the devices to attain the equilibrium by converging to the pure action strategy.

Figure 6. Convergence to pure Nash equilibrium using LRI, for different cooperation levels (x_i).

Figure 7 depicts the effect of the channel variation on the time of convergence of the learning algorithm. The higher the speed of variation of the channel, the longer the algorithm takes to converge. Additionally, due to the random aspect of the channel, the algorithm might not converge to the same equilibrium. In addition, for the fast-fading case, the algorithm does not converge at all; it continues to seek pure equilibrium but cannot reach it due to the rapid variation of the channel. Henceforth, we need to look at the mixed Nash equilibrium to suggest an alternative solution when a PNE may not exist, or the channel is rapidly varying, slowing-down or preventing convergence.

6.2. Reaching Mixed Equilibrium Using Boltzmann–Gibbs (BG) Algorithm

Different from a pure strategy that denotes the action a device takes with probability 1, a mixed strategy is an assignment of a probability to each pure strategy. In the pure analysis, the equilibrium is reached such that no player would benefit by changing his strategy, while in the mixed analysis each device is indifferent between the actions taken. In the sequel, we investigate the mixed strategy using BG-RL algorithm and the long-term behavior of the devices.

In part I [12] of this work, we rely on the expected throughput value (through outage probability) and the average channel gain. Now, we study the real case by generating random Rayleigh channel gains. We respect the condition of $|h_1|^2 \geq |h_2|^2 \geq |h_3|^2$ and allow random fast fading. To reach the mixed equilibrium, we use the decentralized Boltzmann–Gibbs dynamics above-mentioned in Algorithm 2.

Figure 7. The effect of channel fading on the convergence toward a pure Nash equilibrium using LRI.

Figure 8 provides an illustrative example of p_i^* convergence, under Gibbs-Boltzmann scheme. It discovers the mixed equilibrium strategies in the case of a fast-fading channel. The proposed scheme, contrary to the biiform game, considers random channels and estimated utilities. In the figure, device 2 serves as a relay for most of the time (65% of the time), then device 1 with 45% and finally device 3 with 5%. Device 1 is the strongest but it is the most affected by both devices' interference. By adopting the indifferent mode, it chooses to communicate through cellular for half of the time and D2D for the other half. While device 2, the second strongest device, chooses to communicate through cellular and serve as a relay for most of the time. Device 2 is affected, once in cellular, by device 3 interference only. Thus, to reach high throughput, device 3 needs to stay in D2D mode, this is why device 3 communicates through D2D for more than 90% of the time.

Figure 8. The effect of channel fading on the convergence to a mixed NE using BG dynamics.

Figures 9 and 10 depict the behavior of the devices using the BG algorithm function of γ_{th} and x_i . Differently from Figure 8 where $\gamma_{th} = 40$ dB and $x_i = 0.5, i \in \{1, 2, 3\}$, the algorithm converges to partially mixed equilibria. For $\gamma_{th} = 30$ dB, device 1 decides to

always communicate through cellular and serve as relay to device 3 that decides to always use D2D communication, while device 2 communicates 70% of the time via cellular mode and 30% via D2D. For $\gamma_{th} = 20$ dB, device 1 behaves as before and serves as relay to device 2 that decides to always use D2D communication, while device 3 communicates through both modes. The reason behind this behavior is that device 1, as the strongest device, could attain higher throughput with low γ_{th} even with the presence of other devices' interference and could then serve as a relay. If a device switches to D2D mode, the throughput of device 1 increases and so does the portion of throughput it gives to D2D devices. For the case where device 2 decides, if it serves as a relay, to use 90% of the throughput for itself and give just 10% of it to D2D devices, the mixed equilibrium is reached if device 2 communicates through D2D while device 1 serves as relay and device 3 communicates through both modes. This is because allowing a small fraction of throughput to D2D devices leads to a decrease in their throughput values.

Figure 9. The effect of the SINR threshold (γ_{th}) on the convergence to a mixed NE using BG dynamics.

Figure 10. The effect of cooperation level (x_i) on the convergence to a mixed NE using BG dynamics.

However, choosing a mixed strategy by being indifferent doesn't always guarantee high throughput. For example, in our case, the throughput of the devices while being indifferent is lower than the one while using pure strategies. Thus, we decide to consider the long-term behavior of the devices, by allowing them to seek the pure strategy at each part of time where the channel remains constant before it changes randomly.

6.3. Long-Term Behavior

We turn now to investigate the long-term behavior of the devices equilibria. The long-term behavior in our case refers to what happens to the relaying probability as the number of iterations gets big and as long as the channel is randomly changing. If the channel is experiencing a fast fading, it could not converge to a pure strategy, but it could, at each part of the time where the channel is steady, converge to the pure equilibrium that changes when the channel gain changes too. The long-term behavior is then the average relaying probability of the devices as long as the channel changes.

Figure 11 depicts the long-term behavior of the devices with respect to the transmit power. In the first figure, device 1 uses a higher transmit power, so it could benefit from higher throughput. This allows device 1 to use cellular mode and act as a relay for approximately all the time, while device 2 and device 3 use D2D mode for more than 80%

of the time. In the second figure, device 3 uses high transmit power, so it chooses to serve as a relay for 80% of the transmission period while device 1 transmits through D2D mode for 70% of the time. This is because, when device 3 transmits with high power, it can attain high throughput values but causes high interference effect to device 1 and device 2. That is why device 1 switches to D2D for most of the time to avoid other devices' interference and device 3 stays at the cellular side as the device offering the highest throughput.

Figure 11. Long-term relaying probability under different transmit power schemes.

Figure 12 represents the long-term behavior of the devices with respect to γ_{th} . Having a low SINR-threshold value allows device 1 to reach high throughput even in the presence of interference so, as the strongest device, it serves as a relay most of the time even if the channel quality changes. The same thing happens to device 2 and 3: having a low γ_{th} allows device 3 to serve as relay more than half of the time because it could reach high throughput. Additionally, as it is the weakest device, it is not affected by other devices' interference once on the cellular side.

Figure 12. Long-term relaying probability under fast fading and different SINR threshold values.

Figure 13 shows the long-term behavior of the devices with respect to x_i . If device 1 decides, as a relay, to use 90% of the throughput for itself, device 3 decides then to serve as relay instead of device 1 to ensure high throughput for D2D devices. While if device 3 uses 10% of throughput for itself, its probability of relaying diminishes because its throughput decreases. Depending on the devices' parameters, the learning algorithms allow them to converge to pure and mixed strategies to attain equilibrium. These algorithms are totally decentralized. Each device has no information about the number and the parameters of other devices. The only thing it needs is its own throughput value at each transmission.

Figure 13. Long-term relaying probability under fast fading and different cooperation levels x_i .

7. Conclusion and Perspectives

In this article, we considered the uplink case wherein n devices choose whether to communicate through cellular (e.g., 5G/6G) or via D2D links to maximize their throughput. Cellular devices use NOMA, whilst they may serve neighboring devices using an orthogonal multiple access method (e.g., OFDMA/SC-FDMA). In the first part [12] of this article, we exhibited a biform game framework to predict and assess the devices behavior. In this second part, we proposed two distributed reinforcement learning algorithms to enable devices to self-explore their optimal strategies in a fully distributed manner. The simulation results showed that D2D relaying improves the overall throughput. Decentralized RL was shown to be efficient and rapid when implementing self-configuring features. However, reaching an equilibrium point does not always guarantee a satisfactory quality of service level for the devices. In our future work, we aim to deal with the case of QoS-sensitive (AR/VR/MR/XR, URLLC, etc.) applications, using federated learning.

Author Contributions: Conceptualization, S.D., E.S. and M.G.; formal analysis, S.D., E.S. and M.G.; funding acquisition, M.G. and E.-M.A.; investigation, S.D.; methodology, S.D. and E.S.; project administration, E.S. and M.G.; software, S.D. and E.-M.A.; supervision, E.S. and M.G.; validation, E.S., M.G. and E.-M.A.; writing—original draft, S.D.; writing—review and editing, E.S. and E.-M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by NEST Research Group (ENSEM), Mohammed VI Polytechnic university, and International University of Rabat.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable and insightful comments. Part of this work has been conducted during the internship of Safaa Driouech at TICLab at the International University of Rabat (UIR) under the supervision of Mounir Ghogho and Essaid Sabir.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nawaz, S.J.; Sharma, S.K.; Wyne, S.; Patwary, M.; Asaduzzaman, M. Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future. *IEEE Access* **2019**, *7*, 46317–46350. [\[CrossRef\]](#)
2. Alsharif, M.H.; Kelechi, A.H.; Albreem, M.A.; Chaudhry, S.A.; Zia, M.S.; Kim, S. Sixth Generation (6G) Wireless Networks: Vision, Research Activities, Challenges and Potential Solutions. *Symmetry* **2020**, *12*, 676. [\[CrossRef\]](#)
3. Ali, S.; Saad, W.; Rajatheva, N.; Chang, K.; Steinbach, D.; Sliwa, B.; Wietfeld, C.; Mei, K.; Shiri, H.; Zepernick, H.J.; et al. 6G White Paper on Machine Learning in Wireless Communication Networks. *arXiv* **2020**, arXiv:2004.13875.
4. Dang, S.; Amin, O.; Shihada, B.; Alouini, M.S. What should 6G be? *arXiv* **2019**, arXiv:1906.00741.
5. Aazhang, B.; Ahokangas, P.; Alves, H.; Alouini, M.S.; Beek, J.; Benn, H.; Bennis, M.; Belfiore, J.; Strinati, E.; Chen, F.; et al. *Key Drivers and Research Challenges for 6G Ubiquitous Wireless Intelligence (White Paper)*; University of Oulu: Oulu, Finland, 2019.
6. Saad, W.; Bennis, M.; Chen, M. A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems. *IEEE Netw.* **2020**, *34*, 134–142. [\[CrossRef\]](#)
7. Tariq, F.; Khandaker, M.R.A.; Wong, K.K.; Imran, M.A.; Bennis, M.; Debbah, M. A Speculative Study on 6G. *IEEE Wirel. Commun.* **2020**, *27*, 118–125. [\[CrossRef\]](#)
8. Piran, M.; Suh, D. Learning-Driven Wireless Communications, towards 6G. *arXiv* **2019**, arXiv:1908.07335.
9. Janbi, N.; Katib, I.; Albeshri, A.; Mehmood, R. Distributed Artificial Intelligence-as-a-Service (DAIaaS) for Smarter IoE and 6G Environments. *Sensors* **2020**, *20*, 5796. [\[CrossRef\]](#)
10. Xiao, Y.; Shi, G.; Krunz, M. Towards Ubiquitous AI in 6G with Federated Learning. *arXiv* **2020**, arXiv:2004.13563.
11. Khan, L.U.; Pandey, S.R.; Tran, N.H.; Saad, W.; Han, Z.; Nguyen, M.N.H.; Hong, C.S. Federated Learning for Edge Networks: Resource Optimization and Incentive Mechanism. *IEEE Commun. Mag.* **2020**, *58*, 88–93. [\[CrossRef\]](#)
12. Driouech, S.; Sabir, E.; Ghogho, M.; Amhoud, E.M. D2D Mobile Relaying Meets NOMA—Part I: A Biform Game Analysis. *Sensors* **2021**, *21*, 702. [\[CrossRef\]](#)
13. Tembine, H. *Distributed Strategic Learning for Wireless Engineers*; CRC Press: Boca Raton, FL, USA, 2012. [\[CrossRef\]](#)
14. Shahid, A.; Aslam, S.; Lee, K.G. A decentralized heuristic approach towards resource allocation in femtocell networks. *Entropy* **2013**, *15*, 2524–2547. [\[CrossRef\]](#)

15. Lu, X.; Schwartz, H.M. Decentralized learning in two-player zero-sum games: A L-RI lagging anchor algorithm. In Proceedings of the 2011 American Control Conference, San Francisco, CA, USA, 29 June–1 July 2011; pp. 107–112.
16. Elhammouti, H.; Sabir, E.; Benjillali, M.; Echabbi, L.; Tembine, H. Self-Organized Connected Objects: Rethinking QoS Provisioning for IoT Services. *IEEE Commun. Mag.* **2017**, *55*, 41–47. [[CrossRef](#)]
17. Attaoui, W.; Sabir, E. *Combined Beam Alignment and Power Allocation for NOMA-Empowered mmWave Communications*; Ubiquitous Networking; Habachi, O., Meghdadi, V., Sabir, E., Cancès, J.P., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 82–95.
18. Al-Tous, H.; Barhumi, I. Distributed reinforcement learning algorithm for energy harvesting sensor networks. In Proceedings of the 2019 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Sochi, Russia, 3–6 June 2019; pp. 1–3.
19. Park, H.; Lim, Y. Reinforcement Learning for Energy Optimization with 5G Communications in Vehicular Social Networks. *Sensors* **2020**, *20*, 2361. [[CrossRef](#)]
20. Khan, M.; Alam, M.; Moullec, Y.; Yaacoub, E. Throughput-Aware Cooperative Reinforcement Learning for Adaptive Resource Allocation in Device-to-Device Communication. *Future Internet* **2017**, *9*, 72. [[CrossRef](#)]
21. Zia, K.; Javed, N.; Sial, M.N.; Ahmed, S.; Pirzada, A.A.; Pervez, F. A distributed multi-agent RL-based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets. *IEEE Access* **2019**, *7*, 6733–6745. [[CrossRef](#)]
22. Li, Z.; Guo, C.; Xuan, Y. A multi-agent deep reinforcement learning based spectrum allocation framework for D2D communications. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Big Island, HI, USA, 9–13 December 2019; pp. 1–6.
23. Handouf, S.; Sabir, E.; Sadik, M. A pricing-based spectrum leasing framework with adaptive distributed learning for cognitive radio networks. In *International Symposium on Ubiquitous Networking*; Springer: Berlin, Germany, 2015; pp. 39–51.
24. Tathe, P.K.; Sharma, M. Dynamic actor-critic: Reinforcement learning based radio resource scheduling for LTE-advanced. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–4.
25. Sun, P.; Li, J.; Lan, J.; Hu, Y.; Lu, X. RNN Deep Reinforcement Learning for Routing Optimization. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; pp. 285–289.
26. Khodayari, S.; Yazdanpanah, M.J. Network routing based on reinforcement learning in dynamically changing networks. In Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05), Hong Kong, China, 14–16 November 2005; p. 5.
27. Islam, S.R.; Avazov, N.; Dobre, O.A.; Kwak, K.S. Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 721–742. [[CrossRef](#)]
28. Ding, Z.; Lei, X.; Karagiannis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195. [[CrossRef](#)]
29. Tabassum, H.; Ali, M.S.; Hossain, E.; Hossain, M.; Kim, D.I. Non-orthogonal multiple access (NOMA) in cellular uplink and downlink: Challenges and enabling techniques. *arXiv* **2016**, arXiv:1608.05783.