Article

A Study on the Impact of Integrating Reinforcement Learning for Channel Prediction and Power Allocation Scheme in MISO-NOMA System

Mohamed Gaballa [1,*], Maysam Abbod [1] and Ammar Aldallal [2]

1   Department of Electronic & Electrical Engineering, Brunel University London, Uxbridge UB8 3PH, UK
2   Department of Telecommunication Engineering, Ahlia University, Manama 10878, Bahrain
*   Correspondence: mohamedgaballa.gaballa@brunel.ac.uk

Abstract: In this study, the influence of adopting Reinforcement Learning (RL) to predict the channel parameters for user devices in a Power Domain Multi-Input Single-Output Non-Orthogonal Multiple Access (MISO-NOMA) system is inspected. In the channel prediction-based RL approach, the Q-learning algorithm is developed and incorporated into the NOMA system so that the developed Q-model can be employed to predict the channel coefficients for every user device. The purpose of adopting the developed Q-learning procedure is to maximize the received downlink sum-rate and decrease the estimation loss. To satisfy this aim, the developed Q-algorithm is initialized using different channel statistics and then the algorithm is updated based on the interaction with the environment in order to approximate the channel coefficients for each device. The predicted parameters are utilized at the receiver side to recover the desired data. Furthermore, based on maximizing the sum-rate of the examined user devices, the power factors for each user can be deduced analytically to allocate the optimal power factor for every user device in the system. In addition, this work inspects how the channel prediction based on the developed Q-learning model, and the power allocation policy, can both be incorporated for the purpose of multiuser recognition in the examined MISO-NOMA system. Simulation results, based on several performance metrics, have demonstrated that the developed Q-learning algorithm can be a competitive algorithm for channel estimation when compared to different benchmark schemes such as deep learning-based long short-term memory (LSTM), RL based actor-critic algorithm, RL based state-action-reward-state-action (SARSA) algorithm, and standard channel estimation scheme based on minimum mean square error procedure.

Keywords: RL; *Q*-learning; MISO-NOMA; KKT conditions

1. Introduction

The Non-Orthogonal Multiple Access (NOMA) system has been characterized as an inspiring multiple access form for upcoming wireless approaches to enhance the spectral efficiency and throughput [1]. NOMA system can develop the available resources more realistically by efficiently, taking into consideration the users' channel environments and also giving support to several users with distinctive Quality of Service (QoS) needs [2]. The integration of NOMA and multiple antenna techniques can be exploited to improve and reinforce system performance [3], therefore, inspecting Multiple Input-Single Output (MISO) NOMA system can be a good example in the direction of characterizing the expected upgrade in achievable data rates [4]. In downlink NOMA structure, the receiver device can receive a multiplexing of signals transmitted to several user terminals in the NOMA cell, thus eliminating the interference generated by other user devices come to be essential for coordinated detection. Frequently in power domain NOMA (PD-NOMA), multiuser detection can be handled via successive interference cancellation (SIC) [5]. In the SIC procedure, symbols from numerous users are decoded successively on the basis of the

Channel State Information (CSI) and power percentage designated for each user. A broad investigation of CSI for various users is demanding because pilot data that can be exploited in channel prediction, might interfere with symbols from other user terminals, therefore affecting the performance of a conventional prediction scheme, such as the Minimum Mean Square Error (MMSE) estimator [6]. Furthermore, power allocation policy is considered an essential issue for user devices when PD-NOMA is considered [7].

Deep Learning (DL) or Reinforcement Learning (RL) techniques, have the ability to track the differences in the channels among users and BS, thus, they are recently considered a powerful tool for upcoming radio systems [8,9]. Hence, allocating the power factors or estimating the CSI for user devices with the assistance of Machine Learning (ML) algorithms, triggered the authors for more deep investigations into this field in order to enhance the performance and detection process.

1.1. Related Works

Different techniques were introduced by authors in [10] to realize the optimal MMSE channel estimator in the Reconfigurable Intelligent Surfaces (RIS)-based MISO system. In the first technique, the authors suggest an analytical linear estimator to adjust the phase shift matrix of the RIS during the training phase, and the estimator based on that technique is shown to produce sensible accuracy compared to the least-squares method when the statistical properties of the applied channel and noise are considered. In the other approach, authors have expressed the channel prediction problem as an image denoising problem, then they introduce a Convolutional Neural Network (CNN) to achieve the denoising and predict the optimal MMSE channel parameters. Numerical outcomes have clarified that the proposed estimator based CNN algorithm can offer improved performance compared to the linear estimation method and low computational intricacy is preserved.

Toward enhancing the link reliability, a neural network model for a wireless channel estimator is proposed in [11] to be used with uncoded space-time diversity procedure in Multi Input Multi Output (MIMO) system. Based on the neural network ML structure, a channel estimator is suggested, and a mathematical scheme is presented to derive an optimum power transmission factors that can assist in lessening the channel prediction bandwidth utilization. Simulation results revealed that the channel estimator based on the proposed neural network structure can deliver an improvement in Bit Error Rate (BER) and Mean Square Error (MSE) compared to the standard MMSE channel estimation technique.

In a massive MIMO system and on the basis of a deep autoencoder scheme, authors in [12] performed experimental verifications on two tasks, one task for channel estimation modelling for wireless links, and the other task is belonging to a power allocation policy. The proposed deep learning autoencoder is also used to manage the issue raised from inadequate training datasets that may cause critical overfitting problems and consequently affect the model's reliability. Results based on the autoencoder procedure clarified that the suggested scheme could successfully enhance performance when the extent of the training dataset is mainly within a specified threshold selection.

To get over limitations raised when standard iterative power control techniques are utilized, such as high complexity and unnecessary latency, the work in [13] introduced a deep learning framework to manage these issues. In the presented structure, the outdated and partial CSI is exploited, and a Deep Neural Network (DNN) framework is created to construct an optimization problem to boost the spectral efficiency in device-to-device communication systems. User fairness and energy efficiency constraints were examined, and simulation outcomes showed that the proposed DNN model can attain better spectral and energy efficiency compared to the MMSE procedure when numerous channel correlation factors are considered.

Based on CSI, the position of each user device with respect to BS, and the path loss, a deep learning framework labelled PowerNet is introduced in [14]. The authors attempt to prove that it is possible to avoid the time consumption involved with intricate channel estimation procedures, and at the same time, power control can be managed. Different

from traditional DNNs that employ a fully connected structure, the presented PowerNet method utilizes a CNN layers to recognize the interference model through several links in wireless networks. Simulation outcomes revealed that the suggested PowerNet scheme can realize a stable performance without explicit channel estimation.

Recently, approximating the channel parameters or predicting the power factors with the assistance of Reinforcement learning (RL), is investigated by many researchers. The authors of [15] proposed an end-to-end channel estimation framework for a downlink multiuser multiple antenna system. The authors presented an RL-based actor-critic scheme for channel estimation without the assumption of ideal CSI. The authors mainly depend on the agent to bring and utilize the pilot symbols into the estimation process and then employ the estimated channel parameters to create downlink beamforming matrices. To satisfy the purpose of maximizing the sum rate reward, network parameters are adjusted based on the deep policy gradient method. The results proved that the suggested channel estimation algorithm can provide convergence and stable performance under various channel statistics and can perform better than the typical MMSE procedure when the sum rate metric is examined.

In [16], the authors developed a Deep Reinforcement Learning (DRL) method for device-to-device pairing to understand the correlation patterns between wireless networks. The introduced RL algorithm is adopted to explore the joint channel selection and power control problem for device-to-device pairing and to boost the weighted sum rate. Based on the suggested DRL learning procedure, each device-to-device pair can make use of the outdated and local information to understand the network parameters and perform decisions independently. Results showed that without a global CSI, the suggested DRL scheme is capable to attain a stable performance close to that achieved using standard analytical approaches.

The combination between a DNN as a tool for channel prediction and an optimized power scheme is explored in [17] for the purpose of multiuser detection in the NOMA system. The DNN based Long Short-Term Memory (LSTM) network is developed for channel prediction based on complex data processing. The DNN network is trained on the basis of both the correlation between successive training sequences and the normalised channel statistics. The efficiency of the suggested DNN based LSTM for channel prediction is inspected using different fading models and simulation outcomes, in terms of different performance metrics, have proved that the presented DNN scheme for channel estimation can provide a consistent performance compared to the MMSE procedure even when cell capacity is expanded.

1.2. Research Gap and Significance

Based on the preceding works, many of the proposed schemes that consider predicting the channel parameters task are mainly focused on implementing several deep neural networks (DNN) while applying RL approaches, which in turn leads to an increase in the number of hidden layers with a massive number of neurons in each layer. The significance of this study is to illuminate that we can eliminate the need for such DNN approaches, and instead, we can adopt the RL based developed Q-learning algorithm to predict the channel coefficients for each user device in MISO-NOMA cell, and at the same time, a notable improvement in system performance and network convergence is realized. The most prominent gain of the developed channel estimator scheme is that it can enhance the system performance without the need for hidden layers or an external training set.

In addition, several RL algorithms have been proposed to explicitly address the issues associated with channel state information (CSI), beamforming, and power allocation. To the best of the authors' knowledge, there is no study that explores the incorporation between Q-learning algorithm for channel prediction and the power allocation policy as an integrated scheme for multiuser detection in downlink MISO-NOMA system in fading channels.

Furthermore, it is worth mentioning that unlike deep learning algorithms, that mainly depend on learning from a training data set, the proposed Q-learning algorithm in our

study is developed to dynamically enhance the system performance and adjust to the variations in the channel based on the feedback from the environment.

1.3. Contributions to Knowledge

The channel prediction problem in downlink NOMA systems was considered in numerous works. In addition, there have been several works that apply machine learning (ML) to handle the channel estimation task in wireless communication systems. However, most of the current research on channel prediction in the NOMA systems based on ML is introduced via deep neural networks. To the best of the author's knowledge, currently, there is no research that manages the channel approximation task in a multiuser multi-input single-output NOMA system through an RL based Q-learning algorithm. The RL based Q-learning algorithm is developed based on maximizing the sum rates for all users in the network such that it can be used efficiently to predict the channel parameters for each user in the MISO-NOMA cell.

In addition, in this work, a structured mathematical analysis is introduced to formulate a non-complex analytical form for the power allocation for user devices in the examined MISO-NOMA system based on boosting the sum rate of the system while considering the constraints of the total power budget in the system, and the QoS for each user. Furthermore, the performance of the MISO-NOMA system is investigated when both the developed Q-learning algorithm for channel estimation and the derived power allocation scheme are jointly implemented. In this work, the contributions can be summed up as shown:

- In this study, a framework is proposed to illuminate how RL based Q-learning algorithm is developed based on maximizing the sum rates for all users in a MISO-NOMA system in order that it can be used dynamically to predict the channel parameters for each user in the MISO-NOMA cell.
- As a reference comparison, four further simulation environments are established. (1) the standard minimum mean square error (MMSE) based channel prediction scheme (Neumann et al.); (2) the DNN algorithm based on LSTM network for channel prediction applied in [17], (3) the RL based actor-critic procedure for channel prediction applied in [15], (4) the fourth simulation environment is dependent on applying RL based State-Action-Reward-State-Action (SARSA) procedure (Ahsan et al. and Mu et al.). The simulation outcomes of these environments are compared with the results of our proposed RL based Q-learning scheme, and the results emphasized that dependability can be assured by our developed Q-model for predicting channel parameters even when the number of devices in the cell is increased.
- To validate the efficacy of the developed Q-learning algorithm for channel prediction, the developed Q-model is investigated using Rayleigh and Rician fading channels.
- Evaluate the beneficial impact of cooperatively integrating the RL based Q-learning algorithm for channel prediction and the derived power allocation scheme for the purpose of multiuser recognition in the power domain MISO-NOMA system.
- The optimized power allocation scheme and the fixed power allocation scheme are both compared when the developed Q-learning scheme is implemented as a channel estimator.

The remainder of this paper is structured as follows. Section 2 describes the system model. Analysis of the optimization problem is presented in Section 3. The optimization framework and procedure are discussed in Section 4. The RL structure is introduced in Section 5. Section 6 discusses the Q-learning algorithm-based channel prediction. The RL-based Q-model architecture and channel estimation algorithm are summarized in Section 7. The simulation environment is described in Section 8, and simulation results are presented in Section 9. Lastly, conclusions are shown in Section 10.

Notation: bold lower-case letters denote vectors, bold upper-case letters denote matrices, and lower-case letters denote scalars. The subscript on a lower-case letter x_i represent ith element of vector x. $E(\cdot)$ refers to the expectation and $(\cdot)^{\mathrm{T}}$ refers to the transpose of the

vector. For two real numbers a $\leq$ b, [a, b] is the set for all real numbers in the range from a to b.

2. System Model

2.1. Multiuser Environment

In this work, a multiuser environment with a single Base Station (BS) and multiple user devices (UDs) is considered. The BS is supplied with N antennas and all the UDs are supplied with a single antenna. The network is assumed to work with equal length time intervals and each time interval includes one transmission, which contains either uplink or downlink transmissions. The pilot-assisted channel prediction is considered in this work, where pilot symbols can be identified by BS and UDs [15,17]. Each user device initially transmits its pilot symbols to BS via an uplink channel. Then, prior to data transmission, the BS can inspect the pilot symbols and the available network information to facilitate estimating the downlink CSI. The main aim of this work is to model the channel prediction task and to manage the power allocation scheme. We can refer to the matrix of downlink channel coefficients from BS with N antennas to UD i as:

$$\boldsymbol{H}_i = [\boldsymbol{h}_{1i}; \ \boldsymbol{h}_{2i}; \ldots; \boldsymbol{h}_{Ni}] \tag{1}$$

where $\boldsymbol{h}_{ji}$ represents the vector channel parameters from jth antenna at BS to the ith UD, with $j \in [1, 2, \ldots, N]$ and $i \in [1, 2, \ldots, M]$, where N is the number of antennas at BS and M is the number of users in MISO-NOMA cell. Furthermore, we can denote the data signal transmitted to UD i as

$$\boldsymbol{s}_i = [s_{i1}, \ s_{i2}, \ \ldots, s_{iK}] \tag{2}$$

where K is the length of the signal. Then, the matrix of all the UD's sequences can be expressed as

$$\boldsymbol{S} = [\boldsymbol{s}_1; \boldsymbol{s}_2; \ldots; \boldsymbol{s}_M] \tag{3}$$

The received kth signal at jth UD can be denoted as:

$$\boldsymbol{y}_{kj} = \sum_{i=1}^{N} \boldsymbol{h}_{ij} s_{ki} + z_{kj} \tag{4}$$

where z_{kj} denotes the AWGN with zero mean and variance σ^2 at jth UD through kth signal duration. The received kth symbol at all UDs is:

$$\boldsymbol{Y}_k = \sum_{i=1}^{N} \boldsymbol{h}_i s_{ki} + \boldsymbol{z}_k \tag{5}$$

where

$$\boldsymbol{Y}_k = [\boldsymbol{y}_{k1}; \boldsymbol{y}_{k2} ; \ldots ; \boldsymbol{y}_{kM}] \tag{6}$$

$$\boldsymbol{z}_k = [z_{k1}; \ z_{k2}; \ldots; z_{kM}] \tag{7}$$

Many of the current works depend on pilot symbols to approximate the uplink channel parameters and then utilize channel reciprocity to realize the prediction of downlink channel weights [15,18]. These schemes for CSI prediction may not be reliable, especially in cases of inadequate channel reciprocity owing to hardware constraints. Furthermore, this kind of estimator may introduce estimation errors in case the uplink and downlink channel parameters are not stationary within a certain transmission.

In the developed Q-learning procedure, we plan to get assistance from the pilot symbols, and network information to explicitly predict the downlink channel parameters. The set of estimated channel coefficients among BS and M UDs can be indicated as

$$\hat{\mathbf{H}} = \left[\hat{\mathbf{H}}_1; \ \hat{\mathbf{H}}_2; \ldots; \hat{\mathbf{H}}_M\right] \tag{8}$$

where $\hat{\mathbf{H}}_i$ is the predicted matrix channel coefficients between BS that contains N antennas and ith UD, and can be expressed as follows:

$$\hat{\mathbf{H}}_i = \left[\hat{h}_{1i}; \, \hat{h}_{2i}; \dots; \hat{h}_{Ni} \right] \tag{9}$$

where $\hat{h}_{ji}$ represents the predicted channel parameters between jth antenna at BS and the ith UD.

2.2. MISO-NOMA Environment

The fundamental idea of NOMA is to achieve non-orthogonal resource allocation between users while increasing the processing at the receiver side [19]. With non-orthogonal resource allocation, NOMA can attain massive connectivity and accomplish higher spectral efficiency. Existing research on the NOMA system mainly focuses on the code domain and power domain. In the code domain NOMA, distinct spread-spectrum codes are designated to different users and then multiplexed over the same time-frequency resource block. In the power domain NOMA (PD-NOMA) [19], the transmitter superimposes signals with different power levels to be sent to several users on the shared spectrum. At the receiver, each user can decode his own desired signal by means of successive interference cancellation (SIC).

In this subsection, the downlink MISO-NOMA system is explored where user devices and BS are linked by different fading channels. NOMA cell is assumed where one BS with two antennas is implemented to assist user devices (UDs), and each device terminal has one antenna. In PD-NOMA [19], user devices receive the superimposed signal sent from BS which involves target and interfering signals sent through the same resources. Thus, combining different signals supported by unique power portions is critical to distinguish signals and strengthen the successive interference cancellation (SIC) technique. The system structure for the basic components implemented in the examined MISO-NOMA system is shown in Figure 1.
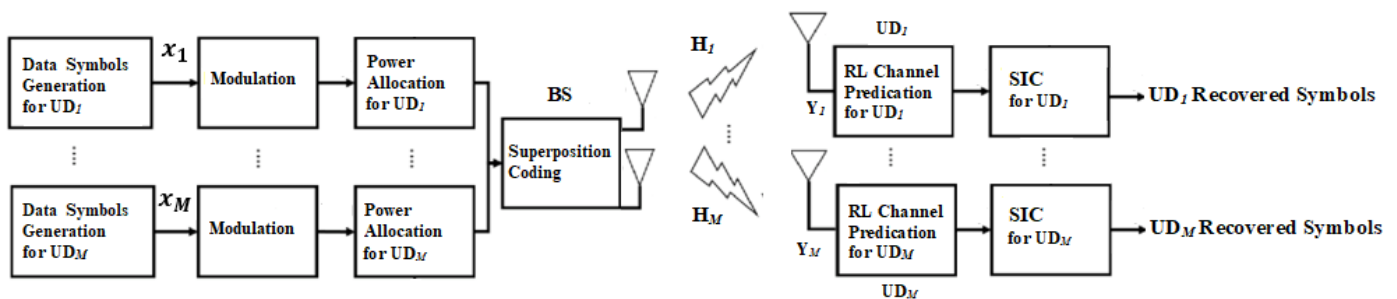

Figure 1. MISO-NOMA system basic Structure based RL channel prediction.

In our observed MISO-NOMA cell, three user devices are considered in the cell, and the examined user devices are identified corresponding to their fading channels and the distances from BS. Fading channels with Rayleigh distribution are adopted to characterize the channel model for every user. The user terminal at the boundary of the cell is realized as a far user, while the nearest user equipment is designated as a near user terminal. The examined cell contains three user devices and the fading path can be distinguished for every user as follows [3]: $h_n \sim \left(0, d_n^{-k} \right)$ for near users, $h_m \sim \left(0, d_m^{-k} \right)$ for the middle user, and $h_f \sim \left(0, d_f^{-k} \right)$ for the user at the edge of the cell, where h_i implies a vector represents the fading path coefficients among BS and user i. Path loss exponent is represented by k, and AWGN is considered with noise power indicated as σ^2. In terms of channel gains, the relation between user devices can be indicated as $|h_n|^2 > |h_m|^2 > \left| h_f \right|^2$ [20] and overall power transmitted from BS to all users in the cell is labelled as P_t. Every user device contains a receiving element that can activate the SIC process to get rid of signals related to other devices with bad channel environments. In contrast, signals related to user devices

with good link conditions may not be separated and regarded as interference. According to the aforementioned assumptions, the superposition-coded signal x sent from BS can be stated as follows [3,17]:

$$x = \sqrt{P_t}\left(\sqrt{\eta_n}x_n + \sqrt{\eta_m}x_m + \sqrt{\eta_f}x_f\right) \tag{10}$$

where η_f, η_m and η_n represent power factors given for a far device, middle device, and near device separately. Furthermore, x_f, x_m and x_n refer to the signal vectors related to far, middle, and near users respectively. The received downlink signal at a far device in the MISO-NOMA cell can be shown as:

$$y_f = xh_{f1} + xh_{f2} + z_f \tag{11}$$

where h_{f1} represents the channel coefficients among a far device and the 1st antenna at BS, h_{f2} represents the channel coefficients among the far device and 2nd antenna at BS and z_f is AWGN noise component at the far device with mean zero and variance σ^2. The far user is signified by weak link condition, and signal x_f is usually given further power percentage by BS where $\eta_f > \eta_m > \eta_n$. The obtained signal at a far device can be formulated as:

$$y_f = \sqrt{P_t\eta_f}x_f\left(h_{f1} + h_{f2}\right) + \left(\sqrt{P_t\eta_m}x_m + \sqrt{P_t\eta_n}x_n\right)\left(h_{f1} + h_{f2}\right) + z_f \tag{12}$$

The 1st term in (12) implies the target signal for far device and the 2nd term indicates the interference term from other user devices. The possible bit rate for a far device could be shown as [3,21]:

$$R_f = \log_2\left(1 + \frac{\left|h_{f1} + h_{f2}\right|^2 P_t\eta_f}{\left|h_{f1} + h_{f2}\right|^2 P_t(\eta_n + \eta_m) + \sigma^2}\right) \tag{13}$$

Typically, the near user device has a good link status alongside BS, therefore, a low power factor can be assigned to x_n, and the near user received signal can be stated as

$$y_n = \sqrt{P_t\eta_n}x_n(h_{n1} + h_{n2}) + \left(\sqrt{P_t\eta_m}x_m + \sqrt{P_t\eta_f}x_f\right)(h_{n1} + h_{n2}) + z_n \tag{14}$$

In Equation (14), the 1st term represents the anticipated signal, and the 2nd term implies interference from other devices. It can be noted from Equation (14), that the interference can be principal since the far user may be assigned a further power percentage. Thus, at a near device, SIC is accomplished, where direct decoding for the far user signal x_f is implemented first, then eliminated from the aggregate signal. After that, the middle device signal x_m is decoded and gets rid of it from the resultant signal and the possible rate for a near user R_n can be shown as:

$$R_n = \log_2\left(1 + \frac{|h_{n1} + h_{n2}|^2 P_t\eta_n}{\sigma^2}\right) \tag{15}$$

3. Optimization Problem Characterization

The key objective here is to maximize the sum rates for user devices in the MISO-NOMA cell. Sum rate maximization is considered based on optimizing the power coefficients for each user terminal in compliance with the status of the channel between each user and the BS. In downlink MISO-NOMA, the objective function or the sum rates for M user devices can be formulated as [3,22]:

$$R_{sum} = \sum_{i=1}^{M} \log_2\left(1 + \frac{|h_{i1} + h_{i2}|^2 P_t\eta_i}{|h_{i1} + h_{i2}|^2 P_t \sum_{j=1}^{i-1}\eta_j + \sigma^2}\right) \tag{16}$$

In the optimization problem, the constraints can be presented as follows:

3.1. Power Constraint

The power designated for every user device in the cell is a fraction of the whole power P_t sent from BS, therefore the power percentage for each device must conform with [22]:

$$\sum_{i=1}^{M} \eta_i \leq 1 \tag{17}$$

where η_i is the power percentage allocated for the *i*th user.

3.2. QoS Constraints

In our analysis, we consider that all the user devices in the examined MISO-NOMA cell need to satisfy a QoS requirement where the minimum rate R_{min} is required to be realised in the system [22,23], this constraint can be expressed as follows:

$$Log_2(1 + SINR_i) \geq R_{min} \tag{18}$$

where $SINR_n$ is the signal-to-interference plus noise ratio for *i*th user and R_{min} is the minimum required transmission rate in the examined MISO-NOMA cell. The expression in (18) can be redeveloped as follows [24]:

$$|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 \rho \left(\eta_i - (2^{R_{min}} - 1) \sum_{j=1}^{i-1} \eta_j \right) > (2^{R_{min}} - 1) \tag{19}$$

where ρ represents the SNR and η_j is the power percentage given for *j*th user device.

4. Optimization Framework

The main aims in this part include the following: (1) present the objective function and the constraints in a standard form, (2) find a general expression for the 1st and 2nd derivative of the objective function, (3) based on the mathematical analysis and the derived formulas, we can inspect that $\frac{\partial^2 R_{Sum}}{\partial \eta_i^2}$ is a negative function, which validates that the objective function is a concave with distinctive global maximum, and (4) finally, we deduce the optimal power factors for each user based on applying the Lagrange function and the KKT necessary conditions.

On the basis of the objective function in (16) and the constraints in (17) & (19) and the fact that there are two antennas at the BS and one antenna at each user terminal, the standard optimization problem can be generally reformulated as follows [24,25]:

$$\max_{\eta} R_{sum} = \sum_{i=1}^{M} \log_2 \left(\frac{|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t \sum_{j=1}^{i-1} \eta_j + \sigma^2 + |\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t \eta_i}{|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t \sum_{j=1}^{i-1} \eta_j + \sigma^2} \right) \tag{20}$$

such that

$$\sum_{j=1}^{M} \eta_j \leq 1$$

$$(2^{R_{min}} - 1) - \rho |\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 \left(\eta_i - (2^{R_{min}} - 1) \sum_{j=1}^{i-1} \eta_j \right) \leq 0$$

$$\eta_i \geq 0 \quad \forall i = 1, 2, \ldots, M$$

In this part, the power optimisation framework is accomplished with regards to three user devices in the MISO-NOMA cell, therefore, the examined constraints can be represented as shown [25,26]:

$$\psi_1(\eta) = \eta_n + \eta_m + \eta_f - 1 \tag{21}$$

$$\psi_2(\eta) = (2^{R_{min}} - 1) - \rho \left| \boldsymbol{h}_{f1} + \boldsymbol{h}_{f2} \right|^2 \Big(\eta_f - (2^{R_{min}} - 1)(\eta_m + \eta_n) \Big) \tag{22}$$

$$\psi_3(\eta) = (2^{R_{min}} - 1) - \rho |\boldsymbol{h}_{m1} + \boldsymbol{h}_{m2}|^2 \Big(\eta_m - (2^{R_{min}} - 1)(\eta_n) \Big) \tag{23}$$

Since the constraints $\psi_1(\eta)$, $\psi_2(\eta)$ & $\psi_3(\eta)$ are linear in terms of η, they are considered convex.

Typically, to prove that the objective function R_{Sum} is concave with a distinctive global maximum, we need to find the first derivative $\frac{\partial R_{Sum}}{\partial \eta_i}$ and the second derivative $\frac{\partial^2 R_{Sum}}{\partial \eta_i{}^2}$ of the objective function [3,24]. The first derivative of the objective function can be deuced in general form as follows [23]:

$$\begin{aligned}
\frac{\partial R_{Sum}}{\partial \eta_i} = \frac{1}{\ln 2} & \left(\frac{|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t}{|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t \sum_{j=1}^{i} \eta_j + \sigma^2} \right) \\
& - \frac{1}{\ln 2} \sum_{k=1}^{M-i} \left\{ \left(\frac{(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t)^2 \eta_{i+k}}{(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \sum_{j=1}^{i+k} \eta_j + \sigma^2)} \right) \right. \\
& \left. \times \left(\frac{1}{(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \sum_{j=1}^{i+k-1} \eta_j + \sigma^2)} \right) \right\}
\end{aligned} \tag{24}$$

Similarly, the second derivative of the objective function can be derived in general form as follows [23,24]:

$$\begin{aligned}
\frac{\partial^2 R_{Sum}}{\partial \eta_i{}^2} = -\frac{1}{\ln 2} & \left\{ \left(\frac{(|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t)^2}{(|\boldsymbol{h}_{i1} + \boldsymbol{h}_{i2}|^2 P_t \sum_{j=1}^{i} \eta_j + \sigma^2)^2} \right) \right. \\
& - \sum_{k=1}^{M-i} \left\{ \left(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \right)^3 \eta_{i+k} \right. \\
& \times \left(\frac{[2(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \sum_{j=1}^{k+i-1} \eta_j + \sigma^2) + |\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \eta_{i+k}]}{(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \sum_{j=1}^{i+k} \eta_j + \sigma^2)^2} \right) \\
& \left. \left. \times \left(\frac{1}{(|\boldsymbol{h}_{(i+k)1} + \boldsymbol{h}_{(i+k)2}|^2 P_t \sum_{j=1}^{i+k-1} \eta_j + \sigma^2)^2} \right) \right\} \right\}
\end{aligned} \tag{25}$$

Based on the above mathematical analysis and the derived formulas, we can inspect that $\frac{\partial^2 R_{Sum}}{\partial \eta_i{}^2}$ is a negative function, which verifies that the objective function is a concave with a distinctive global maximum [3,24,27]. To derive the optimal power factors, the Lagrange function and the KKT necessary conditions can be applied [28].

$$\mathcal{L}\Big(\eta_n, \eta_m, \eta_f, \mu_1, \mu_{2,}, \mu_3\Big) = R_{Sum} - \mu_1 \psi_1(\eta) - \mu_2 \psi_2(\eta) - \mu_3 \psi_3(\eta) \tag{26}$$

where μ_1, μ_2, and μ_3 represent Lagrange multipliers for the 3 users' scenario.

- Optimality conditions can be written as follows [3,24,27]:

$$\frac{\partial R_{Sum}}{\partial \eta_n} - \mu_1 \frac{\partial \psi_1(\eta)}{\partial \eta_n} - \mu_2 \frac{\partial \psi_2(\eta)}{\partial \eta_n} - \mu_3 \frac{\partial \psi_3(\eta)}{\partial \eta_n} = 0 \tag{27}$$

$$\frac{\partial R_{Sum}}{\partial \eta_m} - \mu_1 \frac{\partial \psi_1(\eta)}{\partial \eta_m} - \mu_2 \frac{\partial \psi_2(\eta)}{\partial \eta_m} - \mu_3 \frac{\partial \psi_3(\eta)}{\partial \eta_m} = 0 \tag{28}$$

$$\frac{\partial R_{Sum}}{\partial \eta_f} - \mu_1 \frac{\partial \psi_1(\eta)}{\partial \eta_f} - \mu_2 \frac{\partial \psi_2(\eta)}{\partial \eta_f} - \mu_3 \frac{\partial \psi_3(\eta)}{\partial \eta_f} = 0 \tag{29}$$

Given the fact that $|\mathbf{h}_n|^2 > |\mathbf{h}_m|^2 > \left|\mathbf{h}_f\right|^2$, we can demonstrate that the analyzed constraints are feasible [3] and after a few mathematical manipulations the closed form for the power factors η_f, η_m, and η_n can be deduced as follows [27]:

$$\eta_f = \left(\frac{(2^{R_{min}} - 1)}{2^{R_{min}}}\right)\left(1 + \frac{1}{\rho\left|\mathbf{h}_{f1} + \mathbf{h}_{f2}\right|^2}\right) \tag{30}$$

$$\eta_m = \left(\left(\frac{(2^{R_{min}} - 1)}{2^{R_{min}}}\right)\left(1 + \frac{1}{\rho|\mathbf{h}_{m1} + \mathbf{h}_{m2}|^2}\right) - \left(\frac{2^{R_{min}} - 1}{2^{R_{min}}}\right)^2\left(1 + \frac{1}{\rho|\mathbf{h}_{f1} + \mathbf{h}_{f2}|^2}\right)\right) \tag{31}$$

$$\eta_n = 1 - \left(\eta_m + \eta_f\right)$$

$$\eta_n = \frac{1}{(2^{R_{min}})}\left(\left(\frac{1 + \rho|\mathbf{h}_{f1} + \mathbf{h}_{f2}|^2}{(2^{R_{min}})\rho|\mathbf{h}_{f1} + \mathbf{h}_{f2}|^2}\right) + \left(\frac{(2^{R_{min}} - 1)}{\rho|\mathbf{h}_{m1} + \mathbf{h}_{m2}|^2} - \frac{1}{\rho|\mathbf{h}_{f1} + \mathbf{h}_{f2}|^2}\right)\right) \tag{32}$$

5. Reinforcement Learning Framework

Typically, RL is developed on the basis of a Markov Decision Process (MDP) design, that contains basic elements [29,30]: a state space '*S*', which is the set of states or observations in the environment and these states can be observed by the agent. An action space '*A*', which is the set of actions that can be selected by the agent at each state. An instantaneous reward '*R*', which is the direct reward that is given to the agent after selecting an action $a \in A$ to transfer to a state $s \in S$. Policy '*P*' represents the mapping criteria to move from the current observed state to a new state based on the action that will be taken by an agent. Another important element in the RL process is the State-action value function $Q(s, a)$, which is formally described as the expectation or the average of cumulative discounted rewards when an action $\mathbf{a} \in A$ is selected by an agent in the state $\mathbf{s} \in S$ when a certain policy is considered. Furthermore, RL can be considered a method of understanding the agent's interaction in a stochastic environment by successively selecting actions during a sequence of time periods. Therefore, the main aim of reinforcement learning is to train an agent to carry out a certain task within an uncertain environment [30].

The interaction between the agent and the environment can be described as follows: at each time period, the agent can recognize the observations or states in the environment, and based on the current observation, the agent can identify and carry out a specific action. Then, an immediate reward will be sent from the environment to the agent. The reward is a measure of how effective the action is, when the agent performs a certain action to achieve a specific goal [31]. Basically, at each learning time interval, the RL agent interacts with the environment by following a particular policy that controls the transition between state space to action space.

Based on the aforementioned discussion and as shown in Figure 2, the RL agent can be essentially represented by two elements: a policy and a learning algorithm [32]. The policy is the mapping criterion that chooses actions on the basis of the observations or status observed in the environment. Usually, the policy can be represented as a function with tunable parameters, such as DNN, while the learning algorithm constantly improves the parameters of the policy based on observations, actions, and rewards [33]. In general, the objective of the learning algorithm is to realize the best possible policy that can maximize the expected cumulative long-term reward received during the task.
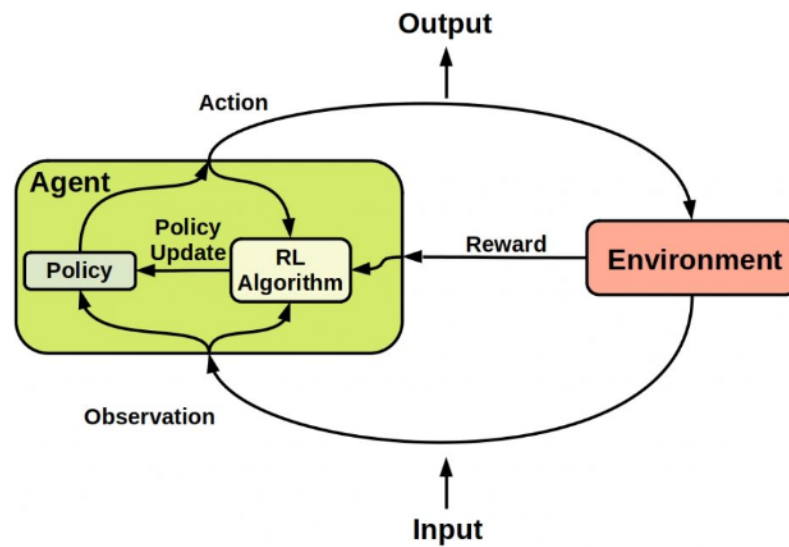
Figure 2. Reinforcement Learning Framework.

6. Channel Estimation Based Q-Learning Algorithm

In the considered channel prediction scheme, it is assumed that the action spaces are discrete, therefore, we manage to use an RL-based Q-learning procedure as one of the candidates of RL schemes for parameters update in our examined cell [34,35]. The Q-learning algorithm is categorized as a model-free, and off-policy reinforcement learning procedure, also a Q-learning agent is characterized as a value-based RL agent that has the role of updating a specific critic value function to enhance the future rewards. At a certain state, the agent can inspect and select the action for which the expected reward is maximized. In this section, RL based Q-learning is employed for channel prediction tasks in the MISO-NOMA cells where pilot symbols are also adopted to assist in the channel estimation process [36]. Therefore, it is assumed that there is coordination between BS and user devices such that the pilot symbols can be recognized at the BS and user terminals. In our work, we have considered the BS as the Q-learning agent, and we assume that the BS will start estimating the channel parameters for each user after user devices complete sending the pilot signals [37]. Therefore, in our developed RL based Q-learning algorithm can be utilized to estimate the CSI after the BS receives the pilot signals.

The scenario for the channel prediction process based on the developed Q-learning model can be outlined in this way [38]. Firstly, at the start of each transmission time slot, user devices can send pilot symbols to BS across the uplink channel. Secondly, on the basis of the developed RL based Q-learning algorithm and availability of network information such as user's distance and path loss, BS (agent) can predict the downlink CSI for user devices. Thirdly, BS will generate the superposition coding signal and performs downlink data transmission. Finally, the receiver of each user terminal will receive the downlink transmitted data and the estimated channel parameters based on Q-learning algorithm will be utilized to decode the desired signal. In addition, each user device can feedback the signal-to-interference plus noise ratio (SINR) or the achieved rate to the BS to enhance the detection process.

In this study, the main objective of the developed RL based Q-learning algorithm is to maximize the downlink sum rate and reduce the estimation loss. Instead of estimating the received signal, we primarily concentrate on incorporating the developed Q-learning model in the NOMA system for the purpose of channel estimation [39]. The RL-based Q-agent is designed to estimate the channel parameters by interacting with the environment, hence strict orthogonal pilot symbols are not required as shown in the standard procedures. Throughout the learning iteration, the Q-learning agent decides on the action that can enhance the approximated state-action value function $Q(s, a)$ therefore, the expected long-term reward can be also maximized in the neural networks. It is worth

mentioning that when increasing the number of learning iterations, updating Q-values becomes more sufficient, and an improved channel approximation and sum rate reward can also be achieved [34,36,40].

In the proposed Q-learning scheme, the sum rate is presented at the learning time interval t as R_t, hence, the instantaneous sum rate at time instant t can be shown as follows [15,34]

$$R_t = \sum_{i=1}^{M} \log(1 + SINR_{it}) \tag{33}$$

where $SINR_{it}$ is the signal-to-interference plus noise ratio of user i at time instant t and M is the number of users in the MISO-NOMA cell. In this work, the optimum goal of the developed Q-learning algorithm is to maximize the total discounted reward R^γ starting from time instant t, which can be denoted as

$$R_t^\gamma = \sum_{k=t}^{\infty} \gamma^{k-t} R_{k+1} \tag{34}$$

where R_t^γ is the discounted reward at time slot t, and γ is the discount factor. Substituting the sum rate from (33) into (34), the discounted sum rate reward, can be expressed as [41]:

$$R_t^\gamma = \sum_{l=t}^{\infty} \gamma^{l-t} \sum_{i=1}^{M} \log\left(1 + SINR_{i(l+1)}\right) \tag{35}$$

As previously stated, the Q-learning agent is the BS, whose aim is to boost the accumulative transmission sum rate. Therefore, two value functions can be inspected while considering the RL maximization problem [34,36,42], the first one is the state value function $V(s)$

$$V(s) = E[R^\gamma / (S_t = s)] \tag{36}$$

and the other one is the state-action value function $Q(s, a)$

$$Q(s, a) = E[R^\gamma / (S_t = s, A_t = a)] \tag{37}$$

where E denotes the expected value given that the agent follows a certain policy within the applied procedure. Due to unspecified transition probabilities and limited observed states, an optimal policy is difficult to achieve. Therefore, the Q-learning procedure is developed to approximately achieve the best possible policy. In the developed Q-learning procedure, the state-action value function $Q(s, a)$ values are learned via trial and error and are updated according to the following formula [15,34,36,42]:

$$Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha\left[R(s, a) + \gamma\max_{a' \in A}Q(s', a')\right] \tag{38}$$

where α is the learning rate, s' denotes the new state, and a' is the new action that will be considered by the agent from the action space A to maximize the new state-action value function $Q(s', a')$.

7. Q-Learning Network Architecture

Basically, in data transmission, the frame transmitted includes data and pilot symbols. It is supposed that the implemented channel model is stationary throughout one frame transmission of data and pilot signals and the channel parameters are varying from one frame to another. The basic architecture of the channel prediction scenario based on the developed Q-learning procedure employed in our examined network is illustrated in Figure 3, which primarily consists of several stages [17,43].
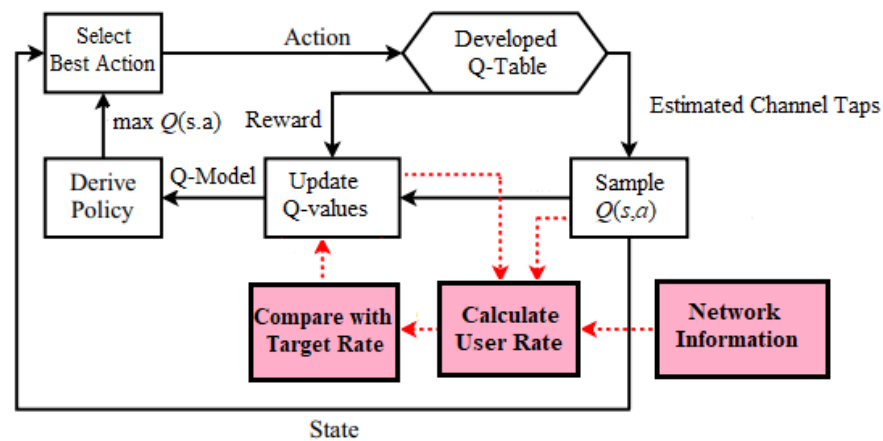
Figure 3. The architecture of the proposed Channel prediction scheme based developed Q-Learning algorithm.

In the first stage, initial channel parameters will be created based on a distinct Rayleigh channel models. In the second stage, we initialize the $\boldsymbol{Q}$ table and initialize the reward matrix $\boldsymbol{R}$ with zero values. The signal-to-interference plus noise ratio (SINR), and the minimum required rate R_t, can be calculated for every user device in the MISO-NOMA cell with the aid of the availability of the network information such as the initial assigned power percentage for each user terminal, and the entire power transmitted from BS P_T. Primarily, the Q-values can be adjusted based on the difference between the assigned target rate $\boldsymbol{R}_T$ and the initial generated user rate for each device. In the third stage, the best action will be explored and implemented by the $\boldsymbol{Q}$ agent, and then updating the values for the Q-table that represent the observation action pair $Q(s, a)$. Furthermore, the values for the reward matrix $\boldsymbol{R}$ will be dynamically assigned according to the actions executed by the Q-agent.

In the fourth stage, the state action value function $\boldsymbol{Q}(\boldsymbol{s}, \boldsymbol{a})$ that represent the values for the Q-table will be modified according to a Q-learning procedure with the aid of the following parameters, the discount factor γ, the assigned immediate reward matrix $\boldsymbol{R}$, and the learning rate $\boldsymbol{\alpha}$. Throughout the learning phase, the generated state action values $Q(s, a)$ will be sampled to calculate the new channel rate and at the same time update the Q-table until the optimum rate or the terminal state is achieved.

Dataset Preparation

Essentially, path loss and the distance between every user terminal and the BS need to be specified in the dataset to facilitate the random generation of the channel weights for every user device in the examined MISO-NOMA network [43]. In the beginning, pilot symbols are created, transmitted, and identified at the BS and at the receiver of every device. Additionally, power factors for every device in the cell need to be initially assigned. The channel weights for every device in the cell are initialized to set up the Q-table values, and during the algorithm iterations, the Q-values are modified according to a Q-learning procedure [34–36].

Throughout the learning process and for the sake of updating the Q-table, the discount factor γ, learning rate $\boldsymbol{\alpha}$, the target rate $\boldsymbol{R}_T$, current state, and the terminal state should be identified. In our developed Q-learning algorithm, the Q-agent will choose the next state at random and set it as the next state, then the Q-learning agent will inspect all possible actions available to move to the next state. Next, the Q-learning agent will carefully identify the best action $\boldsymbol{a}$, that satisfies the maximum value for $Q(s, a)$ to move to the new state. After moving to the new state, a reward value will be assigned to the agent as a measure of how successful this transition was in order to move to the new state [44]. During the update phase, we compute ΔQ, which represents the difference between the new generated

value function and the preceding value function of $Q(s, a)$. Then, update the resultant $Q(s, a)$ value in the Q-table according to the following formula.

$$Q(s, a) = Q(s, a) + \alpha \cdot \Delta Q \tag{39}$$

Based on the updated Q-values in Q-table and the updated channel gain, a new achieved rate can be calculated and compared to the target rate for each user device in the cell. In the developed Q-learning algorithm, once the optimum rate or the terminal state is reached, the developed Q-matrix will be employed to compose the channel taps for each user device. The developed Q-learning procedure for channel approximation can be summarized as presented in Algorithm 1.

Algorithm 1: Developed Q-learning Channel Prediction Structure.

---

1. Initialize the Q table values and initialize the reward matrix R with zeroes.

Inputs

2. Number of Iterations and the size for the channel parameters for every user device.
3. Initial distance "d_i" of every user device from the BS.
4. Path loss parameter "ϑ".
5. Design random pilot symbols.
6. Initialize the random channel parameters for each user "h_{ij}" based on fading model, $j \in [1, 2, \ldots, N]$ and $i \in [1, 2, \ldots, M]$. N is the number of antennas at BS and M is the number of devices in the cell.
7. Designate the power percentage "η_i" for each user.
8. Determine system bandwidth "B", Total transmit power "P_T", and noise spectral density "N_o"
9. Assign the desired channel parameters "h_{id}" and the target rate "R_T"

Procedure

10. Based on the channel gain $\left|h_{ij}\right|^2$, total transmit power "P_T", and initial power factor for each user "η_i", signal to interference noise ratio "$SINR_i$", minimum required rate "R_i" can be calculated for each device.
11. At each iteration, compare the initial generated rate "R_i" with the target rate "R_T".
12. Update the values for the Q-table that represent the current state and action pair $Q(s, a)$.

Q-algorithm

13. identify discount factor "γ", learning rate "α", the current state, and the terminal state.
14. Choose the next state at random and set it as the next new state.
15. Inspect all possible actions "a_i" to move to the new state.
16. Select the best action $a_i \in A$, which satisfies the maximum value for the Q-value function argmax $Q(s, a)$ to move to the new state.
17. Identify the immediate Reward "R", based on the action implemented to move to the new state.
18. Based on the following: (1) maximum Q-value $Q(s, a)$ obtained in (16), (2) the corresponding reward "R", (3) the discount factor "γ", then $Q(s, a)$ can be updated based on bellman's equation

$Q(s, a) \leftarrow R + \gamma$ argmax $Q(s, a)$

Outputs

19. Based on the updated $Q(s, a)$ values in Q-table, the channel coefficients "h_{ij}" and channel gain $\left|h_{ij}\right|^2$ can be updated and a new user rate can be calculated and compared to the target rate "R_T".
20. Compute the difference "ΔQ" between the updated value function $Q_{new}(s, a)$ and the previous $Q(s, a)$.
21. Based on (20), $Q(s, a)$ value in the Q-table can be further updated according to $Q(s, a) \leftarrow Q(s, a) + \alpha \cdot \Delta Q$
22. Check whether the terminal state has been reached or the episode has been completed.
23. Compose predicted channel taps $\hat{h}_i$

---

8. Simulation Environment

Characterization of the simulation parameters and settings is discussed in this section. The examined downlink MISO-NOMA system contains three distinct user devices and one BS in which the BS is supplied with two antennas and every user device in the cell is provided with a single antenna. In the examined NOMA structure, the modulated signals in downlink transmission are superimposed and transferred by BS to user devices via independent Rayleigh or Rician fading channels that are influenced by AWGN with noise power density assigned as $N_0 = -174$ dBm/Hz and the path loss is set to 3.5. MATLAB software is utilized as a simulation tool to satisfy the following aims, (1) inspect, characterize, and evaluate the performance of the developed RL based Q-learning algorithm when implemented as a channel estimator in the considered MISO-NOMA system, (2) investigate the reliability of incorporating the developed Q-algorithm as channel estimator scheme with the optimized power scheme in the examined MISO-NOMA network, and performance metrics are considered to explore the impact of this integration. (3) optimized power allocation scheme and fixed power allocation scheme are both compared when the developed Q-learning scheme is utilized as a channel estimator in the cell. Monte-Carlo simulations are performed with $N = 10^5$ iterations, and at the outset of each iteration, pilot symbols are randomly generated and recognized at the BS and each device. The main simulation parameters are summarized in Table 1.

Table 1. Summary of Simulation Parameters.

Parameter	Value
Simulation Tool	MATLAB
Modulation type	QPSK
Number of Users	3, [2–10]
System Bandwidth B	1000 kHz
Fading channel	(Rayleigh, Rician)
Path loss exponent	3.5
Number of Iterations	10^5
Noise PSD N_0	-174 dBm/Hz
Learning Rate α	0.1
Discount factor γ	0.9

The presented Simulation figures are generated based on the assumption that the channel coefficients are not available at each user device. Thus, in order to examine the effectuality of the developed RL based Q-learning procedure, and for the sake of comparison, four further simulation environments are established, (1) standard minimum mean square error (MMSE) based channel prediction scheme [45]; (2) DL algorithm based on LSTM network for channel prediction applied in [17], (3) RL based actor-critic procedure for channel prediction applied in [15], (4) the fourth simulation environment is dependent on applying RL based State-Action-Reward-State-Action (SARSA) procedure (Ahsan et al., Mu et al. and Jiang et al.). Throughout the simulations, we point out to MMSE technique as conventional NOMA, to denote that user devices are applying the MMSE technique for predicting the channel state information (CSI) prior to reconstructing the desired signal.

In the simulation environment, NOMA parameters are generated on the basis of the LTE standard [46,47], and channel parameters are created to initially model the Rayleigh fading channels based on the ITU models. In our developed Q-learning algorithm, at the end of the training episode, or if the terminal state is reached, the updated $Q(s, a)$ values in the Q-table will be employed as a practical channel coefficients for the user devices. Different power percentages are initially assigned for every user device according to channel gain and based on the existing distance from the BS. Power factors η_n, η_m, and

η_f are specified for near, middle, and far users respectively. In a fixed power allocation setup, we designate $\eta_f = 0.65$, $\eta_m = 0.25$, and $\eta_n = 0.1$. In the optimized power structure (OPS), power factors are allocate d for user devices in proportion to the analytical formula concluded previously for every device in Section 4. In the simulation files, the transmission distance for each user device with respect to BS is assigned as follows: $d_f = 900$ m, $d_m = 400$ m, and $d_n = 100$ m. Data and pilot symbols are modulated using Quadrature phase shift keying (QPSK) as the modulation format and the applied transferred power is mostly varying from 0 to 30 dBm.

9. Simulation Results and Discussion

Simulation outcomes that clarify the comparison between the developed RL based Q-learning algorithm and the conventional NOMA scheme that applies MMSE method to predict the channel coefficients for each device are shown in Figure 4 in terms of BER versus power transmitted. The predicted channel parameters using both schemes are employed for the signal detection for each user device and the simulated results are shown where fixed power allocation (FPA) is considered. When the developed Q-algorithm is applied for channel estimation, each user device in the examined MISO-NOMA cell provides a noticeable improvement in lowering the BER compared to the MMSE procedure. At particular BER values such as 10^{-2}, the attained power saving by the Q-learning algorithm is within 2 dBm for far and middle user devices, while a power reduction within 1 dBm is recorded for the near user.
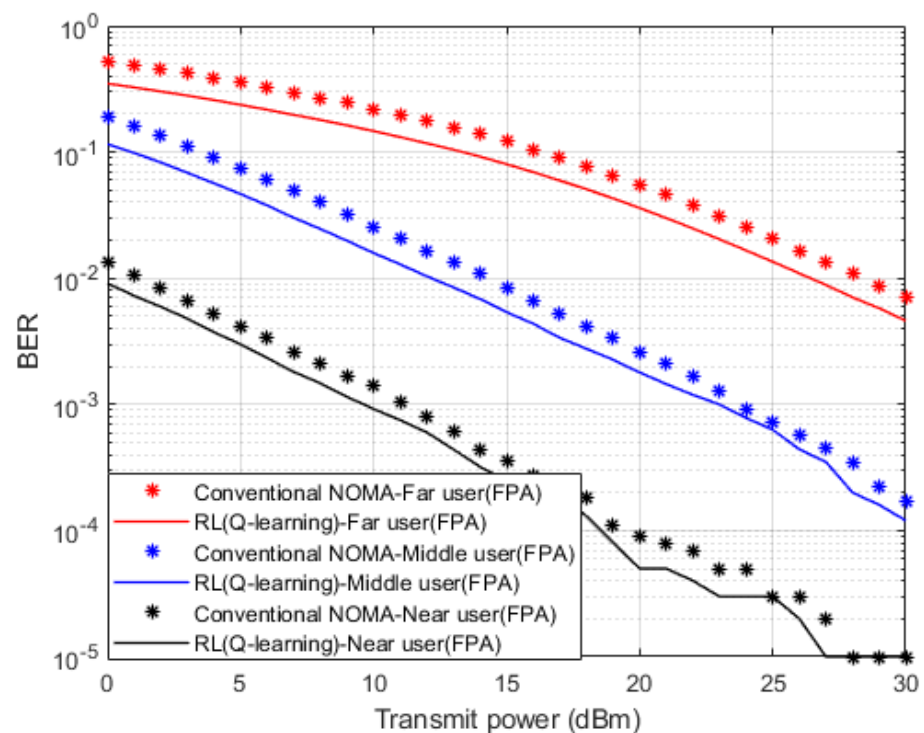

Figure 4. BER vs. power (*Q*-learning & Conventional NOMA (MMSE)).

In terms of the outage probability against applied power, Figure 5 illustrates the results for the inspected user devices in the MISO-NOMA cell when the developed Q-learning and standard MMSE are considered as a channel estimator schemes. Far, and middle devices simulation outcomes indicate about 2 dBm enhancement in saving power to realize 10^{-2} outage probability when the developed Q-learning algorithm scenario is applied compared to the MMSE procedure. Similarly, a near user with the developed Q-learning algorithm displays a 1 dBm improvement in power saving with respect to the MMSE scheme. This enhancement in power saving verifies the advantage of the developed Q-model as a channel estimator compared to the MMSE technique.
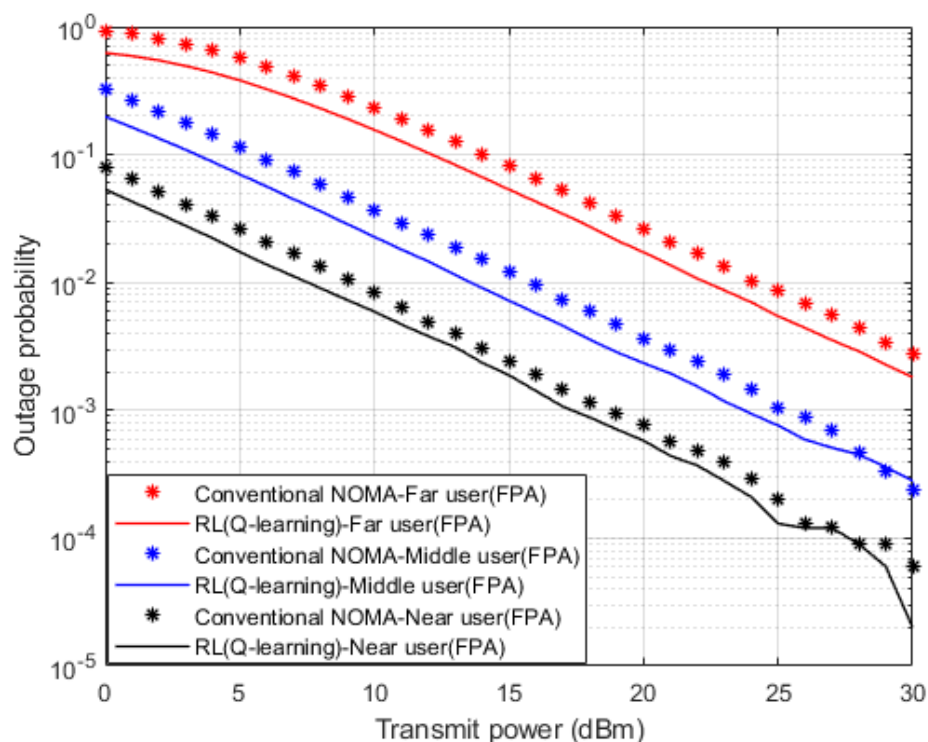
Figure 5. Outage Prob. vs. power (*Q*-learning & Conventional NOMA (MMSE)).

In Figure 6, we implement three baselines for comparisons: (1) standard minimum mean square error (MMSE) based channel prediction scheme [45]; (2) DL algorithm based on LSTM network for channel prediction applied in [17]; and RL based actor-critic procedure for channel prediction applied in [15]. This figure shows simulation results for the sum rate for all the user devices in the MISO-NOMA network versus applied power. Based on the simulation outcomes, it is evidently shown that the developed RL based Q-learning algorithm reveals superiority over standard MMSE procedure by 12 b/s/Hz approximately. Furthermore, the developed Q-learning scheme performs an enhancement over the DL based LSTM procedure presented in [17] by 2 b/s/Hz. For the third benchmark in [15], we generate the simulation environment according to the following: the actor and critic networks are both composed of two hidden layers with 400 and 300 nodes, respectively. The learning rate for actor and critic networks are 10^{-4} and 10^{-3} respectively. The discount factor γ is set to be 0.9 and has a buffer size of 10^5 [15]. Our developed RL based Q-learning procedure, shows superiority over the RL based actor-critic procedure at low power levels while starting from 23 dBm the actor-critic procedure starts showing some enhancement in terms of sum rates compared to the Q-learning process. These findings can validate that the developed Q-learning algorithm can be a competitive scheme compared to other algorithms that mainly depend on hidden layers to predict channel parameters.

Simulation outcomes for the sum rate against different number of users in the applied MISO-NOMA cell are illustrated in Figure 7, where the reference power is chosen to be 1 dBm. In addition to our proposed Q-learning algorithm, three distinct channel prediction methods are investigated as a benchmark comparison: (1) standard minimum mean square error (MMSE) based channel prediction scheme [45]; (2) DL algorithm based on LSTM network for channel prediction applied in [17]; and RL based actor-critic technique for channel estimation applied in [15]. As revealed from the results, our developed RL based Q-learning algorithm can achieve a substantial greater sum rate with respect to standard MMSE procedure, by at least 2 b/s/Hz. It can be observed that as the number of user devices in the cell is increasing, the suggested RL based Q-learning algorithm still shows dominance in accomplishing higher rates with respect to MMSE and DL based LSTM channel estimation methods. Similar to Figure 6, the RL actor-critic procedure applied

in [15] is created in our MISO-NOMA environment with the following parameters: the actor and critic networks are both composed of two hidden layers with 400 and 300 nodes, respectively. The learning rate for actor and critic networks are 10^{-4} and 10^{-3} respectively. The discount factor γ is set to be 0.9 and has a buffer size of 10^5 [15]. As shown in the results, the developed Q-learning scheme is showing an advantage over the actor-critic scheme with up to 6 users in the cell. Then, the hidden layers feature in the actor-critic procedure starts producing some sort of improvement in the sum rates compared to the Q-learning algorithm while the number of user terminals in the cell is increasing. Overall, these outcomes reveal that dependability can be assured by the suggested Q-learning algorithm even when the user devices in the cell are increased. In addition, it is worth saying that while increasing the user devices in the system, the interference will also grow up, thus the sum rate could be degraded.

Figure 8 illustrates simulation outcomes for the achievable capacity for every device in the examined MISO-NOMA system when both the developed Q-learning algorithm and MMSE channel estimation procedures are implemented. The attained rate for near devices reveals substantial improvement by 10 b/s/Hz over far and middle users' rates. The superiority of the near user in terms of the achievable rate is anticipated, due to the stable channel situation for the near user compared to other devices in the system. Additionally, the suggested Q-learning algorithm still can deliver few visible improvements compared to the MMSE technique for far and middle users' environments, this slight improvement is associated with the interference and inadequate link conditions for far and middle devices.

In addition to the three baselines comparisons implemented in Figures 6 and 7, we also create and implement RL based State-Action-Reward-State-Action (SARSA) algorithm [48–50] in Figures 9–11 for the purpose of more investigations and benchmark comparisons. The features and parameters of the SARSA algorithm are adapted in order that the SARSA procedure can be used as a channel estimator and compare the results of SARSA algorithm with the results obtained based on our developed Q-learning algorithm.
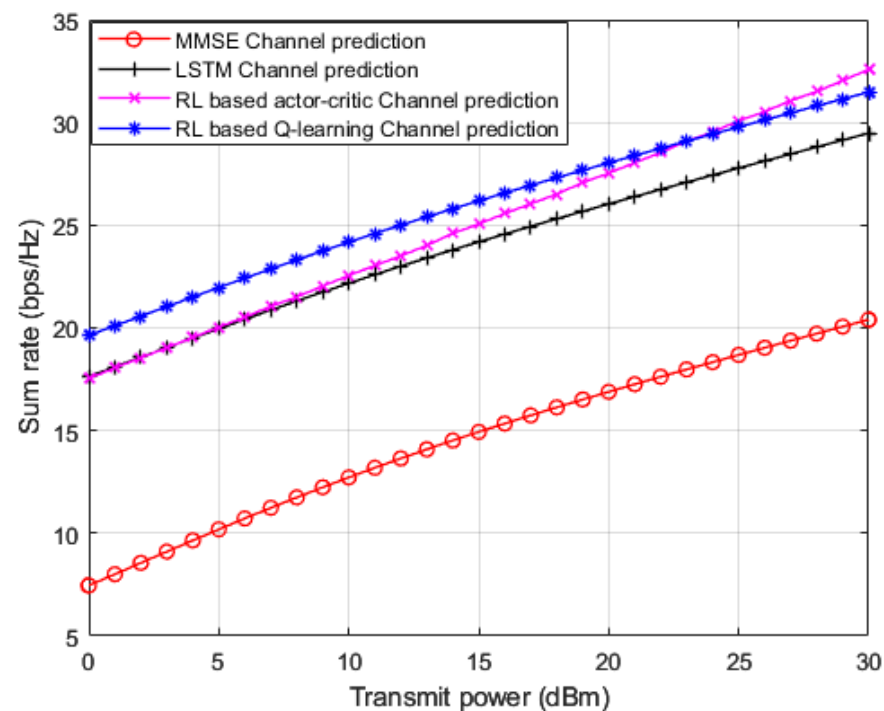

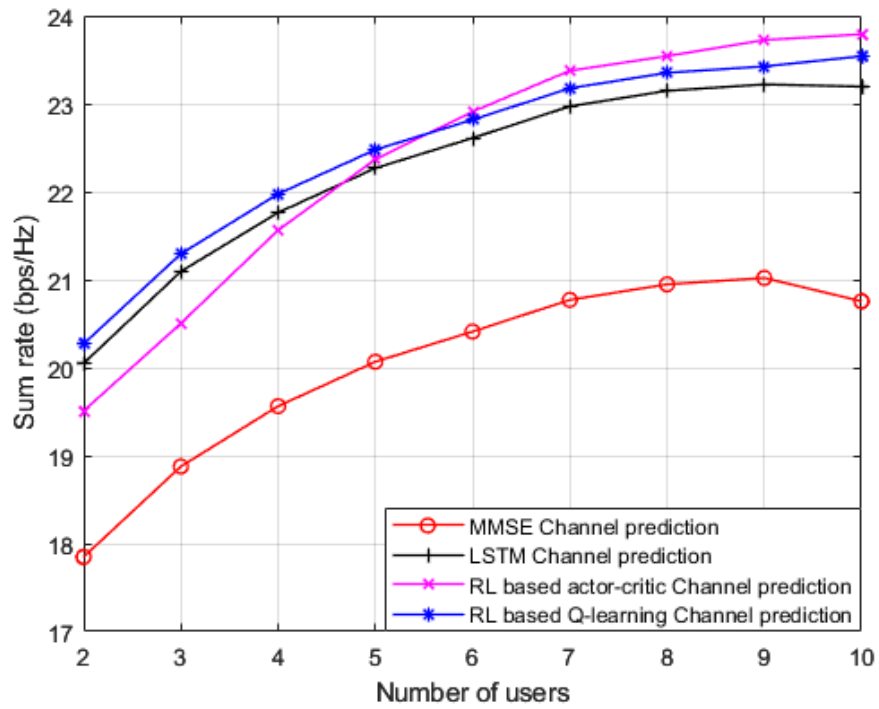**Figure 6.** Sum rate vs. power (MMSE, LSTM, RL actor-critic, RL *Q*-learning).

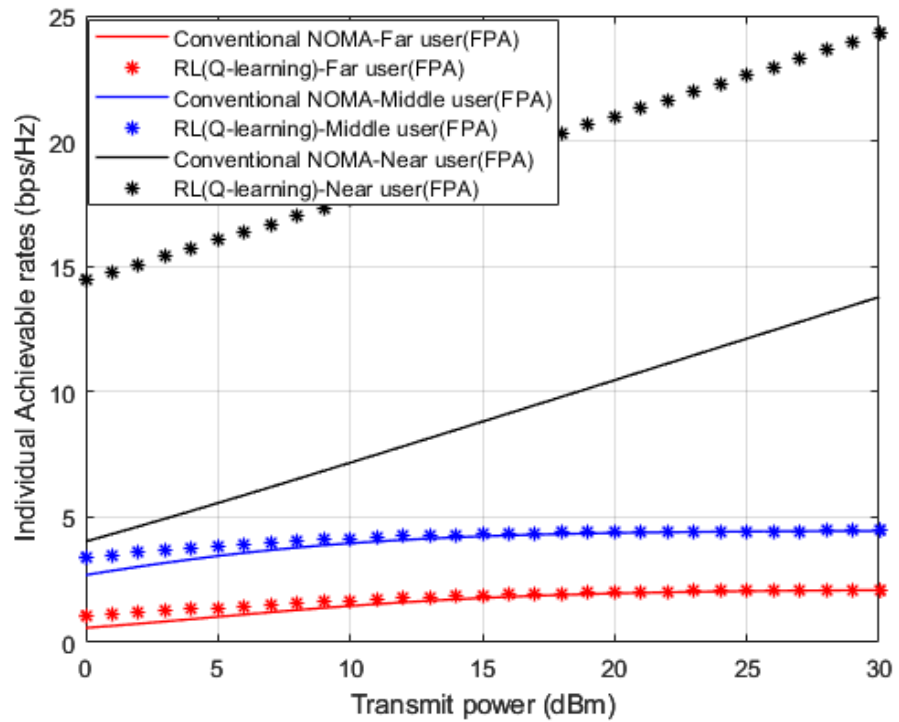**Figure 7.** Sum rate vs. number of users (MMSE, LSTM, RL actor-critic, RL *Q*-learning).

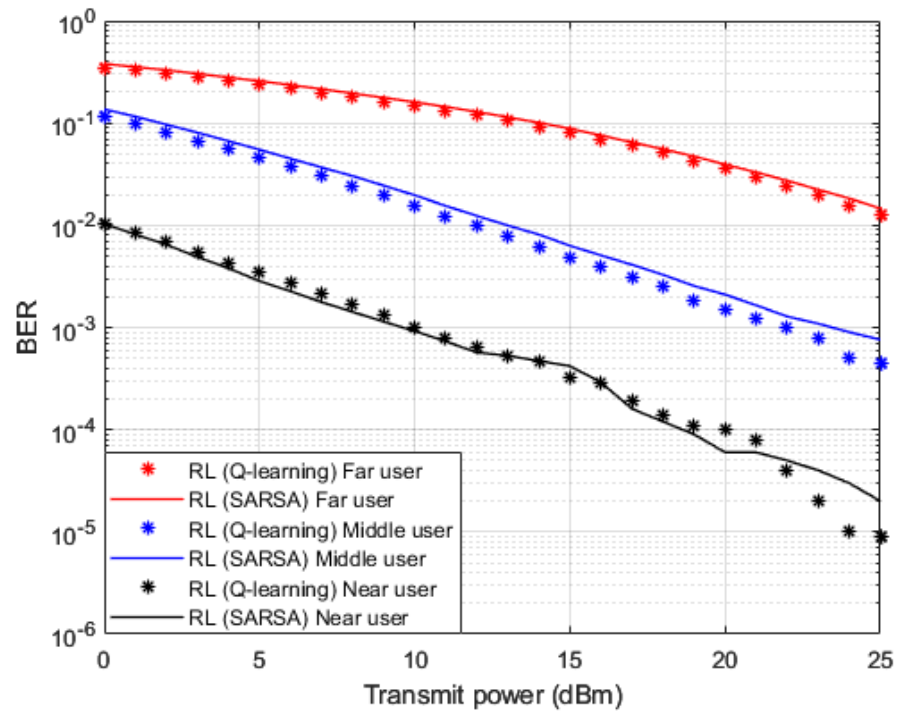**Figure 8.** Individual rate vs. power (*Q*-learning, Conventional NOMA (MMSE)).

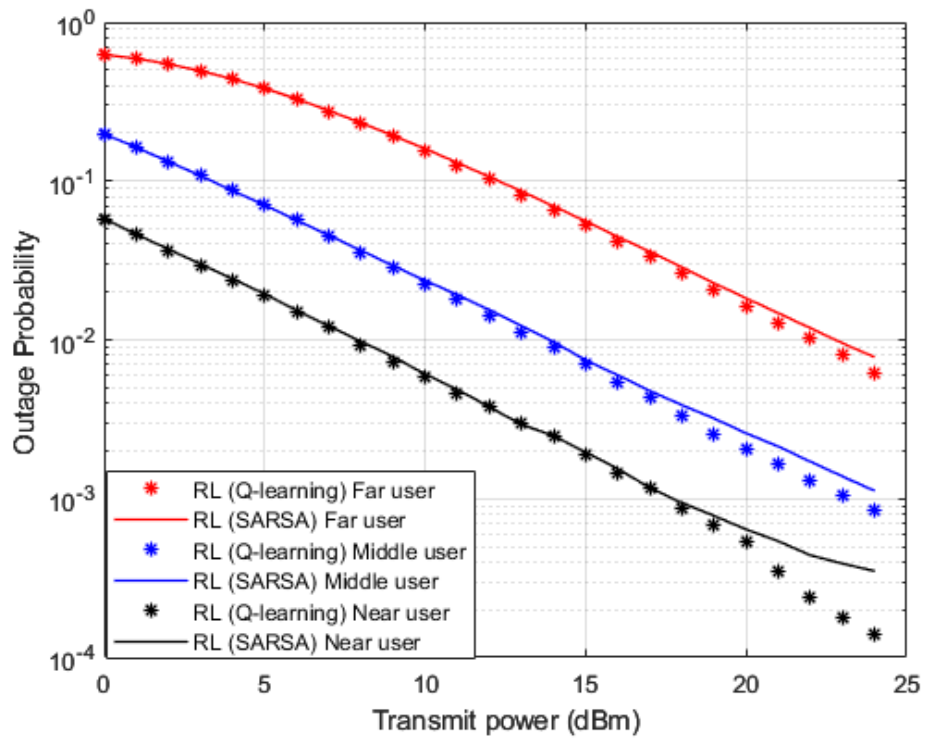**Figure 9.** BER vs. power (*Q*-learning, SARSA).

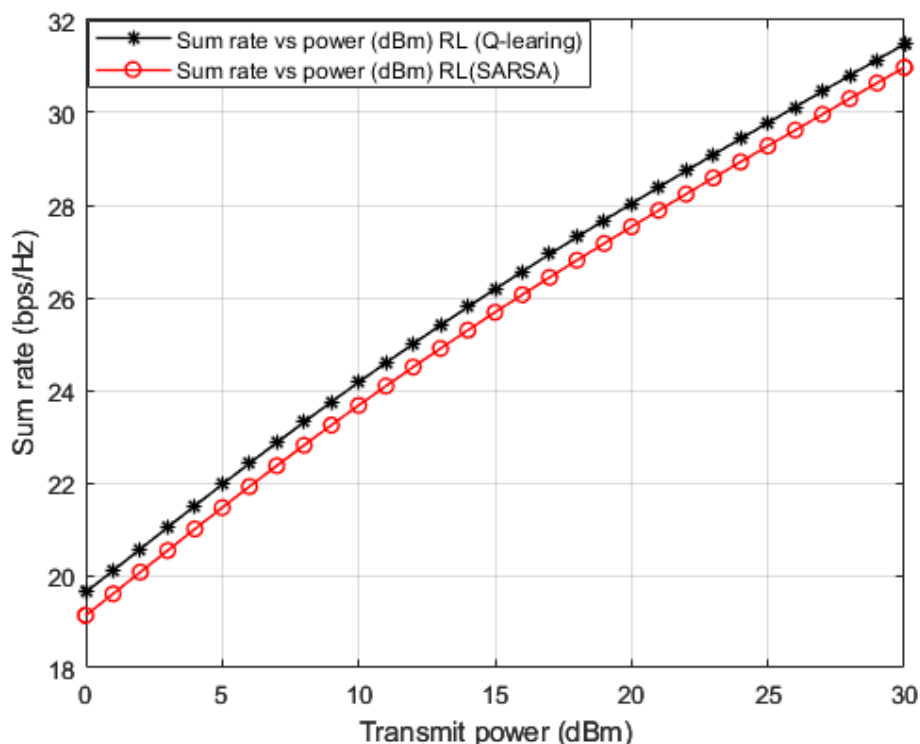**Figure 10.** Outage Prob. vs. power (*Q*-learning, SARSA).

Figure 11. Sum rate vs. power (*Q*-learning, SARSA).

The Q-learning algorithm and SARSA algorithm are two efficient RL algorithms, they are both table-based procedures with a Q-table to record equivalent Q-values of each state-action pair. However, when the size of state space increases, it will need a considerable amount of memory. Similar to the Q-learning algorithm, the SARSA algorithm also has exploration and exploitation processes, and it also needs a Q-table to record $Q(s_t, a_t)$ value corresponding to state s_t and action a_t. Differently, the running steps of the SARSA algorithm are as follows. First, according to the action selection scheme, the gent at the current state s_t, will select the action a_t. Then, the agent gets an immediate reward R based on the corresponding $Q(s_t, a_t)$ value. Finally, s_t will transfer to s_{t+1} and the agent will choose the next action a_{t+1}. Hence, the SARSA algorithm is a bit different from the Q-learning procedure, where the Q-value in the SARSA method is updated based on the action a_t implemented by the agent at the state s_t. While in the Q-learning algorithm, the action with the greatest Q-value in the next state s_{t+1} is employed to update Q-table.

In Figures 9 and 10, where BER and outage probability metrics are simulated against transmitted power, both our developed Q-learning and SARSA algorithms show comparable performance. However, at high power levels, the suggested Q-learning algorithm shows little improvement compared to the SARSA algorithm, which may be justified that the Q agent deciding the greedy action, which is the action that provides the maximum Q-value for the state. More investigations for the comparison between SARSA and the developed Q-learning algorithms are shown in Figure 11. Sum rates versus applied power are simulated in Figure 11, and it is noticed that the suggested Q-learning scheme provides an advantage over the SARSA algorithm, and a power saving is recorded by 1–2 dB approximately.

The proposed Q-learning method and traditional MMSE technique will be further examined when the Rician channel is applied for the path between BS and each user device. Rician channel is a stochastic model for wireless transmission where the signal reaches the receiver device via various scattered paths. Figure 12, illustrate simulation outcomes for BER against power transmitted when the Rician fading channel is applied. In the Rician simulation environment, we assign parameter $K = 10$, where K is described as the fraction of the signal power of the line-of-sight path to the signal power of the remaining scattered components. In addition, maximum doppler shift = 100 and sample rate = 9600 Hz are

used. Results for the Rician channel indicate that the Q-learning algorithm still can provide some sort of enhancement in decreasing the BER compared to the MMSE procedure. This slight improvement can be explained by the existence of a line of site component among BS and user terminal which can enhance the work of the MMSE procedure.
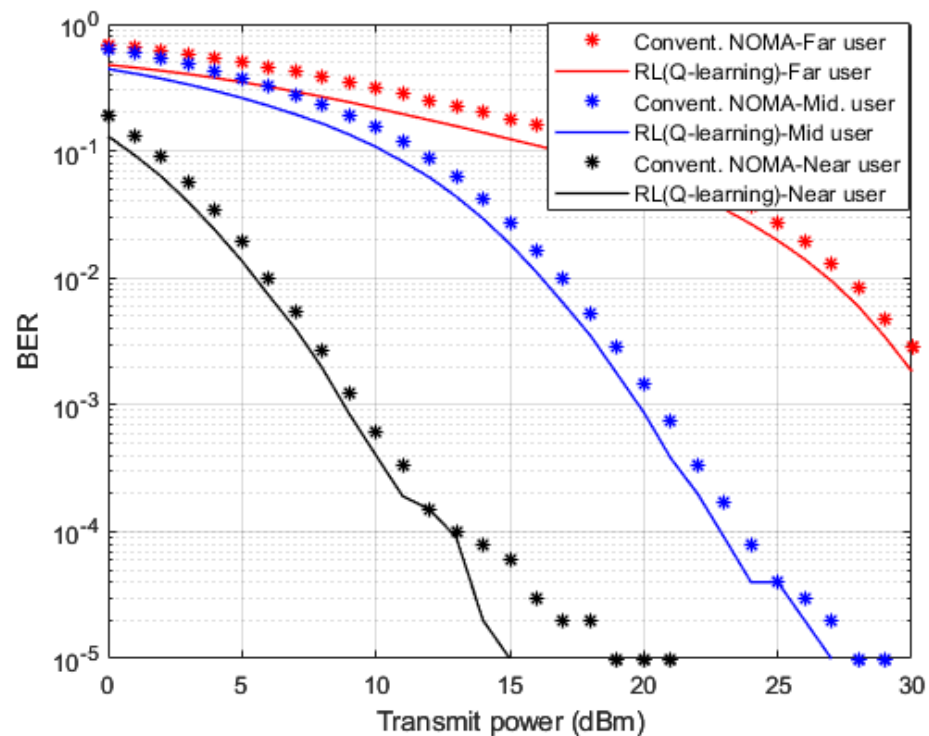

Figure 12. BER vs. Power (*Q*-learning, Conventional NOMA (MMSE)—Rician channel).

In Figure 13, two separate simulation setups are accomplished here to produce these results. In the first setup, the Fixed Power Allocation (FPA) structure is assigned for every user terminal in the MISO-NOMA cell. The second setup depends on the Optimized Power Structure (OPS) applied in accordance with the analytical power scheme that previously concluded in Section 4. FPA or OPS will be applied in conjunction with the suggested Q-learning algorithm as a channel estimator. Simulation outcomes in terms of BER indicate that far and middle users show the dominance of the OPS over the FPA. It can be noted that at specific BER values such as 10^{-2}, the achieved power saving by OPS policy is about 5 dBm for the far user, and 1–2 dBm approximately for the middle user. For near user results, the developed Q-learning algorithm jointly with the FPA scheme provide evident improvement in terms of BER over OPS, this could be clarified that for near device scenario, the stable channel condition provides more advantageous for the performance than the assigned power.

Outage probability results versus power are shown in Figure 14, where OPS and FPA schemes are also implemented. Both arrangements of OPS and FPA are implemented in conjunction with the proposed Q-learning algorithm as a channel estimator in the MISO-NOMA cell. Both far user and middle user results reveal an improvement in outage probability where a power reduction can be observed within 1–2 dBm when OPS is applied compared to the FPA scheme. On the other hand, near user with a Q-learning algorithm and FPA scenario shows a considerable outage improvement compared to the OPS case. A power reduction within 5 dBm is achieved when the FPA scheme is applied. These findings verify the results obtained for BER in Figure 13, which indicate that the FPA scheme is more adequate for user devices with high channel gains.
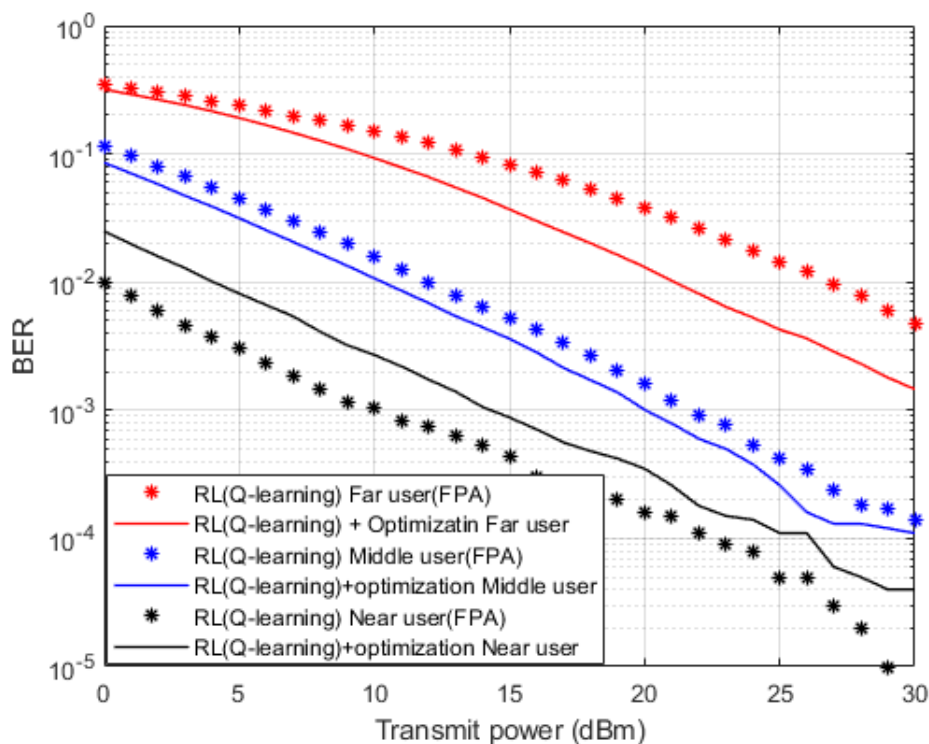
Figure 13. BER vs. Power (*Q*-learning, Optimization, FPA).
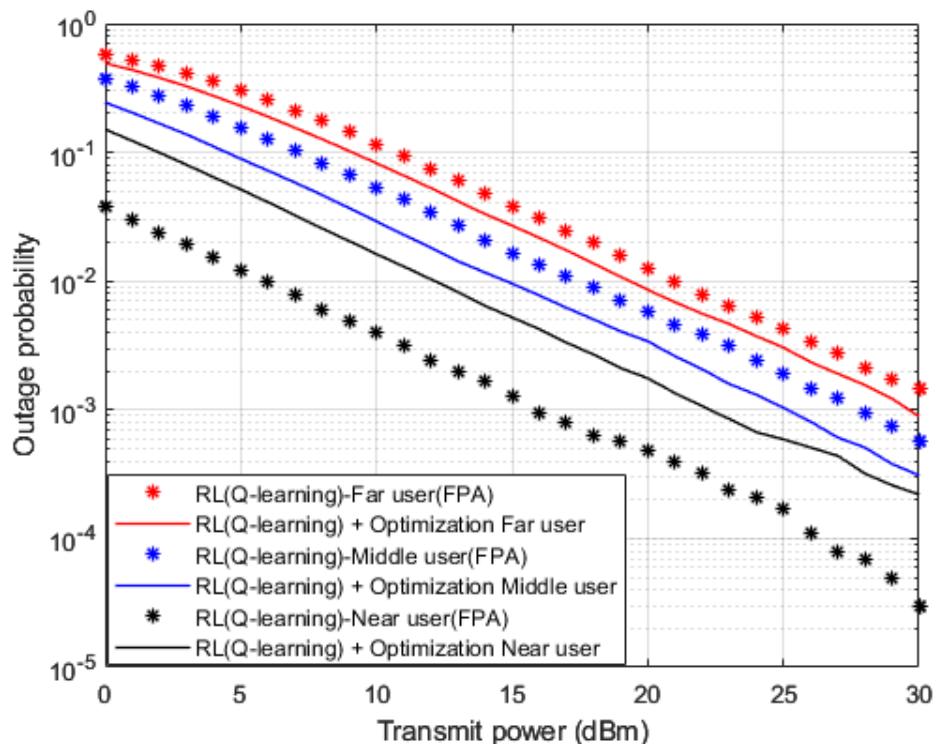

Figure 14. Outage Prob. vs. Power (*Q*-learning, Optimization, FPA).

In Figure 15, attainable rates for user devices are simulated against power transmitted when OPS and FPA schemes are applied in conjunction with the proposed Q-learning algorithm that is applied as a channel estimator. Results for far and middle devices point out that OPS provides 1 b/s/Hz improvement compared to the FPA scheme. This limited improvement might be clarified where the management of the power allocation for devices is not necessarily sufficient enough to alleviate the influence of interference particularly for

far and middle devices that mainly experience unstable links environments. As expected, results for near user device reveal superiority in achieved rate with respect to middle and far devices with at least 10 b/s/Hz. Furthermore, the results for the near user with FPA indicate a noticeable improvement compared to OPS, which validates the results obtained in Figures 13 and 14.
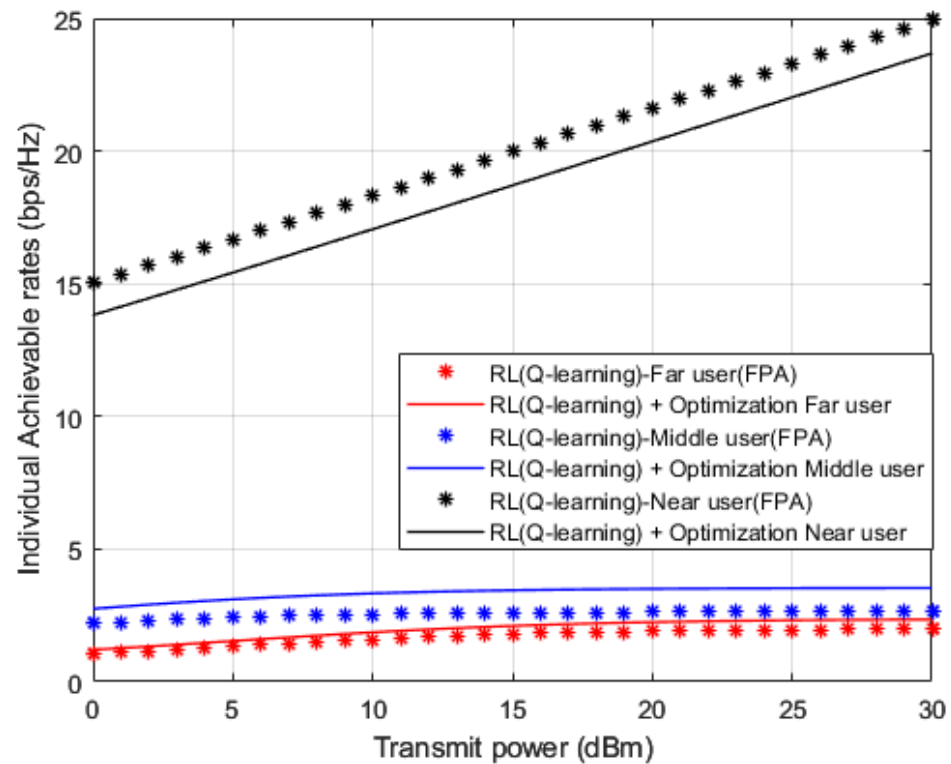

Figure 15. Individual rate vs. Power (*Q*-learning, Optimization, FPA).

In the end, we can further provide the analysis of the computational complexity as follows: The complexity of the reinforcement learning algorithm mainly depends on the size of the state space and the size of the action space [51]. According to [51], we can approximate the computational complexity of the Q-learning algorithm as $\mathcal{O}(SAH)$ per iteration, where S is the number of states, A is the number of actions, and H is the number of steps per episode. According to the state space and action space defined in our simulation environment, the amount of work per iteration can be approximated as $\mathcal{O}(NMK)$ [51,52], where N represents a number of antennas in BS, M represent a number of user devices in the cell, and K represents the size of channel coefficients. On the other hand, the computational complexity for the benchmark scheme implemented in [15], is described as follows: the sizes of the input layer, the first hidden layer, the second hidden layer, and the output layer for each network implemented in [15] is denoted as I, h_1, h_2, and U respectively. Thus, the total number of parameters in each network can be denoted as $\theta = I + h_1 + h_2 + U$, therefore, the complexity of this scheme regarding the channel estimation task can be approximated as $\mathcal{O}(MN_A(I + h_1 + h_2 + U)$ [15], where M represents the number of user terminals and N_A represent the number of antennas at BS. According to [53], the corresponding computational complexity for the traditional channel estimation method based MMSE can achieve a relatively low complexity, $\mathcal{O}(M^{2.37})$ [45,53], but, at the cost of performance degradation. Based on the aforementioned analysis, it can be shown that the complexity of the developed RL based Q-learning algorithm is competitive compared to other procedures.

10. Conclusions

In this study, the influence of adopting a developed RL based Q-learning algorithm to distinctly predict the channel parameters for every user device in the MISO-NOMA system is analyzed. In the developed Q-learning algorithm, the Q-model is created on the basis of the initialized channel statistics then updated based on the interaction between the Q-agent and the environment to maximize the received downlink sum rate and minimize the estimation loss. The efficacy of the developed Q-learning procedure is investigated by inspecting the performance of the proposed algorithm against different benchmark channel estimation schemes. The first benchmark scheme is based on standard MMSE procedure, the second approach is applying DL based LSTM network, the third scheme is implementing RL based actor-critic algorithm, and the fourth benchmark scheme is using RL based SARSA algorithm. In addition, the reliability of the proposed Q-learning procedure is validated by analyzing the behavior of the developed Q-learning algorithm in different fading channels. Furthermore, we provided a scenario that explores how the proposed channel prediction method based on Q-learning algorithm and the derived power allocation structure are both cooperatively employed for multiuser recognition in the MISO-NOMA network. Simulation results emphasized that dependability can be ensured by the developed Q-model even when the number of users in the cell is increased. Furthermore, the simulation outcomes in terms of BER, Outage probability, and individual user rate have demonstrated that the developed Q-learning algorithm for channel estimation jointly with an optimized power scheme can both realize consistent performance.

Author Contributions: Conceptualization, M.G.; Methodology, M.A.; Validation, A.A.; Formal analysis, M.G. and M.A.; Supervision, M.A. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

AWGN	Additive White Gaussian Noise
BER	bit error rate
BS	Base Station
CSI	Channel state information
DL	Deep Learning
DNN	Deep Neural Network
FPA	Fixed Power Allocation
OPS	Optimized Power structure
KKT	Karush-Kuhn–Tucker
LSTM	Long Short-Term Memory
LTE	Long Term Evolution
ML	Machine Learning
MSE	Mean Square Error
MMSE	Minimum Mean Square Error
MUD	Multiuser detection
PD-NOMA	Power Domain Non-Orthogonal Multiple Access
QoS	Quality of Service
SIC	Successive interference cancellation

MISO	Multi-input single-output
SARSA	State-Action-Reward-State-Action
RL	Reinforcement Learning

References

1. Dai, L.; Wang, B.; Ding, Z.; Wang, Z.; Chen, S.; Hanzo, L. A survey of non-orthogonal multiple access for 5G. *IEEE Commun. Surveys Tuts.* **2018**, *20*, 2294–2323.
2. Ding, Z.; Liu, Y.; Choi, J.; Sun, Q.; Elkashlan, M.; Chih-Lin, I.; Poor, H.V. Application of non-orthogonal multiple access in LTE and 5G networks. *IEEE Commun. Mag.* **2017**, *55*, 185–191. [CrossRef]
3. Gaballa, M.; Abbod, M.; Jameel, A. Power Optimization Analysis using Throughput Maximization in MISO Non-Orthogonal Multiple Access System. In Proceedings of the 2021 IEEE Globecom Workshops, Madrid, Spain, 7–11 December 2021.
4. Gaballa, M.; Abbod, M.; Albasman, M. Power Allocation & MRC Analysis for Single Input Multi Output Non-Orthogonal Multiple Access System. In Proceedings of the 2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics), Melbourne, Australia, 6–8 December 2021.
5. Wang, X.; Zhu, P.; Li, D.; Xu, Y.; You, X. Pilot-Assisted SIMO-NOMA Signal Detection with Learnable Successive Interference Cancellation. *IEEE Commun. Lett.* **2021**, *25*, 2385–2389. [CrossRef]
6. AbdelMoniem, M.; Gasser, S.M.; El-Mahallawy, M.S.; Fakhr, M.W.; Soliman, A. Enhanced NOMA system using adaptive coding and modulation based on LSTM neural network channel estimation. *Appl. Sci.* **2019**, *9*, 3022. [CrossRef]
7. Fu, Y.; Salaün, L.; Sung, C.W.; Chen, C.S. Distributed Power Allocation for the Downlink of a Two-Cell MISO-NOMA System. In Proceedings of the IEEE 87th Vehicular Technology Conference, Porto, Portugal, 3–6 June 2018; pp. 1–6.
8. Emir, A.; Kara, F.; Kaya, H.; Li, X. Deep learning-based flexible joint channel estimation and signal detection of multi-user OFDM-NOMA. *Phys. Commun.* **2021**, *48*, 101443. [CrossRef]
9. Jeon, Y.-S.; Li, J.; Tavangaran, N.; Poor, H.V. Data-aided channel estimator for MIMO systems via reinforcement learning. In Proceedings of the ICC 2020-2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020.
10. Kundu, N.K.; McKay, M.R. Channel Estimation for Reconfigurable Intelligent Surface Aided MISO Communications: From LMMSE to Deep Learning Solutions. *IEEE Open J. Commun. Soc.* **2021**, *2*, 471–487. [CrossRef]
11. Mthethwa, B.; Xu, H. Deep Learning-Based Wireless Channel Estimation for MIMO Uncoded Space-Time Labeling Diversity. *IEEE Access* **2020**, *8*, 224608–224620. [CrossRef]
12. Li, L.; Zhang, Z.; Yang, L. Influence of Autoencoder-Based Data Augmentation on Deep Learning-Based Wireless Communication. *IEEE Wirel. Commun. Lett.* **2021**, *10*, 2090–2093. [CrossRef]
13. Kim, D.; Jung, H.; Lee, I.H. Deep Learning-Based Power Control Scheme with Partial Channel Information in Overlay Device-to-Device Communication Systems. *IEEE Access* **2021**, *9*, 122125–122137. [CrossRef]
14. Zhang, T.; Mao, S. Energy-Efficient Power Control in Wireless Networks with Spatial Deep Neural Networks. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 111–124. [CrossRef]
15. Chu, M.; Liu, A.; Jiang, C.; Lau, V.K.N.; Yang, T. Wireless Channel Prediction for Multi-user Physical Layer with Deep Reinforcement Learning. In Proceedings of the 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), Helsinki, Finland, 19–22 June 2022.
16. Tan, J.; Liang, Y.C.; Zhang, L.; Feng, G. Deep Reinforcement Learning for Joint Channel Selection and Power Control in D2D Networks. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 1363–1378. [CrossRef]
17. Gaballa, M.; Abbod, M.; Aldallal, A. Investigating the Combination of Deep Learning for Channel Estimation and Power Optimization in a Non-Orthogonal Multiple Access System. *Sensors* **2022**, *22*, 3666. [CrossRef] [PubMed]
18. Ma, W.; Qi, C.; Zhang, Z.; Cheng, J. Sparse Channel Estimation and Hybrid Precoding Using Deep Learning for Millimeter Wave Massive MIMO. *IEEE Trans. Commun.* **2020**, *68*, 2838–2849. [CrossRef]
19. Ghous, M.; Hassan, A.K.; Abbas, Z.H.; Abbas, G.; Hussien, A.; Baker, T. Cooperative Power-Domain NOMA Systems: An Overview. *Sensors* **2022**, *22*, 9652. [CrossRef]
20. Dai, L.; Wang, B.; Yuan, Y.; Han, S.; Chih-Lin, I.; Wang, Z. Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends. *IEEE Commun. Mag.* **2015**, *53*, 74–81. [CrossRef]
21. Tang, Z.; Wang, J.; Wang, J.; Song, J. On the achievable rate region of NOMA under outage probability constraints. *IEEE Commun. Lett.* **2019**, *23*, 370–373. [CrossRef]
22. Yang, Z.; Xu, W.; Pan, C.; Pan, Y.; Chen, M. On the Optimality of Power Allocation for NOMA Downlinks With Individual QoS Constraints. *IEEE Commun. Lett.* **2017**, *21*, 1649–1652. [CrossRef]
23. Ding, Z.; Yang, Z.; Fan, P.; Poor, H.V. On the Performance of Non-Orthogonal Multiple Access in 5G Systems with Randomly Deployed Users. *IEEE Signal Process. Lett.* **2014**, *21*, 1501–1505. [CrossRef]
24. Gaballa, M.; Abbod, M.; Jameel, A.; Khaled, N. Throughput Maximization & Power Optimization Analysis in Non-Orthogonal Multiple Access System. In Proceedings of the 2021 IEEE 4th 5G World Forum, Montreal, QC, Canada, 13–15 October 2021.

25. Li, S.; Derakhshani, M.; Lambotharan, S. Outage-constrained robust power allocation for downlink MC-NOMA with imperfect SIC. In Proceedings of the IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–7.

26. Zhu, J.; Wang, J.; Huang, Y.; He, S.; You, X.; Yang, L. On Optimal Power Allocation for Downlink Non-Orthogonal Multiple Access Systems. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2744–2757. [CrossRef]

27. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.

28. Ben-Tal, A.; Nemirovski, A. *Lecture on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*; MPS-SIAM: Philadelphia, PA, USA, 2018.

29. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.

30. Rothmann, M.; Porrmann, M. A Survey of Domain-Specific Architectures for Reinforcement Learning. *IEEE Access* **2022**, *10*, 13753–13767. [CrossRef]

31. Naeem, M.; De Pietro, G.; Coronato, A. Application of Reinforcement Learning and Deep Learning in Multiple-Input and Multiple-Output (MIMO) Systems. *Sensors* **2022**, *22*, 309. [CrossRef] [PubMed]

32. Amiri, R.; Almasi, M.A.; Andrews, J.G.; Mehrpouyan, H. Reinforcement learning for self-organization and power control of two-tier heterogeneous networks. *IEEE Trans. Wireless Commun.* **2019**, *18*, 3933–3947. [CrossRef]

33. Jeon, Y.S.; Lee, N.; Poor, H.V. Robust Data Detection for MIMO Systems with One-Bit ADCs: A Reinforcement Learning Approach. *IEEE Trans. Wirel. Commun.* **2019**, *19*, 1663–1676. [CrossRef]

34. Nie, J.; Haykin, S. A Q-learning-based dynamic channel assignment technique for mobile communication systems. *IEEE Trans. Veh. Technol.* **1999**, *48*, 1676–1687.

35. Ghavamzadeh, M.; Kappen, H.; Azar, M.; Munos, R. Speedy Q-Learning. In *Advances in Neural Information Processing Systems*; Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K.Q., Eds.; Curran Associates: Red Hook, NY, USA, 2011; Volume 24.

36. Zhai, Q.; Bolić, M.; Li, Y.; Cheng, W.; Liu, C. A Q-Learning-Based Resource Allocation for Downlink Non-Orthogonal Multiple Access Systems Considering QoS. *IEEE Access* **2021**, *9*, 72702–72711. [CrossRef]

37. Dong, M.; Tong, L.; Sadler, B.M. Optimal Insertion of Pilot Symbols for Transmissions over Time-Varying Flat Fading Channels. *IEEE Trans. Signal Process.* **2004**, *52*, 1403–1418. [CrossRef]

38. da Silva, M.V.; Montejo-Sánchez, S.; Souza, R.D.; Alves, H.; Abrão, T. D2D Assisted Q-Learning Random Access for NOMA-Based MTC Networks. *IEEE Access* **2022**, *10*, 30694–30706. [CrossRef]

39. da Silva, J.V.; Souza, R.D.; Alves, H.; Abrao, T. A NOMA-based Q-learning random access method for machine type communications. *IEEE Wireless Commun. Lett.* **2020**, *9*, 1720–1724. [CrossRef]

40. Mete, E.; Girici, T. Q-learning based scheduling with successive interference cancellation. *IEEE Access* **2020**, *8*, 172034–172042. [CrossRef]

41. Zhou, Y.; Zhou, F.; Wu, Y.; Hu, R.Q.; Wang, Y. Subcarrier Assignment Schemes Based on Q-Learning in Wideband Cognitive Radio Networks. *IEEE Trans. Veh. Technol.* **2019**, *69*, 1168–1172. [CrossRef]

42. Xi, B.; Lei, D. Q-Learning-Based Teaching-Learning Optimization for Distributed Two-Stage Hybrid Flow Shop Scheduling with Fuzzy Processing Time. *Complex Syst. Model. Simul.* **2022**, *2*, 113–129. [CrossRef]

43. Gaballa, M.; Abbod, M.; Aldallal, A. Deep Learning and Power Allocation Analysis in NOMA System. In Proceedings of the 2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN), Barcelona, Spain, 5–8 July 2022; pp. 196–201.

44. Youssef, M.-J.; Nour, C.A.; Lagrange, X.; Douillard, C. A Deep Q-Learning Bisection Approach for Power Allocation in Downlink NOMA Systems. *IEEE Commun. Lett.* **2022**, *26*, 316–320. [CrossRef]

45. Neumann, D.; Wiese, T.; Utschick, W. Learning the MMSE Channel Estimator. *IEEE Trans. Signal Process.* **2018**, *66*, 2905–2917. [CrossRef]

46. Cirik, A.C.; Balasubramanya, N.M.; Lampe, L.; Vos, G.; Bennett, S. Toward the Standardization of Grant-Free Operation and the Associated NOMA Strategies in 3GPP. *IEEE Commun. Stand. Mag.* **2019**, *3*, 60–66. [CrossRef]

47. *Recommendation ITU-R M.1225*; Guidelines for Evaluation of Radio Transmission Technologies for IMT-2000. International Telecommunication Union (ITU): Geneva, Switzerland, 1997. Available online: https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.1225-0-199702-I!!PDF-E.pdf (accessed on 15 November 2022).

48. Ahsan, W.; Yi, W.; Qin, Z.; Liu, Y.; Nallanathan, A. Resource Allocation in Uplink NOMA-IoT Networks: A Reinforcement-Learning Approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 5083–5098. [CrossRef]

49. Mu, X.; Zhao, X.; Liang, H. Power Allocation Based on Reinforcement Learning for MIMO System with Energy Harvesting. *IEEE Trans. Veh. Technol.* **2020**, *69*, 7622–7633. [CrossRef]

50. Jiang, H.; Gui, R.; Chen, Z.; Wu, L.; Dang, J.; Zhou, J. An Improved SARSA (λ) Reinforcement Learning Algorithm for Wire-less Communication Systems. *IEEE Access* **2019**, *7*, 115418–115427. [CrossRef]

51. Jin, C.; Allen-Zhu, Z.; Bubeck, S.; Jordan, M. Is Q-learning provably efficient? In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates: Red Hook, NY, USA, 2017; pp. 4863–4873.

52. Li, Z.; Pan, S.; Qin, Y. Multiuser Scheduling Algorithm for 5 G IoT Systems Based on Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2022**. [CrossRef]
53. Liao, Y.; Hua, Y.; Cai, Y. Deep learning based channel estimation algorithm for fast time-varying MIMO-OFDM systems. *IEEE Commun. Lett.* **2019**, *24*, 572–576. [CrossRef]