

## Article

# Optimal User Scheduling in Multi Antenna System Using Multi Agent Reinforcement Learning

Muddasar Naeem <sup>1</sup>, Antonio Coronato <sup>2</sup>, Zaib Ullah <sup>1,\*</sup>, Sajid Bashir <sup>3</sup> and Giovanni Paragliola <sup>1</sup>

<sup>1</sup> Institute of High Performance Computing and Networking, National Research Council of Italy, 80131 Naples, Italy

<sup>2</sup> Centro di Ricerche sulle Tecnologie ICT per la Salute ed il Benessere, Università Giustino Fortunato, 82100 Benevento, Italy

<sup>3</sup> Department of Electrical Engineering, National University of Sciences & Technology, Islamabad 44000, Pakistan

\* Correspondence: zaib.ullah@icar.cnr.it

**Abstract:** Multiple Input Multiple Output (MIMO) systems have been gaining significant attention from the research community due to their potential to improve data rates. However, a suitable scheduling mechanism is required to efficiently distribute available spectrum resources and enhance system capacity. This paper investigates the user selection problem in Multi-User MIMO (MU-MIMO) environment using the multi-agent Reinforcement learning (RL) methodology. Adopting multiple antennas' spatial degrees of freedom, devices can serve to transmit simultaneously in every time slot. We aim to develop an optimal scheduling policy by optimally selecting a group of users to be scheduled for transmission, given the channel condition and resource blocks at the beginning of each time slot. We first formulate the MU-MIMO scheduling problem as a single-state Markov Decision Process (MDP). We achieve the optimal policy by solving the formulated MDP problem using RL. We use aggregated sum-rate of the group of users selected for transmission, and a 20% higher sum-rate performance over the conventional methods is reported.

**Keywords:** reinforcement learning; user scheduling; channel capacity; MIMO; MU-MIMO; next-generation networks; fairness; sumrate



**Citation:** Naeem, M.; Coronato, A.; Ullah, Z.; Bashir, S.; Paragliola, G. Optimal User Scheduling in Multi Antenna System Using Multi Agent Reinforcement Learning. *Sensors* **2022**, *22*, 8278. <https://doi.org/10.3390/s22218278>

Academic Editors: Pietro Manzoni and Davy P. Gaillot

Received: 29 September 2022

Accepted: 24 October 2022

Published: 28 October 2022

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



**Copyright:** © 2022 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

Higher data rates and reliability are a few challenges faced by 5G and beyond technologies. MIMO systems are becoming one of the significant pillars in wireless communication due to their spectral efficiencies and diversity gains [1]. In addition, MU-MIMO systems further exploit both spatial multiplexing and spatial diversity for reliable communication links and higher data rate [2]. MU-MIMO systems performance is superior to that of single-user MIMO communication networks due to its ability to serve more than one user in a Transmission Time Slot (TTS) and frequency band [3].

Moreover, MU-MIMO systems provide a notable advantage over conventional communication systems (e.g., its multiplexing gain is proportional to the number of transmit antennas, even though the user device doesn't need to have many antennas). This minimizes the burden because of limitations of size and user equipment cost [3]. The other benefit is a fewer impact on propagation problems, i.e., antenna correlation and channel rank, since multi-user diversity addresses propagation issues. MU-MIMO is on the other hand, sensitive to Channel State Information (CSI) accuracy due to inter-user interference and may be mitigated as in [4,5].

Recent advances in Deep Neural Network (DNN) have made RL [6] the most important and attractive Artificial Intelligence (AI) technology. RL is a machine learning branch where an agent interacts with a given environment (mostly unknown), chooses actions, and gradually explores the environment's characteristics. RL and DNN have been

used in diverse research and real-life areas such as computer vision, the configuration of resources, self-organized systems, games, natural language processing, communication and networking, robotics, and autonomous control [7–11].

Machine Learning (ML) techniques have been used recently in different aspects of communication systems, i.e., Resource Allocation (RA). For example, deep learning has been applied in [12] for RA in massive MIMO systems. Similarly, applications of RL in RA management of MIMO systems have been found [13–16], but RL-based methods have not been used for scheduling multiple users in uplink MU-MIMO systems.

The proposed work aims at investigating the potential use of RL in MIMO communications with attention on MU-MIMO systems. The motivation behind this work stems from relevant aspects, such as (1) the importance of allocating radio resources using a scheduling mechanism at the base station in future wireless communication networks; (2) the limitations of existing state-of-the-art user scheduling techniques; and (3) the robustness of RL methods in alleviating these shortcomings and giving an efficient performance.

Moreover, the RL, being one of the most appealing technology, learns the dynamics of an environment MU-MIMO scheduling in this case) with its experience. One does not need to set values of different parameters as required in standard scheduling algorithms. Instead, the agent can learn the optimal combination of wireless channel parameters and optimally select the group of users out of all candidate users for transmission due to its decision-making ability.

The paper is organized as follows. Section 2 provides the state-of-the-art on user scheduling in MU-MIMO systems. Section 3 reports a brief overview of the RL methodology and an introduction to MIMO communication. In contrast, Section 4 provides a detailed introduction to the proposed approach with a problem statement, problem formulation, and RL scheduling algorithm. The results are presented in Section 5. Finally, the concluding remarks are given in Section 6.

## 2. Related Work

Optimal use of radio resources is essential to enhance the system capacity, and user schedule can play a crucial role [17]. We reviewed several user selection schemes adopted in the literature using various selection mechanisms. These methods can be broadly categorized into three types. The first group of tools adopts a certain system parameter for the selection of users (e.g., include SLNR and SINR, etc.) [18]. The second category of the scheduling algorithm considers the scenario where the CSI is not available at the Base Station (BS). The third technique addresses the issue of fairness; i.e., consider fairness as the only performance metric to ensure the quality of service [19].

From the game theory perspective, a user-centric access point scheduling for cell-free massive MIMO systems has been investigated in [20]. Authors have developed a user-centric access-point cluster model as a local altruistic game. Moreover, a maximum non-neighbor-set-based concurrent spatial adaptive play technique is to obtain the Nash equilibrium.

Similarly, a user-selection mechanism for MU-MIMO systems in uplink mode is presented in [21], where the authors used antennas and the ZF detector at the receiver in the BS. They consider the scenario of imperfect channel estimation with AWGN and Rician fading channels. The objective of the user selection is to maximize SNR.

Another user scheduling framework for a cooperative nonorthogonal multiple access scenario is developed in [22]. Deep learning technology has been employed to recognize and classify the channels of imperfect CSI. Deep learning was used to enhance the accuracy of CSI. While authors in [23] consider an end-to-end design of MU-MIMO systems in a downlink scenario, including precoding, limited feedback, and pilot sequences. Then, DL method is used to jointly optimize the precoder design at a base station BS and generate user feedback information. The neural network is used at BS to produce pilot sequences and assists the users in obtaining CSI.

A beam-user selection based on machine learning and low complexity hybrid beamforming infrastructure for the multiuser massive MIMO downlink system is presented in [24]. The householder reflectors are employed to produce the orthogonal analog beamforming matrix. The proposed scheme also uses a feedforward neural network and shows reasonable performance in terms of energy efficiency in the ill-conditioned massive MIMO environment, while the joint user selection and optimal transmit power and antenna selection have been discussed in [25]. The problem of joint user selection and optimal transmit power, and antenna selection is formulated to address inter-cell interference in multi-cell massive MIMO networks. A novel power consumption technique is also used to analyze precise power consumption.

The problems of max-product and max-min power allocation have been formulated in [26,27] by using SINR and SLNR mechanism for linear precoder design. DNN is deployed to predict the optimal power allocation based on each user's location and helps to minimize the system's processing time in identifying the optimal power allocation.

An SLNR-based user scheduling approach is presented in [28], where a user's leakage power to other users is considered the major parameter to decide whether the user should be chosen. Another similar approach is proposed in [29] that also addresses user selection. A modification to the leakage-based method regarding the selection of the first user was presented in [30]. Block diagonalization-based technique is proposed in [31], and the authors claim to achieve reasonable fairness and capacity among users. The authors of [2] developed a data detection receiver and joint maximum likelihood modulation classification of the co-scheduled users. In [32], the authors have considered fairness and sum-rate performance metrics by proposing a near-optimal scheduling algorithm.

A resource allocation mechanism is developed in [33] by using a POMDP method for downlink transmit beamforming at BS equipped with multi-antennas. The authors have used the myopic policy in designing the scheme to prevent the high computational complexity of the value iteration technique. A binary FPA is used in [34] for both antenna and user scheduling to obtain sum-rate performance with reduced computational complexity. An ML-based joint infrastructure for a hybrid precoder and user scheduling is proposed in [35] for MU-MIMO to improve the sum rate. The first component is based on cross-entropy, while the latter is based on the Correlation factor. Keysight's electronic system-level software is used to produce a channel matrix.

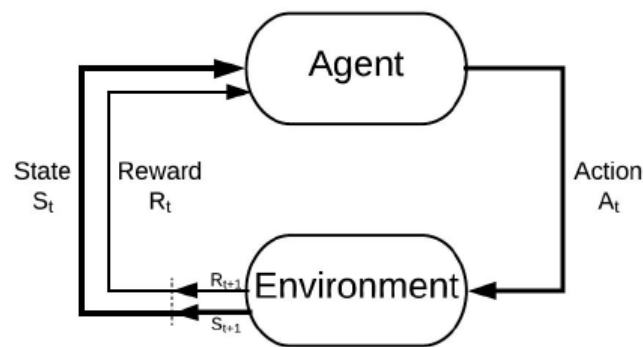
We have reviewed the diverse types of related works. Some works consider the standard parameters for user selection to enhance system capacity, while others apply deep learning and machine learning methods for resource or power allocation. The proposed work is the first concerning using Reinforcement Learning to solve the optimal scheduling in MU-MIMO systems.

### 3. Technical Background

In this section, we report an introduction to RL and to MIMO communication.

#### 3.1. Reinforcement Learning

The goal of an RL agent is to interact with a given environment and learn the dynamics of that environment as demonstrated in Figure 1. In the learning process, an RL agent can take action  $a_t$  out of available actions set in any state  $s_t$  at  $t$  time interval and then receive a corresponding reward. After much trial and error, the agent learns the best action for each environment state. The process of learning optimal action for every state forms an optimal policy. The first step is to formulate an underlying problem into an MDP problem and then use a suitable RL technique to solve the problem.



**Figure 1.** The reinforcement learning problem.

After selecting an appropriate RL method for a modeled MDP, the next step is keeping an exploration-exploitation balance. The RL agent has the option either to use known rewarding actions or can find new actions that may prove more beneficial. So, only one scheme may not be a good strategy, and an RL agent must be able to learn a trade-off between the two.

RL algorithms may be categorized as value-based, policy-based, or a combination of both (e.g., the actor-critic algorithm). This classification can be done between model-free and model-based techniques. The first group of algorithms concerns those techniques that do not require a precise model of the environments to be controlled, such as the Q-learning method as given in Algorithm 1. At the same time, the second class of methods exploits the model and provides an analytical solution to the MDP that describes the environment, such as dynamic programming. Dynamic programming has two variants, i.e., value iteration and policy iteration. The former is summarized in Algorithm 2.

Therefore, the first class of tools, thus, rely on trial and error to update their knowledge and experience about the given environment. The RL agent has to interact with the environment repeatedly to learn the environment. A few examples of this category are the temporal difference and Monte Carlo.

---

**Algorithm 1:** Q-learning algorithm.

---

```

1. Initialize
   Q arbitrarily
   Q (terminal) = 0
   Repeat
     initialize s
     Repeat
       choose  $a' \in \epsilon - greedily$ 
       take action a, observe  $r, s'$ 
        $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$ 
        $s \leftarrow s'$ 
     s is terminal
   until convergence
  
```

---

**Algorithm 2:** Value iteration algorithm.

---

```

Initialize  $V$  arbitrarily
Repeat
   $\Delta \leftarrow 0$ 
  For each  $s \in S$ 
     $v \leftarrow V(s)$ 
     $V(s) \leftarrow \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$ 
     $\Delta \leftarrow \max(\Delta, |v - V(s)|)$ 
until  $\Delta < \theta$  (a small positive number)
output a deterministic policy,  $\pi$ , such that
 $\pi(s) = \operatorname{argmax}_a \sum_{s',r} p(s',r|s,a)[r + \gamma V(s')]$ 

```

---

In our problem, first, each user selects a resource block and sends this information to BS so that BS can choose a group of users based on the received information. We modeled a resource block as a single state MDP, and Bayesian tools are suitable to solve such problems. The probability distribution for RV  $X$  is employed to make an inference about an RV  $X$  in Bayesian RL, and later extraction on distribution is done for inferences [36]. Such a process is performed by following these steps:

1. Take a prior distribution  $P(X)$ ;
2.  $P(X)$  is the belief about RV  $X$  with no data observation;
3. take a statistical model  $P(Y|X)$ ;
4.  $P(Y|X)$  is the statistical dependence and belief about RV  $Y$  given the  $X$ ;
5. Make observation on data  $Y = y$ ;
6. Find the posterior distribution using the Bayes rule as in [36].

$$P(X|Y = y) = \frac{P(y|X)P(X)}{\int P(y|X')P(X')dX'} \quad (1)$$

### 3.2. MIMO Communication

MIMO is different from single antenna systems as data transmission, and reception in MIMO is done on multiple antennas. Moreover, MIMO introduces signaling degrees of freedom, also known as the spatial degree of freedom, and it is absent in single antenna systems [37]. Exploiting spatial degrees of freedom may be done for “multiplexing”, “diversity”, or a combination of both.

A communication system with a transmitter with many antennas and a receiver with many antennas is considered a single-user MIMO system. Similarly, a communication system with a single transmitter, but many devices on the other side, each with one or more antennas, is categorized as MU-MIMO system. More introduction to MU-MIMO is given in Section 4.

Another important extension of MIMO is the massive MIMO framework. It may be considered a massive MIMO network consisting of more than one MU-MIMO infrastructure. This type of network employs three concepts of beamforming, spatial multiplexing, and spatial diversity.

Precoding is a key process in MIMO employed to map  $K$  data streams to  $Tr_x$  transmitting antennas. In single-user MIMO,  $K$  data streams belong to single-user, but in MU-MIMO  $K$  data streams are intended for  $K$  users with a single symbol per user. Some well-known precoding techniques are SVD precoding or optimal unitary precoding, codebook-based precoding, ZF precoding, and DPC. The first three are linear, while DPC is a non-linear coding method.

## 4. System Model

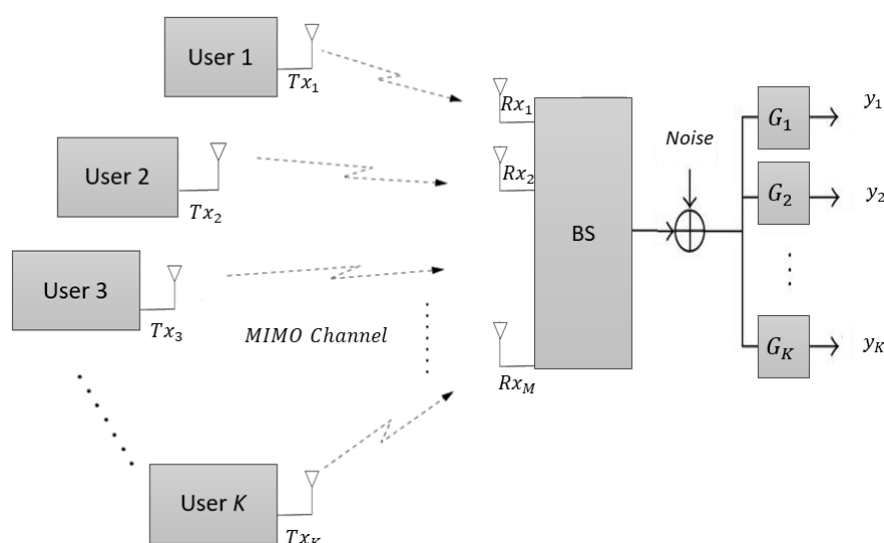
Before explaining the system model’s main components, such as the problem statement, its formulation into RL framework, and then RL-based scheduling, we first briefly explain BS, which is an important component of wireless communication. 5G and 6G

massive MIMO concept aims to enhance throughput and optimize the energy efficiency of the wireless communication system [38].

BSs can be considered 5G cell internet towers. The number of cellular BSs keeps increasing due to the increasing demand for cellular devices. Moreover, a rise in cell phone users for data-heavy operations results in a strain on the existing towers. So, more BSs can help to enhance transmission between devices and cell tower antennas. The 5G technology employs mmWave signals that do not cover large areas as in 3G/4G networks. Therefore, specialized BSs are required to tackle the 5G mobile traffic.

#### 4.1. Problem Statement

An MU-MIMO system in uplink configuration has a single BS that has  $Rx_M$  receive antennas and  $k$  users each with a single transmit antenna as shown in Figure 2. When ( $k > Rx_M$ ), then BS needs to choose a group of users to allow them for transmission, and the selected group of users should be equal to available receive antennas ( $Rx_M$ ) at BS.



**Figure 2.** MU-MIMO uplink system with  $K$  users each with a single antenna and a BS with  $Rx_M$  receive antennas.

This scenario requires a user scheduling mechanism at BS before allowing the user to transmit. We consider the Rayleigh fading channel, and at each TTS,  $N$  users may be scheduled to uplink, while  $N \leq Rx_M$ . The data received by the BS can be written as given in Equation (2) [3].

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \quad (2)$$

where  $\mathbf{H}(Rx_M \times N) = \sum_{i=1}^K \sqrt{P_i} \mathbf{h}_i x_i$  is Rayleigh fading channel and  $\mathbf{x} = [x_1, x_2, \dots, x_k]^T$  and  $[\cdot]^t$  denotes the transpose. While  $x_i \in \mathbb{C}^{N \times 1}$  is the transmitted signal,  $P_i$  large-scale received power of one of the selected users, respectively, and  $n \sim \mathcal{N}(0, \sigma^2)$  is the AWGN.

#### 4.2. Problem Formulation

As discussed in Section 3, an RL problem needs to be modeled as an MDP, that includes states, actions, and rewards.

- **State**—Each underlying problem has a state or set of states that an RL agent may visit/explore. We map a state at time  $t$  as the combination of multiple parameters. First, we consider that each candidate user can determine its transmit beamforming vector using CSI, which is locally available [39]. Secondly, we use the combination of SINR, Gram Schmidt Orthogonalization, and SLNR as given in Equation (3). Each parameter is weighted equally, normalized between zero and one, and works as prior



information for the BS. The former information is available to each user while the latter is received through BS.

$$s_t = (SINR_m, g_m, SLNR_m) \quad (3)$$

where  $SINR_m$ ,  $SLNR_m$  and  $g_m$  indicate values of SINR, SLNR and Gram Schmidt Orthogonalization, respectively, for  $m_{th}$  user form a state at time  $t$ , may be calculated as given in Equation (4) according to [19], Equation (5) according to [30] and Equation (8), respectively.

$$SINR_m = \frac{\|h_m w_m\|^2}{\sigma^2 + \sum_{g(i) \in G} \|h_m w_{g(i)}\|^2} \quad (4)$$

where  $G$  is the subgroup of selected users, while  $h_m$  and  $w_m = h^*(hh^*)^{-1}$  are channel and precoding vector of  $m_{th}$  user, respectively.

$$SLNR_m = \frac{\|h_m w_m\|^2}{\sigma_m^2 + \sum_{m \neq i} \|h_m w_m\|^2} \quad (5)$$

where

$$w_m \propto \text{eigenvector}_{\lambda_{max}}((\sigma_m^2 I + H_m^* H_m)^{-1} h_m^* h_m) \quad (6)$$

The  $\lambda_{max}$  indicates the maximum SLNR of a user.

To calculate  $g_m$  for all users, the component of  $h_m$  orthogonal to subspace spanned by  $[g_1 \dots g_{i-1}]$  as in [32].

$$g_m = h_m - \sum_{j=1}^{i-1} \frac{h_m * g_j^*}{\|g_j\|^2} * g_j \quad (7)$$

$$= h_m \left(1 - \sum_{j=1}^{i-1} \frac{g_j * g_j^*}{\|g_j\|^2}\right) \quad (8)$$

When  $i = 1$ , then  $g_m = h_m$ .

- **Action**—Each user (agent) has to choose a resource block to transmit its data to BS. Therefore, an action is chosen by an agent  $m$  at time step  $t$  and can be written as given in Equation (9) as

$$a_t = (U_{1,rb_1}, U_{2,rb_1}, \dots, U_{m,rb_n}, \dots, U_{k,rb_z}) \quad (9)$$

where  $U_{m,rb_n}$  represents that  $m_{th}$  user selects  $n_{th}$  resource block for data transmission.

- **Reward**—The natural objective of any user selection technique is to enhance the system capacity and optimal utilization of available radio resources. A reward function defines the goal in a RL problem. In every time step, the environment feedbacks to the RL agent a number as a reward. The agent's goal is to maximize the total reward it receives over the long run. The reward or feedback thus defines the good and bad actions for the agent [6]. Sumrate is the metric used to indicate system performance in MU-MIMO systems. Therefore, the reward for the RL agent will be the aggregated sum-rate for all selected users as given in Equation (10) according to [28].

$$R = C_{sum} = \sum_{j \in G} \log_2(1 + SINR_j) \quad (10)$$

where  $G$  is the group of users selected for transmission.

After having formulated an MU-MIMO user scheduling problem into a MDP by defining states, actions, and rewards, next we move toward the solution using the Bayesian RL method. The advantage of a Bayesian RL technique is that an agent can use the initial

knowledge available in Equation (3), helping an agent learn and converge faster than classic RL approaches. Equation (11) needs to be solved for Bayesian RL.

$$V^{\pi^*}(x, b) = \max_a \sum_{x'} Pr(x' | x, b, a) [X'_r + \gamma V^{\pi^*}(x', b_{xax'})] \quad (11)$$

where the state as defined in Equation (3) is represented with  $X$  and distribution over the unknown  $\theta$  is with  $b$ , respectively. The  $Pr(x' | x, b, a)$  is used for transition probability. The probabilities may also be utilized as the posterior distribution  $P(\theta|x)$  by employing  $\theta$  given the initial information as given in Equation (1) and rewritten below in Equation (12).

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \quad (12)$$

We considered  $P(x|\theta)$  and  $P(\theta)$  as the Bernoulli and Beta distribution, respectively. As  $P(x|\theta)$  and  $P(\theta)$  are Bernoulli and Beta distributions, respectively, and since Beta distribution is a conjugate prior to the Bernoulli distribution, therefore,  $P(\theta|x)$  is considered as Beta distributed. This indicates that when the aggregated sumrate increased after inclusion of a user, then  $(P(\theta|x))$  should be Beta( $\alpha + 1, \beta$ ). Similarly, when the aggregated sumrate decreases after inclusion of a user, then the  $(P(\theta|x))$  should be Beta( $\alpha, \beta + 1$ ).

To address the exploration–exploitation issue, we use Thompson Sampling [40]. It is a scheme for decision-making problems, and actions are performed sequentially to keep a balance between known actions (exploitation) and exploring new actions (exploration). In Thompson Sampling, the probability  $P(\theta|x)$  is sampled from the prior. Then, the user with the highest sampled probability is moved to the group of users who will be allowed for transmission by the BS in each TTS.

#### 4.3. RL-Based User Scheduling

This subsection focuses on the proposed RL-based user scheduling scheme. The detailed implementation of the methodology is shown in Figure 3 and a step-by-step explanation is given below.

- Each user determines its transmit beamforming vector to quantify the amount of interference from the other users and resource block for data transmission as defined in Equation (9).
- As we are using the Rayleigh fading channel, which is to say that the channel gain from any of  $Rx_M$  antennas to a user is described by ZMCSCG RV and this becomes a model that is suitable for narrow-band networks functioning in non-line-of-sight scenarios [41].
- BS After receiving feedback from users in the form of a priority resource block and amount of interference, calculate  $(SINR, SLNR, g)$  at each time TTS. Following BS selects the subset of users to equal the antenna number at BS to allow users for transmission based on all estimated information.
- After selecting a group of users, the aggregate sum-rate is calculated according to Equation (10). This sum-rate acts as feedback (reward), and based on the obtained reward for the selected users; each user can choose the most suitable resource block for transmission, which results in optimal utilization of available resources and enhancement in system capacity.



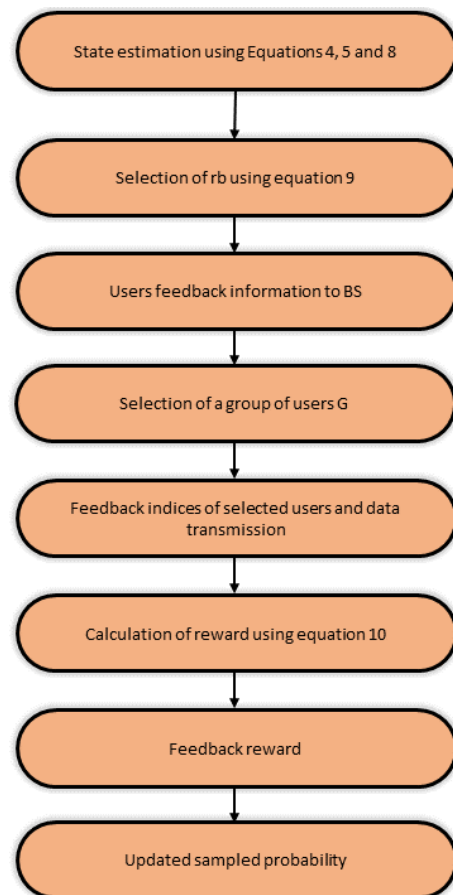
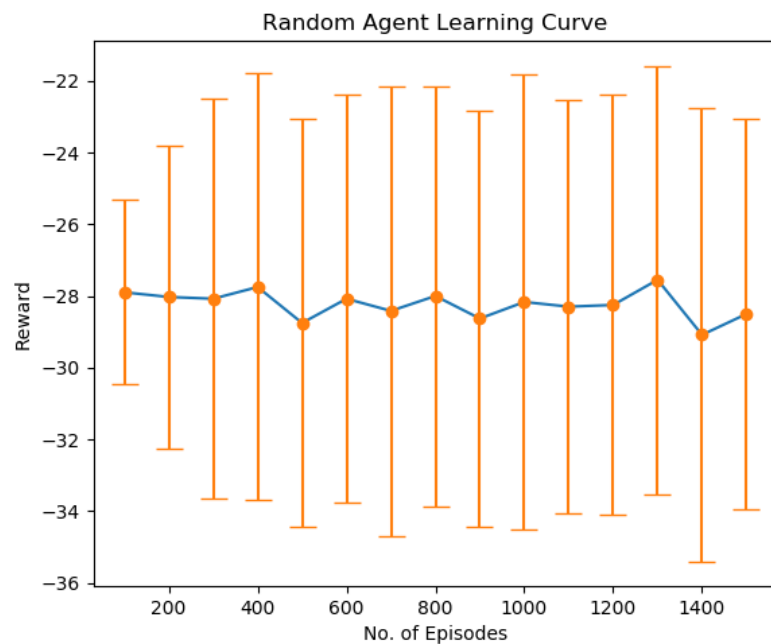


Figure 3. Flowchart of scheduling mechanism.

## 5. Results

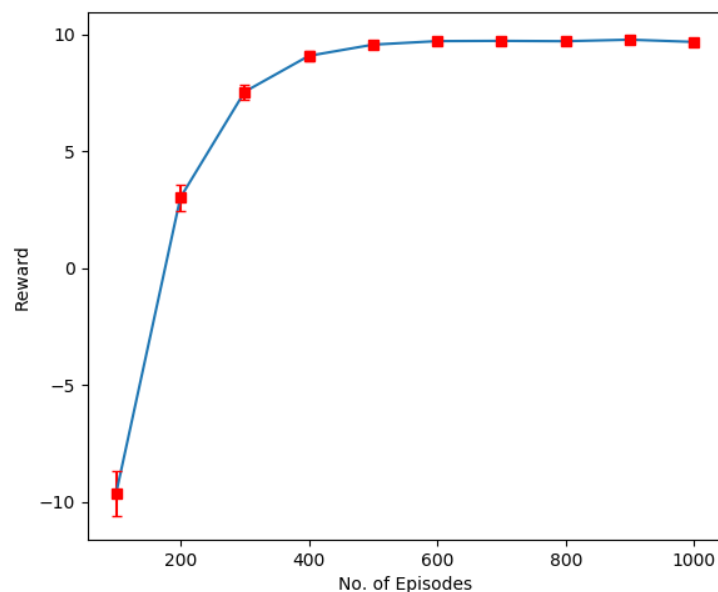
This section reports the simulation results to show the performance of the proposed methodology. For simplicity, consider that each user has a single antenna and a BS with four and five receive antennas. Extension to more transmit and receive antennas may be studied in [42]. We consider a uniform transmit power for all data streams. The proposed scheme is compared with the Random, Low Complexity (LOWC) algorithm, SINR, and SLNR-based scheduling techniques for the Rayleigh fading channel. The performance metric we considered for evaluating an algorithm is the aggregated sum-rate of all the selected users when the BS has four receive antennas, then the number of users in the subset of the selected users will be four.

The first result shown in Figure 4 indicates the result when the user randomly selects a resource block for data transmission without considering the overall system capacity and other users' priorities. We can see that the learning is mainly imbalanced and results in degraded performance. This is because the agent chooses action randomly and does not learn the environment's dynamics. The experiment was conducted for 30 users, each with a single transmit antenna and four receiving antennas at BS.



**Figure 4.** Sumrate performance of Random algorithm.

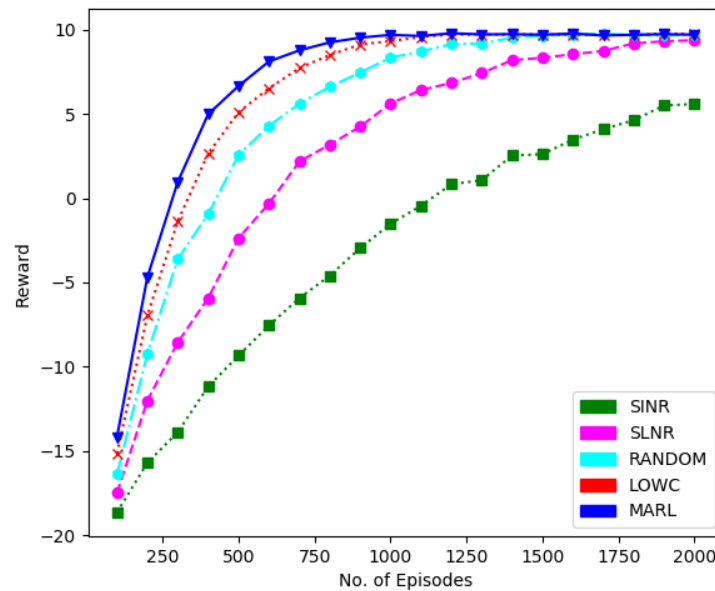
On the contrary, in Figure 5, when the users consider the aggregated reward, which is the sum rate of all selected users, the overall performance is enhanced due to faster learning and convergence. Now, the agent is not taking actions randomly. Instead, it is continuously learning from the environment. At each time slot, a subset of users is selected for service, and the sum rate (reward) for the selected group of users is computed. Each time the better value of the sum rate over the previous one serves as positive feedback, and the agent learns that the chosen action was correct. While in case of negative feedback, the agent avoids choosing that particular action in that specific situation (state) again. The result is obtained for configuration four receiving antennas at BS and 20 users, each with a single antenna.



**Figure 5.** Sumrate performance of RL algorithm for 30 users, four receive antennas at BS and one transmit antenna at each user.

Furthermore, in Figure 6, we have presented a performance comparison in terms of sum-rate to demonstrate the feasibility of the RL method over the other methods. We have

considered 30 users, each with a single antenna and five receiving antennas at BS. It is evident that the proposed technique performs better than the others, while the random algorithm performs worst.

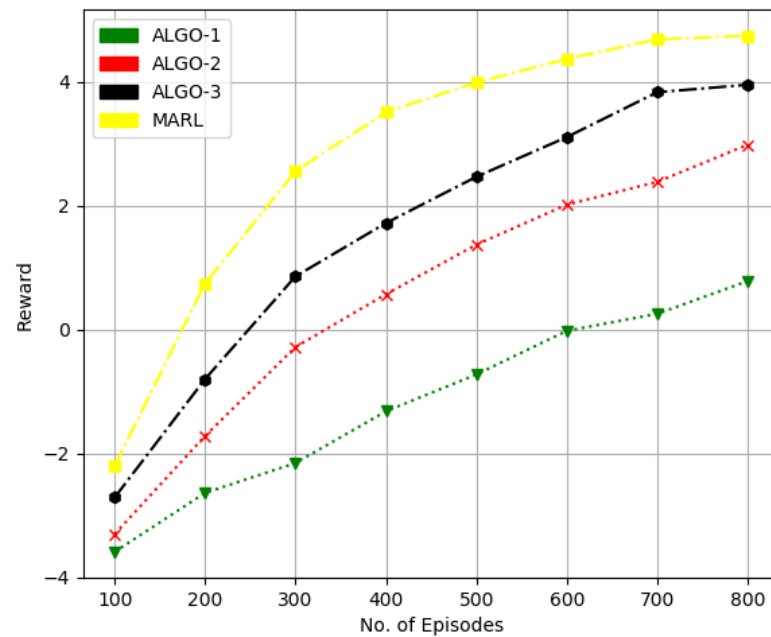


**Figure 6.** Sumrate performance comparison of different techniques for 30 users, five receive antennas at BS and one transmit antenna at each user.

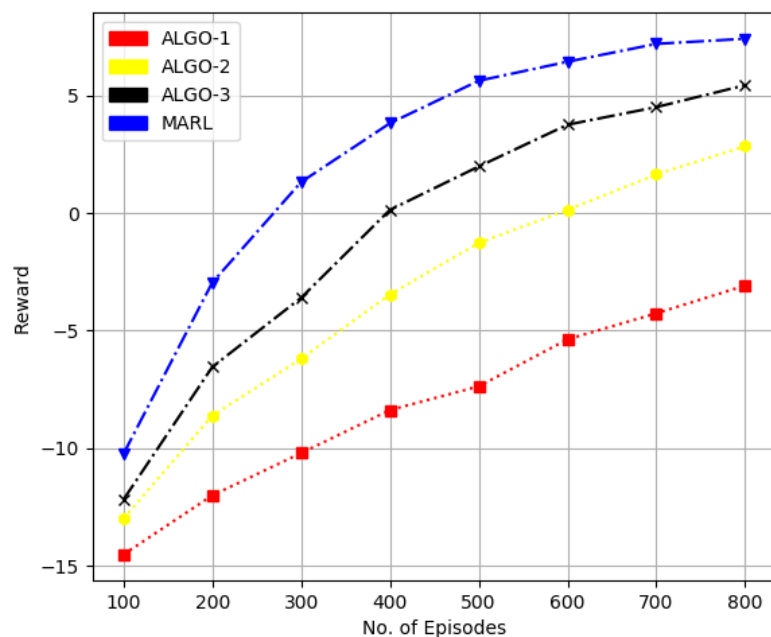
Furthermore, we have found that the selection of the first user in the SLNR and SINR scheduling method (i.e., the user with maximum channel gain) is not realistic. We have investigated how the first user selection can be made arbitrary. Even a random choice of the first user will not affect the aggregated sum-rate performance of the SLNR and SINR algorithms.

In Figures 7 and 8, we considered the performance of the proposed method and state of the algorithms referred as ALGO-1, ALGO-2 and ALGO-3 in Figures 7 and 8 were presented in [33–35], respectively. The work in [33] uses the POMDP method for downlink transmit beamforming at BS equipped with multi-antennas. While a binary FPA is employed in [34] for both antenna and user scheduling to obtain sum-rate performance with comparatively less computational complexity. The framework in [35] applies a joint infrastructure for a hybrid precoder, and user scheduling is proposed.

In Figure 7, we considered ten candidate users, each with a single transmit antenna and four receive antennas at BS. At the same time, we increased the total number of candidate users to 20 in Figure 7 while keeping other parameters the same. We can quantify that the RL-based method obtains about 20% higher sum-rate performance than that of the ALGO-3 and quite significantly better performance over the other two schemes.



**Figure 7.** Performance comparison in terms of sumrate for 10 users, four receive antennas at BS and one transmit antenna at each user.



**Figure 8.** Performance comparison in terms of sumrate for 20 users, four receive antennas at BS and one transmit antenna at each user.

## 6. Conclusions

We have considered the user selection problem in an MU-MIMO system and introduced the multi-agent RL methodology as one of the solutions. We first modeled the user scheduling problem as MDP by defining states, actions, and rewards. We developed an optimal scheduling policy by optimally selecting a group of users for transmission with the help of RL.

The simulation results demonstrate that the proposed methodology provides a significant performance enhancement and indicates that AI-based techniques could play a vital role in communication systems. We have conducted experiments using four and five receive antennas at BS for 10, 20, and 30 users. A 20% higher sum-rate performance of the

proposed scheme is reported for 10 and 20 users, while it achieves a slightly better sum-rate for 30 users and takes around 500 fewer episodes for convergence.

**Author Contributions:** Conceptualization, M.N.; Investigation, Z.U. and S.B.; Methodology, A.C.; Resources, G.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

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

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BS	Base Station
RV	Random Variable
RA	Resource Allocation
SINR	Signal-to-Interference plus Noise Ratio
POMDP	Partially Observable MDP
SLNR	Signal-to-Leakage plus Noise Ratio
SNR	Signal-to-Noise Ratio
AWGN	Additive White Gaussian Noise
ZF	Zero Forcing
DPC	Dirty Paper Coding
SVD	Singular Value Decomposition
ML	Machine Learning
CSI	Channel State Information
FPA	Flower Pollination Algorithm
TTS	Transmission Time Slot
DL	Deep Learning
DNN	Deep Neural Networks
RL	Reinforcement Learning
MIMO	Multiple Input Multiple Output
MU-MIMO	Multi User MIMO
MDP	Markov Decision Process

## References

1. Foshini, G. On limits of wireless communication in a fading environment when using multiple antenna. *Wirel. Pers. Commun.* **1998**, *6*, 315–335.
2. Sareddeen, H.; Mansour, M.M.; Jalloul, L.M.; Chehab, A. Efficient near optimal joint modulation classification and detection for MU-MIMO systems. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 3706–3710.
3. Jindal, N.; Goldsmith, A. Dirty-paper coding versus TDMA for MIMO broadcast channels. *IEEE Trans. Inf. Theory* **2005**, *51*, 1783–1794. [[CrossRef](#)]
4. Lee, J.; Jindal, N. Dirty paper coding vs. linear precoding for MIMO broadcast channels. In Proceedings of the 2006 Fortieth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 29 October–1 November 2006; pp. 779–783.
5. Peel, C.B.; Hochwald, B.M.; Swindlehurst, A.L. A vector-perturbation technique for near-capacity multi-antenna multiuser communication-part I: Channel inversion and regularization. *IEEE Trans. Commun.* **2005**, *53*, 195–202. [[CrossRef](#)]
6. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
7. Di Sarno, C.; Formicola, V.; Sicuranza, M.; Paragliola, G. Addressing Security Issues of Electronic Health Record Systems through Enhanced SIEM Technology. In Proceedings of the 2013 International Conference on Availability, Reliability and Security, Regensburg, Germany, 2–6 September 2013; pp. 646–653. [[CrossRef](#)]
8. Coronato, A.; Di Napoli, C.; Paragliola, G.; Serino, L. Intelligent Planning of Onshore Touristic Itineraries for Cruise Passengers in a Smart City. In Proceedings of the 2021 17th International Conference on Intelligent Environments (IE), Dubai, United Arab Emirates, 21–24 June 2021; pp. 1–7. [[CrossRef](#)]
9. Coronato, A.; de Pietro, G.; Paragliola, G. A Monitoring System Enhanced by Means of Situation-Awareness for Cognitive Impaired People. In Proceedings of the 8th International Conference on Body Area Networks, Boston, MA, USA, 30 September–2 October 2013; ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering): Brussels, Belgium, 2013; pp. 124–127. [[CrossRef](#)]

10. Paragliola, G.; Coronato, A.; Naeem, M.; De Pietro, G. A Reinforcement Learning-Based Approach for the Risk Management of e-Health Environments: A Case Study. In Proceedings of the 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Las Palmas de Gran Canaria, Spain, 26–29 November 2018; pp. 711–716. [[CrossRef](#)]
11. Naeem, M.; Rizvi, S.T.H.; Coronato, A. A Gentle Introduction to Reinforcement Learning and its Application in Different Fields. *IEEE Access* **2020**, *8*, 209320–209344. [[CrossRef](#)]
12. Van Chien, T.; Canh, T.N.; Björnson, E.; Larsson, E.G. Power control in cellular massive MIMO with varying user activity: A deep learning solution. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 5732–5748. [[CrossRef](#)]
13. Nie, J.; Haykin, S. A Q-learning-based dynamic channel assignment technique for mobile communication systems. *IEEE Trans. Veh. Technol.* **1999**, *48*, 1676–1687.
14. Bennis, M.; Niyato, D. A Q-learning based approach to interference avoidance in self-organized femtocell networks. In Proceedings of the 2010 IEEE Globecom Workshops, Miami, FL, USA, 6–10 December 2010; pp. 706–710.
15. Santos, E.C. A simple reinforcement learning mechanism for resource allocation in lte-a networks with markov decision process and q-learning. *arXiv* **2017**, arXiv:1709.09312.
16. Kong, P.Y.; Panaitopol, D. Reinforcement learning approach to dynamic activation of base station resources in wireless networks. In Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, 8–11 September 2013; pp. 3264–3268.
17. Kurve, A. Multi-user MIMO systems: The future in the making. *IEEE Potentials* **2009**, *28*, 37–42. [[CrossRef](#)]
18. Naeem, M.; Bashir, S.; Khan, M.U.; Syed, A.A. Performance comparison of scheduling algorithms for MU-MIMO systems. In Proceedings of the 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 12–16 January 2016; pp. 601–606.
19. Naeem, M.; Bashir, S.; Khan, M.U.; Syed, A.A. Modified SINR based user selection for MU-MIMO systems. In Proceedings of the 2015 International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, 12–13 December 2015; pp. 1–4.
20. Wei, C.; Xu, K.; Xia, X.; Su, Q.; Shen, M.; Xie, W.; Li, C. User-centric access point selection in cell-free massive MIMO systems: A game-theoretic approach. *IEEE Commun. Lett.* **2022**, *26*, 2225–2229. [[CrossRef](#)]
21. Carvajal, H.; Orozco, N.; Cacuango, S.; Salazar, P.; Rosero, E.; Almeida, F. A Scheduling Scheme for Improving the Performance and Security of MU-MIMO Systems. *Sensors* **2022**, *22*, 5369. [[CrossRef](#)]
22. Yin, Z.; Chen, J.; Li, G.; Wang, H.; He, W.; Ni, Y. A Deep Learning-Based User Selection Scheme for Cooperative NOMA System with Imperfect CSI. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 7732029. [[CrossRef](#)]
23. Jang, J.; Lee, H.; Kim, I.M.; Lee, I. Deep Learning for Multi-User MIMO Systems: Joint Design of Pilot, Limited Feedback, and Precoding. *IEEE Trans. Commun.* **2022**, *20*, 4044–4057. [[CrossRef](#)]
24. Ahmed, I.; Shahid, M.K.; Khammari, H.; Masud, M. Machine Learning Based Beam Selection with Low Complexity Hybrid Beamforming Design for 5G Massive MIMO Systems. *IEEE Trans. Green Commun. Netw.* **2021**, *5*, 2160–2173. [[CrossRef](#)]
25. Salh, A.; Audah, L.; Abdullah, Q.; Shah, N.S.M.; Shipun, A. Energy-efficient low-complexity algorithm in 5G massive MIMO systems. *Comput. Mater. Contin.* **2021**, *67*, 3189–3214. [[CrossRef](#)]
26. Perdana, R.H.Y.; Nguyen, T.V.; An, B. Deep Learning-based Power Allocation in Massive MIMO Systems with SLNR and SINR Criteria. In Proceedings of the 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), Jeju Island, Korea, 17–20 August 2021; pp. 87–92.
27. Perdana, R.H.Y.; Nguyen, T.V.; An, B. Deep neural network design with SLNR and SINR criteria for downlink power allocation in multi-cell multi-user massive MIMO systems. *ICT Express* **2022**. [[CrossRef](#)]
28. Xia, X.; Wu, G.; Liu, J.; Li, S. Leakage-based user scheduling in MU-MIMO broadcast channel. *Sci. China Ser. F Inf. Sci.* **2009**, *52*, 2259–2268. [[CrossRef](#)]
29. Xia, X.; Wu, G.; Fang, S.; Li, S. SINR or SLNR: In successive user scheduling in mu-mimo broadcast channel with finite rate feedback. In Proceedings of the 2010 International Conference on Communications and Mobile Computing, Shenzhen, China, 12–14 April 2010; pp. 383–387.
30. Naeem, M.; Khan, M.U.; Bashir, S.; Syed, A.A. Modified leakage based user selection for MU-MIMO systems. In Proceedings of the 2015 13th International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 14–16 December 2015; pp. 320–323.
31. Zhao, L.; Li, B.; Meng, K.; Gong, B.; Zhou, Y. A novel user scheduling for multiuser MIMO systems with block diagonalization. In Proceedings of the 2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, 8–11 September 2013; pp. 1371–1375.
32. Naeem, M.; Bashir, S.; Ullah, Z.; Syed, A.A. A near optimal scheduling algorithm for efficient radio resource management in multi-user MIMO systems. *Wirel. Pers. Commun.* **2019**, *106*, 1411–1427. [[CrossRef](#)]
33. Sharifi, S. A POMDP Framework for Antenna Selection and User Scheduling in Multi-User Massive MIMO Systems. Ph.D. Thesis, Ontario Tech University, Oshawa, ON, Canada, 2022.
34. Mohanty, J.; Pattanayak, P.; Nandi, A.; Baishnab, K.L.; Talukdar, F.A. Binary flower pollination algorithm based user scheduling for multiuser MIMO systems. *Turk. J. Electr. Eng. Comput. Sci.* **2022**, *30*, 1317–1336. [[CrossRef](#)]
35. Rajarajeswarie, B.; Sandanalakshmi, R. Machine learning based hybrid precoder with user scheduling technique for maximizing sum rate in downlink MU-MIMO system. *Int. J. Inf. Technol.* **2022**, *14*, 2399–2405. [[CrossRef](#)]



36. Ghavamzadeh, M.; Mannor, S.; Pineau, J.; Tamar, A. Bayesian Reinforcement Learning: A Survey. *Found. Trends Mach. Learn.* **2016**, *8*, 102–109. [[CrossRef](#)]
37. Naeem, M.; De Pietro, G.; Coronato, A. Application of reinforcement learning and deep learning in multiple-input and multiple-output (MIMO) systems. *Sensors* **2021**, *22*, 309. [[CrossRef](#)] [[PubMed](#)]
38. He, S.; Du, J.; Liao, Y. Multi-User Scheduling for 6G V2X Ultra-Massive MIMO System. *Sensors* **2021**, *21*, 6742. [[CrossRef](#)] [[PubMed](#)]
39. Lee, B.O.; Je, H.W.; Sohn, I.; Shin, O.S.; Lee, K.B. Interference-Aware Decentralized Precoding for Multicell MIMO TDD Systems. In Proceedings of the IEEE GLOBECOM 2008—2008 IEEE Global Telecommunications Conference, New Orleans, LA, USA, 30 November–4 December 2008; pp. 1–5. [[CrossRef](#)]
40. Dearden, R.; Friedman, N.; Andre, D. Model-Based Bayesian Exploration. *arXiv* **2013**, arXiv:1301.6690.
41. Tse, D.; Viswanath, P. *Fundamentals of Wireless Communication*; Cambridge University Press: Cambridge, UK, 2005.
42. Yoo, T.; Goldsmith, A. On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming. *IEEE J. Sel. Areas Commun.* **2006**, *24*, 528–541.