future internet

MDPI

Article

Resource Allocation for D2D Communications in Multi-Slice NOMA-Based Cellular Networks

Lijun Dong, Jingjing Wu * and Yitong Yang

School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
* Correspondence: wujingjing@cse.neu.edu.cn

Abstract

Significant challenges will be encountered in next-generation cellular networks to achieve both high spectral efficiency (SE) and diverse quality of service (QoS) requirements simultaneously, particularly under stringent bandwidth and power budgets within highly dynamic and dense topologies. To address these challenges, we formulate an optimization problem in a multi-slice non-orthogonal multiple access (NOMA) system with underlay device-to-device (D2D) communications. This problem aims to maximize SE and satisfy user QoS demands by jointly optimizing power allocation and resource block (RB) assignment. To solve this non-convex and NP-hard problem, we propose a resource allocation mechanism based on joint optimization and cooperative multi-agent deep reinforcement learning (MADRL). Specifically, we construct an optimization framework based on successive convex approximation (SCA) and the Lagrange duality method to derive an analytical iterative solution for the optimal power allocation under a given RB assignment, thereby avoiding the inherent discretization error of the action space in pure learning methods. Furthermore, we propose a cooperative multi-agent algorithm based on dueling double deep Q-Network (CMAD3QN) to address the discrete RB assignment problem. Simulation results demonstrate that, compared with benchmark schemes, the proposed scheme exhibits faster convergence speed and significantly enhances system spectral efficiency while ensuring slice isolation and resource constraints.

Keywords: non-orthogonal multiple access; network slicing; device-to-device communications; resource allocation; joint optimization; D3QN

1. Introduction

Cellular networks have become the core component of modern wireless communication systems, supporting a diverse range of services from basic voice calls to high-bandwidth multimedia streaming [1]. With the continuous emergence of novel services and network applications [2], both the number of connected devices and the volume of data traffic are expected to increase dramatically. Messou et al. predict that the global number of connected devices will exceed 40 billion by 2034 [3]. The ever-growing data volume has exacerbated the scarcity of spectrum resources. Consequently, overcoming the bottleneck issue of SE under limited bandwidth budgets has emerged as a core research topic for the next-generation wireless networks [4].

To address the aforementioned challenges, it is imperative to apply new key technologies in next-generation networks [5]. Among these technologies, NOMA has been identified by the International Telecommunication Union as one of the future technological trends for 2030 and beyond. Unlike orthogonal multiple access (OMA), which allocates

Academic Editor: Paolo Bellavista

Received: 31 March 2026
Revised: 29 April 2026
Accepted: 2 May 2026
Published: 6 May 2026

resources to users in an orthogonal manner, NOMA employs superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver to suppress co-channel interference [6]. This allows multiple users to multiplex the same frequency resource non-orthogonally in the power domain, thereby enhancing SE [7]. Similarly, D2D communication is regarded as a highly promising solution for improving SE, capable of facilitating the diversion of core network traffic and reducing end-to-end latency [8]. Compared with traditional cellular communication, D2D communication establishes direct data links between communication pairs without relaying through a base station (BS) [9]. This enables the full utilization of the favorable channel conditions in close proximity and creates more opportunities for spectrum reuse. D2D communication can be categorized into two operating modes. The overlay mode, where D2D transmissions occupy dedicated resources, and the underlay mode, where D2D pairs share wireless resources with cellular users (CUs). Although the overlay mode achieves interference-free communication, it comes at the cost of lower spectral efficiency. On the other hand, if the cross-link interference can be strictly controlled, the underlay mode can potentially achieve higher spectral efficiency [10]. To meet the massive connectivity demands of the next-generation networks, our research focuses on integrating underlay D2D communication into NOMA-based cellular systems.

Furthermore, different application types impose varying QoS requirements on the connected users [11]. The traditional "one-size-fits-all" unified network architecture is unable to simultaneously meet the dual demands of guaranteeing differentiated QoS and enhancing resource utilization [12]. By leveraging software-defined networking (SDN) [13] and network function virtualization (NFV) [14] technologies, network slicing virtualizes the physical network into multiple logically isolated slices, ensuring that each application type can obtain the necessary resources. Although the network isolation achieved through slicing inherently causes some spectral efficiency loss, integrating NOMA can effectively alleviate this issue, thereby enabling a balance between high resource utilization and the differentiated QoS requirements of diverse services [15].

Although the integration of multi-slice NOMA systems with underlay D2D communications can simultaneously achieve high SE and differentiated services, it inevitably increases intra-cell co-channel interference. To mitigate interference within cellular networks, effective power allocation and RB assignment become crucial [16]. However, finding an appropriate balance for resource allocation is challenging. On the one hand, it is necessary to ensure the stability of both links and the system, while the constraints of various devices differ across diverse scenarios. On the other hand, the limited resources of cellular networks and the increasing user demand for data traffic represent conflicting requirements. Consequently, designing efficient resource allocation algorithms is essential to ensure fairness among multiple users and optimize overall system performance. Traditional convex optimization-based methods, while theoretically rigorous, suffer from high computational complexity, making them difficult to adapt to highly dynamic networks. Conversely, pure deep reinforcement learning (DRL) methods often necessitate discretization when handling continuous variables, leading to quantization errors [17]. To address these issues, we propose a resource allocation mechanism based on joint optimization and CMAD3QN to exploit the complementary strengths of both paradigms. Specifically, by employing SCA and Lagrange duality, we derive the optimal power allocation for all CUs, which entirely eliminates the discretization error of the action space. Meanwhile, the high-dimensional discrete resource block assignment is delegated to the CMAD3QN framework. This distributed architecture effectively mitigates the curse of dimensionality and ensures robust decision-making in complex interference environments. In summary, the main contributions of this study are outlined as follows:

- NOMA-enabled multi-slice cellular network communication model is established, where underlay D2D communication is introduced to reuse spectrum resources. This model maps users to different slices according to their differentiated QoS requirements. Within each slice, NOMA is employed to enhance SE, thereby maximizing spectrum reuse gains while guaranteeing service isolation.
- We formulate a joint optimization problem for wireless RB assignment and power allocation. Specifically, the RB assignment encompasses both the RB assignment for CUs and the access control for D2D pairs. The objective is to maximize the total system SE while meeting the QoS requirements for each slice and the power budgets.
- To address this non-convex problem, we propose a distributed resource allocation mechanism based on Joint Optimization and CMAD3QN (JOCMAD3QN). This mechanism innovatively integrates theoretical optimization with intelligent learning. Specifically, SCA and Lagrange dual decomposition are employed to precisely solve the continuous power allocation sub-problem, thereby ensuring theoretical optimality. Meanwhile, an improved CMAD3QN algorithm is utilized to handle the complex discrete RB assignment, effectively resolving the challenge of searching high-dimensional state spaces in dynamic environments.
- We conduct extensive simulation experiments to compare the proposed approach with benchmark schemes and demonstrate that the proposed approach exhibits significant advantages in convergence speed and SE, thereby validating its effectiveness.

The remainder of this paper is organized as follows. Section 2 reviews the related work conducted in recent studies. Section 3 presents the system model and the problem formulation. Section 4 elaborates on the implementation details of our proposed algorithm. Section 5 analyzes the simulation results. Finally, Section 6 concludes this paper.

2. Related Work

In recent years, extensive research has investigated the application of NOMA in wireless communications, aiming to address key challenges such as SE, total throughput, and compatibility with traditional networks. In ref. [18], a single base station (BS) downlink scenario with two users was discussed, where equal bandwidth access was granted to both users. The authors in ref. [19] investigated a single-BS downlink scenario, specifically focusing on the case where only two users multiplex the same channel. A simulated annealing-based scheme was developed to solve the non-convex resource allocation problem with the objective of maximizing energy efficiency. Refs. [18,19] focused on simple single BS scenarios with a limited number of users. The authors in refs. [20–22] considered more complex system scenarios. Specifically, He et al. in ref. [20] employed a MIMO-NOMA uplink system based on group-level SIC, where users were grouped according to their distance from the BS to suppress inter-group interference. The authors in ref. [21] considered a tripartite interaction among users, edge servers, and BSs, focusing on a NOMA-based mobile edge computing system within multi-BS heterogeneous networks. This study aimed to minimize total system energy consumption through joint task offloading and resource allocation. Li et al. in ref. [22] proposed an uplink NOMA system integrated with movable antennas (MAs), aiming to maximize the sum rate by jointly optimizing MA positions, decoding order, and power control. To further exploit the potential of network capacity, numerous researchers have focused on integrating D2D communication into NOMA systems. Meng et al. in ref. [23] proposed a novel D2D-assisted cooperative NOMA model, employing a two-stage transmission mechanism to overcome the long-distance communication limitations of D2D users. In ref. [24], the authors achieved maximum energy efficiency of D2D underlay networks by joint power control and channel allocation. In ref. [25], Jose et al. considered a scenario where both D2D and cellular networks employ NOMA

technology. They derived the exact expressions for the system outage probability and optimized the power allocation for both CUs and D2D users to minimize this probability. In ref. [26], Moussa et al. integrated public safety users into 5G cellular networks by utilizing NOMA and underlay D2D technologies. To maximize the overall system throughput, they optimized the CU selection for each public safety cluster and the power allocation for each individual user within the cluster, while satisfying SIC constraints and guaranteeing the QoS requirements for both CUs and public safety users.

To enhance the provision of customized services for CUs with diverse requirements, recent studies have explored the utilization of network slicing in cellular networks. The authors in ref. [27] investigated the coexistence of eMBB, mMTC, and URLLC within a hybrid NOMA uplink system, and proposed a DRL-based algorithm to maximize long-term average efficiency while satisfying various QoS requirements. The coexistence of eMBB and URLLC slices was considered in ref. [28] and ref. [29]. The authors in ref. [28] aimed to maximize network throughput by jointly optimizing CU grouping, wireless subchannel allocation, and D2D admission, while satisfying the technical requirements of both slices. The authors in ref. [29] focused on minimizing end-to-end energy consumption and resource utilization costs, provided that the minimum radio access network data rate requirements for eMBB users and the end-to-end latency requirements for both types of users are met. The authors in ref. [30] pointed out that the existing 3GPP fixed slicing cannot meet the differentiated QoS requirements of heterogeneous networks, and proposed a design principle centered on on-demand adjustment of the number of slices and virtual network function entities. Lotfi et al. in ref. [31] addressed the issue where traditional fixed slicing schemes in 6G cell-free networks failed to adapt to dynamic network states. In ref. [32], Hossain et al. proposed a novel network slicing technique for NOMA-enabled mobile edge computing networks. Users were associated with different slices according to their diverse latency requirements, and a heuristic algorithm was developed to optimize the total uplink transmission energy consumption of the entire wireless network. However, the aforementioned studies have not yet integrated NOMA, D2D communications, and network slicing technologies into a unified design, which is capable of simultaneously achieving differentiated QoS guarantees and improving the utilization efficiency of spectrum resources in the cellular network scenario.

Given the complexity challenges faced by traditional optimization methods in handling highly dynamic and non-convex resource allocation problems, DRL has emerged as a key technology for resource management in next-generation networks due to its robust environment-sensing and decision-making capabilities. Researchers in refs. [30,33–35] leveraged DRL to explore efficient resource allocation strategies for NOMA and D2D integrated networks. For instance, in ref. [33] and ref. [34], DRL-based joint channel and power optimization schemes were proposed for NOMA-enhanced D2D underlay communications and D2D group communications, respectively, significantly improving the total system throughput. To address user fairness, Syed et al. designed an improved DRL algorithm in ref. [35], which effectively balances sum-rate maximization with fairness among NOMA users. DRL has also been widely applied in the dynamic resource orchestration of network slicing. The authors in ref. [30] proposed an RL-based slicing scheme that dynamically adjusts resources to meet the differentiated QoS requirements of users. To cope with more complex network environments, the authors in ref. [36] constructed a multi-level DRL framework under the 6G O-RAN architecture, achieving efficient management of slicing resources. In particular, for 5G edge-cloud networks and industrial internet of things scenarios, the D3QN has garnered significant attention due to its superior convergence performance. The authors in ref. [37] and ref. [38] leveraged D3QN to respectively solve the problems of slicing resource orchestration and multi-priority computation offloading,

and validated the stability of the algorithm when handling high-dimensional state spaces. However, as network scales expand, centralized DRL faces the challenge of state space explosion. Therefore, MADRL has emerged as a research hotspot. The authors in ref. [39] and ref. [40] respectively employed MADRL in grant-free NOMA systems and integrated terrestrial satellite networks, effectively solving the problems of energy efficiency optimization and dynamic resource management through distributed coordination. However, existing DRL methods often rely on action space discretization when handling continuous power variables, which can lead to quantization errors or excessively high action dimensions. To address these issues, this paper proposes a hybrid mechanism that integrates theoretical optimization with CMAD3QN.

3. System Model and Problem Formulation

In this section, we present the system model. Specifically, the network model is first introduced, followed by a detailed description of the signal models for CUs and D2D pairs. Subsequently, a slicing model based on user CUs' requirements is established. Finally, the optimization problem is formulated. Figure 1 illustrates the proposed system scenario, while Table 1 summarizes the notations used throughout this paper.

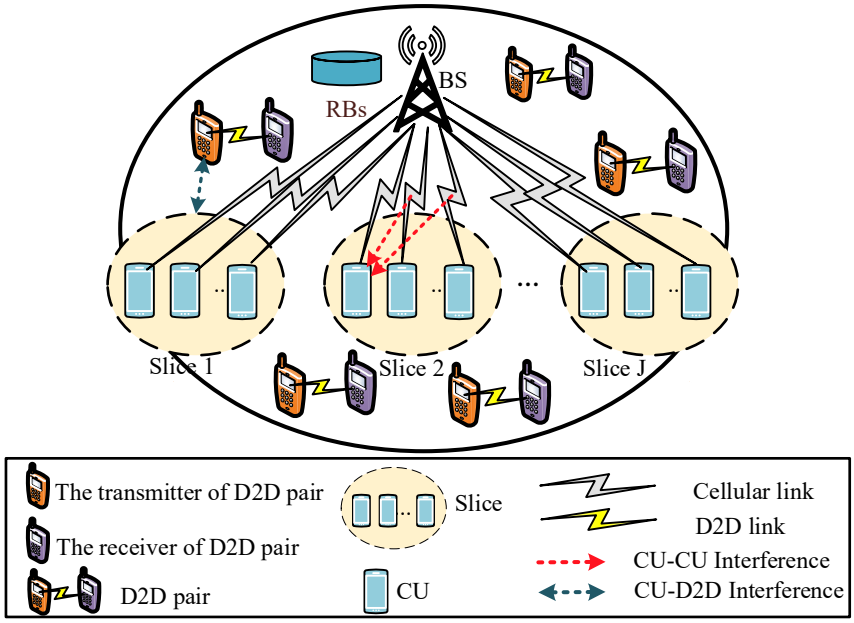

Figure 1. System model of a multi-slice NOMA-based cellular network with underlay D2D communications.

Table 1. List of Notations and Definitions.

Notation	Definition
$\mathcal{U}, \mathcal{D}, \mathcal{R}, \mathcal{J}$	The sets of CUs, D2D pairs, RBs, and network slices
d_{tr}, d_{re}	The transmitter and receiver of the dth D2D pair
B	The bandwidth of one RB
$\mathcal{U}_j$	The subset of CUs in the jth network slice
$u_{i,j}$	The ith CU in the jth network slice
$h^r_{m,n}$	The channel gain between transmitter m and receiver n on the rth RB
$l_{m,n}$	The distance between user m and user n
$\xi_{m,n}$	The lognormal shadowing with standard deviation ξ

Table 1. *Cont.*

Notation	Definition
$g^r_{m,n}$	The multipath fading
$x^r_{i,j}$, x^r_d	The transmit signals of $u_{i,j}$ and the dth D2D pair on the rth RB
$P^r_{i,j}$, P^r_d	The transmit power of $u_{i,j}$ and the dth D2D pair on the rth RB
N_0	The noise power spectral density
λ^j_d	The binary variable for the dth D2D pair shares the RB of the jth network slice
η^r_j	The binary variable for the jth network slice uses the rth RB
R^{QoS}_u, $R^{QoS}_{slice\ j}$	The QoS requirements of the uth CU and the jth network slice
R^{QoS}_{max}, R^{QoS}_{min}	The maximum and minimum QoS requirements of all CUs
$R^{min}_{slice\ j}$, R_0	The minimum QoS constraint of the jth network slice and D2D pairs
$R^r_{i,j}$, $\tilde{R}^r_{i,j}$	The data rate and the lower bound rate of $u_{i,j}$ on the rth RB
μ	The Lagrange multiplier related to the maximum transmit power constraint of BS
$\kappa_{i,j}$	The Lagrange multiplier vector related to the QoS requirements of $u_{i,j}$
$s_j(t)$	The network state of agent (slice) j at time slot t
$a_j(t)$	The action of agent (slice) j at time slot$t$
$f(t)$	The reward function at time slot t

3.1. Network Model

Consider the downlink transmission scenario of a power-domain NOMA-based cellular network. The system model comprises a BS located at the cell center, a set of CUs denoted as $\mathcal{U} = \{1, 2, \cdots, u, \cdots, U\}$ that are uniformly distributed within the cell, and a set of D2D pairs represented by $\mathcal{D} = \{1, 2, \cdots, d, \cdots, D\}$. The dth D2D pair is composed of a D2D transmitter d_{tr} and a D2D receiver d_{re}, constrained by a maximum transmission distance. For the sake of simplicity, it is assumed that the BS, each CU, and each D2D transceiver are equipped with a single antenna. The total system bandwidth is divided into R RBs, denoted by the set $\mathcal{R} = \{1, 2, \cdots, r, \cdots, R\}$, where the bandwidth of each RB is defined as B. Depending on their diverse QoS requirements, the CUs are allocated to J network slices. The set of network slices is denoted as $\mathcal{J} = \{1, 2, \cdots, j, \cdots, J\}$. Specifically, the set of CUs in the jth slice is given by $\mathcal{U}_j = \{1, 2, \cdots, u_j, \cdots, U_j\}$, and the uth CU within the jth slice is denoted as $u_{i,j}$.

The BS communicates with different slices using orthogonal frequency division multiple access (OFDMA). Consequently, the interference between slices is considered negligible. Within each slice, NOMA is employed to multiplex frequency resources. Specifically, multiple CUs in a slice are allocated the same RB to receive data from the BS, and SIC is adopted at the receiver to decode the signals. Meanwhile, the underlay D2D communication mode is adopted, allowing D2D pairs to reuse the RBs assigned to each NOMA group to enhance the network's spectral efficiency. Note that multiple D2D pairs are not permitted to reuse the same frequency resource simultaneously.

3.2. Signal Model

3.2.1. CUs Signal Model

Let $h^r_{m,n}$ denote the channel gain between transmitter m and receiver n on the rth RB, where $m \in \{BS, d_{tr}\}$ and $n \in \{u_{i,j}, d_{re}\}$. Specifically, $h^r_{m,n}$ can be represented as $h^r_{BS,u_{i,j}}$, $h^r_{d_{tr},u_{i,j}}$, $h^r_{d_{tr},d_{re}}$, and $h^r_{BS,d_{re}}$, which respectively represent the channel gains of the rth RB between the BS and $u_{i,j}$, between the D2D transmitter d_{tr} and $u_{i,j}$, between the D2D transmitter d_{tr} and the D2D receiver d_{re}, and between the BS and the D2D receiver d_{re}. The channel gain is expressed as:

$$h^r_{m,n} = \sqrt{PL(l)_{m,n}} \times \sqrt{\xi_{m,n}} \times g^r_{m,n} \tag{1}$$

where $PL(l)_{m,n} = K(l_{m,n}/l_0)^{-v}$ represents the path loss with K denoting the path loss constant, $l_{m,n}$ represents the distance between user m and user n, $l_0 = 1$ km is the reference distance unit, and v denotes the path loss exponent; $\xi_{m,n}$ represents lognormal shadowing fading with a standard deviation of ξ; and $g^r_{m,n}$ denotes multipath fading, which follows a complex Gaussian distribution with zero mean and unit variance, i.e., $g^r_{m,n} \sim \mathcal{CN}(0,1)$.

In this study, each RB is assigned to at most one network slice, and RBs cannot be shared among different network slices. Meanwhile, D2D users are permitted to reuse the RBs assigned to the network slices. The received signal of $u_{i,j}$ on the rth RB is given by:

$$y^r_{i,j} = \underbrace{\sum_{i=1}^{U_j} h^r_{BS,u_{i,j}} \sqrt{P^r_{i,j}} x^r_{i,j}}_{signal\ from\ BS} + \underbrace{\sum_{d=1}^{D} \lambda^j_d h^r_{d_{tr},u_{i,j}} \sqrt{P^r_d} x^r_d}_{interference\ from\ D2D\ pairs} + n \tag{2}$$

where $x^r_{i,j}$ and x^r_d denote the transmitted signals of $u_{i,j}$ and D2D pair d on the rth RB, respectively. $P^r_{i,j}$ and P^r_d denote the transmit power of $u_{i,j}$ and D2D pair d on the rth RB, respectively. $n \sim \mathcal{CN}(0, BN_0)$ represents the additive white Gaussian noise, where N_0 is the noise's power spectral density. λ^j_d is a binary variable indicating whether the dth D2D pair shares the RB of the jth network slice. It is defined as follows:

$$\lambda^j_d = \begin{cases} 1, & \text{if } d\text{th D2D pair share the RB of the } j\text{th slice} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

At the receiver, SIC is employed to detect CUs signals sequentially, aiming to improve system performance by progressively decoding and eliminating the strongest interference signals. CUs with better channel conditions can decode the signals of CUs with poorer channel conditions, thereby eliminating inter-user interference from them. In this study, an ideal SIC scenario is assumed throughout the calculations. Consequently, when calculating the signal-to-interference-plus-noise ratio (SINR) of a CU, interference from CUs with lower channel gains can be ignored, and only interference from CUs with higher channel gains needs to be considered. Assuming $\left|h^r_{BS,1}\right|^2 < \left|h^r_{BS,2}\right|^2 < \cdots < \left|h^r_{BS,U_j}\right|^2$, the SINR of $u_{i,j}$ on the rth RB is given by:

$$\gamma^r_{i,j} = \frac{P^r_{i,j}\left|h^r_{BS,u_{i,j}}\right|^2}{\underbrace{\sum_{k=i+1}^{U_j} P^r_{k,j}\left|h^r_{BS,u_{k,j}}\right|^2}_{Interference\ from\ stronger\ CU} + \underbrace{\sum_{d=1}^{D} \lambda^j_d P^r_d \left|h^r_{d_{tr},u_{i,j}}\right|^2}_{Interference\ from\ D2D\ pairs} + BN_0} \tag{4}$$

Based on Equation (4), the data rate of $u_{i,j}$ is given by:

$$R_{i,j} = \sum_{r=1}^{R} \eta^r_j B log_2\left(1 + \gamma^r_{i,j}\right) \tag{5}$$

where η^r_j is a binary variable indicating whether the jth network slice uses the rth RB, defined as follows:

$$\eta^r_j = \begin{cases} 1, & \text{if } j\text{th slice utilizes the } r\text{th RB} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

3.2.2. D2D Signal Model

D2D pairs multiplex the frequency resources of slices in an underlay mode. Since multiple D2D pairs are not permitted to multiplex the same frequency resource simultaneously,

the mutual interference between D2D pairs is neglected. The received signal of the dth D2D pair on the rth RB can be formulated as:

$$y_d^r = \underbrace{h_{d_{tr},d_{re}}^r \sqrt{P_d^r} x_d^r}_{Signal\ from\ D2D\ pairs} + \underbrace{h_{BS,d_{re}}^r \sum_{j=1}^{J} \eta_j^r \sum_{i=1}^{U_j} \sqrt{P_{i,j}^r} x_{i,j}^r}_{Interference\ from\ BS} + n \tag{7}$$

The SINR is given by:

$$\gamma_d^r = \frac{P_d^r \left| h_{d_{tr},d_{re}}^r \right|^2}{\sum_{j=1}^{J} \eta_j^r \sum_{i=1}^{U_j} P_{i,j}^r \left| h_{BS,d_{re}}^r \right|^2 + BN_0} \tag{8}$$

The data rate is:

$$R_d = \sum_{j=1}^{J} \sum_{r=1}^{R} \lambda_d^j B log_2(1 + \gamma_d^r) \tag{9}$$

The total data rate of the entire network is given by:

$$R_{total} = \sum_{j=1}^{J} \sum_{i=1}^{U_j} R_{i,j} + \sum_{d=1}^{D} R_d \tag{10}$$

3.3. Slice Model

In our proposed scheme, the QoS metric for CUs is based on their required transmission rates. The BS acquires the QoS requirement of the uth CU via the uplink, denoted as R_u^{QoS}. Subsequently, the BS determines the number J of QoS levels based on statistical characteristics and assigns CUs to one of the J network slices according to their QoS requirements. To ensure that all QoS requirements are satisfied, the QoS interval length R_{slice}^{length} for each network slice is defined as:

$$R_{slice}^{length} = \frac{R_{max}^{QoS} - R_{min}^{QoS}}{J} \tag{11}$$

where R_{max}^{QoS} represents the maximum QoS requirement among all CUs, while R_{min}^{QoS} represents the minimum QoS requirement among all CUs. The QoS interval length ensures that the QoS for all users is satisfied by setting the QoS requirement of each network slice to cover the maximum requirement of the users within that slice. The rules for network slice are as follows:

$$R_{slice\ j}^{min} = R_{min}^{QoS} + j \cdot R_{slice}^{length}, \forall j \in \mathcal{J} \tag{12}$$

$$\mathcal{U}_j = \left\{ u \in \mathcal{U} \middle| R_{slice\ j-1}^{min} < R_u^{QoS} \leq R_{slice\ j}^{min} \right\}, \forall j \in \mathcal{J} \tag{13}$$

where $R_{slice\ j}^{min}$ denotes minimum QoS constraint of the jth network slice. The QoS intervals of the network slices are configured in ascending order of the slice index, i.e., $R_{slice\ 1}^{min} < R_{slice\ 2}^{min} < \cdots < R_{slice\ J}^{min}$. And in order to properly establish the intervals, the initial boundary is specified as $R_{slice\ 0}^{min} = R_{min}^{QoS}$.

3.4. Problem Formulation

SE is a key metric for measuring the information transmission capability of a unit spectrum resource in a communication system, reflecting the utilization efficiency of spectrum resources. Our objective is to maximize SE by jointly optimizing the power allocation

of CUs, RB assignment of CUs, and admission control for D2D, while simultaneously guaranteeing the QoS requirements of users. The SE-Max problem, denoted as $\boldsymbol{P}$, can be formulated as follows:

$$\mathbf{P}: \max_{\{P_{i,j}^r\},\{\eta_j^r\},\{\lambda_d^j\},} SE = \frac{R_{total}}{R \times B} \tag{14}$$

s.t. C1: $R_{i,j} > R_{slice\ j}^{min},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{U}_j$
C2: $R_d \geq R_0,\ \forall d \in \mathcal{D}$
C3: $\lambda_d^j \in \{0,1\},\ \forall d \in \mathcal{D}, \forall j \in \mathcal{J}$
C4: $\eta_j^r \in \{0,1\},\ \forall j \in \mathcal{J}, \forall r \in \mathcal{R}$
C5: $\sum_{j=1}^{J} \lambda_d^j \leq 1,\ \forall j \in \mathcal{J}$
C6: $\sum_{j=1}^{J} \eta_j^r \leq 1,\ \forall r \in \mathcal{R}$
C7: $\sum_{j=1}^{J} \sum_{r=1}^{R} \eta_j^r \leq R$
C8: $0 < \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} P_{i,j}^r \leq P_{BS}^{max}$
C9: $0 < \sum_{r=1}^{R} P_d^r \leq P_d^{max}, \forall d \in \mathcal{D}$

where constraints C1 and C2 specify the QoS requirements for $u_{i,j}$ and D2D pairs, respectively. Constraints C3 and C4 represent the binary decision variables for RB assignment and D2D admission, respectively. Constraint C5 limits the number of D2D pairs sharing the RBs of the jth slice to at most one. Constraint C6 indicates that each RB can be assigned to at most one slice, ensuring resource orthogonality among slices. Constraint C7 ensures that the total number of RBs assigned to all slices does not exceed the total available RBs in the system. Constraint C8 ensures that the total power allocated to all CUs does not exceed the total power budget P_{BS}^{max} of the BS. Constraint C9 defines the maximum power budget P_d^{max} for each D2D user.

Due to the non-convexity of the objective function and the involvement of mixed variables, i.e., continuous power variables and binary allocation variables, this optimization problem is a mixed-integer non-linear programming (MINLP) problem, which is typically NP-hard. To reduce the computational complexity, we observe that D2D admission control is essentially equivalent to determining whether a D2D link reuses the RBs of a particular slice. Therefore, we unify the D2D admission process with the RB assignment of CUs process, treating them collectively as one RB assignment problem. Consequently, in the following sections, we decompose the original problem into two sub-problems: power allocation and RB assignment. These two problems will be solved through a joint optimization framework.

4. Proposed Joint Optimization and CMAD3QN Method

In this section, we propose a joint optimization and CMAD3QN method to solve problem **P**. Specifically, a CMAD3QN scheme is established for RB assignment. Based on this, a dynamic power allocation strategy is proposed for each given RB assignment to maximize the SE in Equation (14).

4.1. Power Allocation for Given RB Assignment

Before introducing the CMAD3QN-based RB assignment, we first present the power allocation solution proposed for a given RB assignment in this subsection. Since the configuration of D2D links is determined by the CMAD3QN and the transmit power of D2D pairs is fixed, the interference from D2D pairs to CUs can be regarded as part of the

background noise. In this situation, the optimal power allocation problem for CUs can be simplified as follows:

$$\mathbf{P1}: \max_{P} \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} R_{i,j}^{r} = Blog_2\left(1+\gamma_{i,j}^{r}\right) \tag{15}$$

s.t. C1: $R_{i,j} > R_{slice\ j}^{min},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{U}_j$

C2:$0 < \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} P_{i,j}^{r} \le P_{BS}^{max}$

where $P = \left\{P_{i,j}^{r}\right\}_{j\in\mathcal{J}, i\in\mathcal{U}_j, r\in\mathcal{R}}$ denotes the power allocation matrix for all CUs. Due to the existence of co-channel interference terms in the objective function, problem **P1** remains a non-convex optimization problem, making it difficult to obtain the global optimal solution. However, by approximating the lower bound of the non-convex function, we can employ the SCA method to solve it. According to ref. [41], the achievable rate $R_{i,j}^{r}$ of $u_{i,j}$ on the rth RB at the t_sth iteration can be tightly approximated by its lower bound $\tilde{R}_{i,j}^{r}$, expressed as:

$$R_{i,j}^{r} \ge \tilde{R}_{i,j}^{r} = B\left[\alpha_{i,j}^{r} log_2\left(\gamma_{i,j}^{r} + \beta_{i,j}^{r}\right)\right] \tag{16}$$

Here, $\alpha_{i,j}^{r}$ and $\beta_{i,j}^{r}$ are expressed as:

$$\alpha_{i,j}^{r} = \frac{\overset{\sim}{\gamma}_{i,j}^{r}}{1+\overset{\sim}{\gamma}_{i,j}^{r}} \tag{17}$$

$$\beta_{i,j}^{r} = log_2\left(1+\overset{\sim}{\gamma}_{i,j}^{r}\right) - \frac{\overset{\sim}{\gamma}_{i,j}^{r}}{1+\overset{\sim}{\gamma}_{i,j}^{r}} log_2 \overset{\sim}{\gamma}_{i,j}^{r} \tag{18}$$

When $\gamma_{i,j}^{r} = \overset{\sim}{\gamma}_{i,j}^{r}$, the inequality in Equation (16) becomes an equality. In our CU power allocation algorithm, $\overset{\sim}{\gamma}_{i,j}^{r}$ in the current iteration is set to be equal to the value of $\gamma_{i,j}^{r}$ obtained from the previous iteration.

In order to convert $\tilde{R}_{i,j}^{r}$ into a concave function, we further replace the power variable $P_{i,j}^{r}$ with $e^{\tilde{P}_{i,j}^{r}}$. Consequently, problem **P1** can be transformed into:

$$\mathbf{P2}: \max_{\tilde{P}} \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} \tilde{R}_{i,j}^{r} \tag{19}$$

s.t. C1 $0 < \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} e^{\tilde{P}_{i,j}^{r}} \le P_{BS}^{max}$

C2: $\sum_{r=1}^{R} \tilde{R}_{i,j}^{r} \ge R_{slice\ j}^{min},\ \forall j \in \mathcal{J}, \forall i \in \mathcal{U}_j$

Theorem 1. *Problem **P2** is a convex optimization problem.*

Proof of Theorem 1. By expanding $\gamma_{i,j}^{r}$ within $\tilde{R}_{i,j}^{r}$ using Equation (4), we can obtain:

$$\tilde{R}_{i,j}^{r} = B\left\{\frac{\alpha_{i,j}^{r}}{\ln 2}\left[\tilde{P}_{i,j}^{r} - \ln\left(\sum_{k=i+1}^{U_j} e^{\tilde{P}_{k,j}^{r}}\left|h_{BS,u_{k,j}}^{r}\right|^2 + N_{i,j}^{r}\right) + \ln\left|h_{BS,u_{i,j}}^{r}\right|^2\right] + \beta_{i,j}^{r}\right\} \tag{20}$$

where $N_{i,j}^r = \sum_{d=1}^{D} \lambda_d^j P_d^r \left| h_{d_{tr},u_{i,j}}^r \right|^2 + BN_0$ denotes the external interference received by $u_{i,j}$, including co-channel D2D interference and environmental noise. It is worth noting that $\tilde{R}_{i,j}^r$ incorporates log and exp functions related to $\tilde{P}_{i,j}^r$. Since the log-sum-exp function is convex, we conclude that problem **P2** is a standard convex optimization problem.

Consequently, the optimal solution to problem **P2** can be obtained through the Lagrange dual decomposition method. First, we construct the Lagrange function $\mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa})$ as:

$$\mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa}) = \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} (1 + \kappa_{i,j}) \tilde{R}_{i,j}^r - \mu \sum_{j=1}^{J} \sum_{i=1}^{U_j} \sum_{r=1}^{R} e^{\tilde{P}_{i,j}^r} + \mu P_{BS}^{max} - \sum_{j=1}^{J} \sum_{i=1}^{U_j} \kappa_{i,j} R_{slice\ j}^{min} \tag{21}$$

where μ and $\boldsymbol{\kappa} = \{\kappa_{i,j}\}_{j \in \mathcal{J}, i \in \mathcal{U}_j}$ are the Lagrange multipliers associated with constraints (**P2**-C1) and (**P2**-C2), respectively. Then, the dual function can be defined as:

$$g(\mu, \boldsymbol{\kappa}) = \max_{\{\tilde{P}_{i,j}^r\}} \mathcal{L}(\tilde{P}_{i,j}^r, \mu, \boldsymbol{\kappa}) \tag{22}$$

Since problem **P2** is strictly convex and satisfies Slater's condition, the Karush–Kuhn–Tucker (KKT) conditions provide necessary and sufficient conditions for achieving the global optimal solution. Specifically, according to the stability conditions of KKT, the optimal power allocation must ensure that the first-order partial derivative of the Lagrange function with respect to the power variable is equal to zero, i.e., $\frac{\partial \mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa})}{\partial \tilde{P}_{i,j}^r} = 0$. With this equality relationship, we can obtain the solution to Equation (22). When calculating the derivative of $\mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa})$ with respect to $\tilde{P}_{i,j}^r$, it is observed that $\tilde{P}_{i,j}^r$ mainly appears in two parts within the summation of the first term of $\mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa})$. First, it appears in the rate expression $\tilde{R}_{i,j}^r$ of $u_{i,j}$ itself on the rth RB. Second, it appears as interference in the rate expressions $\tilde{R}_{w,j}^r$ of users weaker than $u_{i,j}$ denoted as $u_{w,j}, w \in \{1, 2, \cdots, i-1\}$ within the jth slice on the rth RB. Therefore:

$$\frac{\partial \mathcal{L}(\tilde{P}, \mu, \boldsymbol{\kappa})}{\partial \tilde{P}_{i,j}^r} = \frac{\partial \left[(1 + \kappa_{i,j}) \tilde{R}_{i,j}^r + \sum_{w+1}^{i-1} (1 + \kappa_{w,j}) \tilde{R}_{w,j}^r - \mu e^{\tilde{P}_{i,j}^r} \right]}{\partial \tilde{P}_{i,j}^r} \tag{23}$$

By substituting Equation (20) into Equation (23) and setting Equation (23) to zero, we obtain the analytical iterative solution for $\tilde{P}_{i,j}^r$:

$$\tilde{P}_{i,j}^r = \ln \frac{(1 + \kappa_{i,j}) B \alpha_{i,j}^r}{\ln 2 \cdot \left(\mu + \Psi_{i,j}^r \right)} \tag{24}$$

Here, $\Psi_{i,j}^r$ is the interference penalty factor, expressed as:

$$\Psi_{i,j}^r = \sum_{w=1}^{i-1} \frac{(1 + \kappa_{w,j}) B \alpha_{w,j}^r \left| h_{BS,u_{w,j}}^r \right|^2}{\ln 2 \left(\sum_{k=w+1}^{U_j} e^{\tilde{P}_{k,j}^r} \left| h_{BS,u_{w,j}}^r \right|^2 + N_{w,j}^r \right)} \tag{25}$$

https://doi.org/10.3390/fi18050246

However, it is worth noting that $\Psi_{i,j}^{r}$ also contains $\tilde{P}_{i,j}^{r}$. To achieve convergence, we can obtain the result of $\tilde{P}_{i,j}^{r}$ through the following iterative calculation:

$$\tilde{P}_{i,j}^{r(t_p+1)} = \ln \frac{\left(1+\kappa_{i,j}\right) B\alpha_{i,j}^{r}}{\ln 2 \cdot \left(\mu + \Psi_{i,j}^{r(t_p)}\right)} \tag{26}$$

where the superscript t_p denotes the index of the power iteration. Then, we use $\tilde{P}_{i,j}^{r(t_p)}$ from the previous iteration to calculate $\tilde{P}_{i,j}^{r(t_p+1)}$ in the current iteration.

To determine μ and $\boldsymbol{\kappa}$, the dual problem can be formulated as:

$$\mathbf{P3}: \max_{\mu,\boldsymbol{\kappa}} g(\mu,\boldsymbol{\kappa}) \; s.t. \; \mu,\boldsymbol{\kappa} > 0 \tag{27}$$

Since both the objective function and the constraints of problem **P2** are convex, the solution to **P2** is equivalent to the solution of the dual problem **P3**. In the following, we describe the search method for determining the optimal solution to the dual problem using the standard sub-gradient method, where the dual variables μ and $\boldsymbol{\kappa}$ can be updated iteratively as:

$$\mu^{(t_d+1)} = \left[\mu^{(t_d)} - \delta^{(t_d)} \left(P_{BS}^{max} - \sum_{j=1}^{J}\sum_{i=1}^{U_j}\sum_{r=1}^{R} e^{\tilde{P}_{i,j}^{r}}\right)\right]^{+} \tag{28}$$

$$\boldsymbol{\kappa}_{i,j}^{(t_d+1)} = \left[\boldsymbol{\kappa}^{(t_d)} - \delta^{(t_d)} \left(\sum_{r=1}^{R} \tilde{R}_{i,j}^{r} - R_{slice\ j}^{min}\right)\right]^{+} \tag{29}$$

where the superscript t_d denotes the iteration index of the dual problem, $\delta^{(t_d)}$ represents the step size, and $[x]^{+} = \max(0, x)$. If an appropriate step size $\delta^{(t_d)}$ is selected such that $\delta^{(t_d)} \to 0$ as $t_d \to \infty$, this sub-gradient method guarantees convergence. Finally, the optimal transmit power for each CU is obtained by substituting back the variable $P_{i,j}^{r} = e^{\tilde{P}_{i,j}^{r}}$. □

Based on the above analysis, the optimal power allocation for CUs is achieved by decomposing the non-convex optimization problem into three nested loops. Algorithm 1 summarizes this process.

The outer loop is based on the SCA algorithm. First, the power allocation matrix **P** is randomly initialized. In each iteration, the coefficients $\alpha_{i,j}^{r}$ and $\beta_{i,j}^{r}$ are calculated using $\gamma_{i,j}^{r}$ from the previous iteration. Subsequently, the power allocation **P** is updated via Lagrange dual decomposition method, and $\gamma_{i,j}^{r}$ is recalculated based on the new **P**. The loop terminates when the gap falls below a predefined threshold ϵ_s or when the predefined maximum number of iterations is reached. The middle loop employs the Lagrange dual decomposition method to solve the convex problem. It updates the Lagrange multipliers μ and $\boldsymbol{\kappa}$ via the sub-gradient method to satisfy the global power and QoS constraints. The inner loop iteratively calculates the optimal power allocation using the analytical iterative solution while keeping the Lagrange multipliers fixed. When the power difference between two adjacent steps is less than the power tolerance ϵ_p, the process will stop.

Algorithm 1. CUs Power Allocation Algorithm

Require: Maximum number of iterations T_s and threshold ϵ_s for SCA; Step size δ; Tolerances ϵ_d, ϵ_p.
Ensure: Optimal power allocation matrix $\mathbf{P}^{(*)}$.

Initial power allocation $\mathbf{P}^{(0)}$ and calculate the corresponding $\gamma_{i,j}^{r(0)}$;
Set SCA iteration index $t_s = 0$;
repeat
  Set variable $\overset{\sim}{\gamma}_{i,j}^{r(t_s+1)} = \gamma_{i,j}^{r(t_s)}$;
  Compute $\alpha_{i,j}^{r}$ and $\beta_{i,j}^{r}$ from (17) and (18);
  Set dual iteration index $t_d = 0$;
  Initialize Lagrange multipliers μ^0, $\boldsymbol{\kappa}^0$;
  repeat
    Set power iteration index $t_p = 0$;
    repeat
      Update $\tilde{P}_{i,j}^{r(t_p+1)}$ from (26);
      $t_p = t_p + 1$;
    until $\left|\tilde{P}^{(t_p)} - \tilde{P}^{(t_p-1)}\right| \leq \epsilon_p$
    Update $\mu^{(t_d+1)}$ and $\boldsymbol{\kappa}^{(t_d+1)}$ from (28) and (29);
    $t_d = t_d + 1$;
  until $\left(\left|\mu^{(t_d+1)} - \mu^{(t_d)}\right| \leq \epsilon_d \text{ and } \left|\boldsymbol{\kappa}^{(t_d+1)} - \boldsymbol{\kappa}^{(t_d)}\right| \leq \epsilon_d\right)$
  $\mathbf{P}^{(t_s+1)} = exp\,(\tilde{P}^{(\text{optomal\_from\_dual})})$;
  Calculate $\gamma_{i,j}^{r(t_s+1)}$ from (4) with $\mathrm{P}^{(t_s+1)}$;
  $t_s = t_s + 1$;
until $\left(\left|\gamma_{sum}^{(t_s)} - \gamma_{sum}^{(t_s-1)}\right| \leq \epsilon_s \text{ or } t_s > T_s\right)$
return $\mathrm{P}^{(*)} = \mathrm{P}^{(t_s)}$.

4.2. CMAD3QN-Based RB Assignment Strategy

After resolving the continuous power allocation, this subsection focuses on optimizing the remaining discrete variables within the MINLP problem, specifically the RB assignment. Here, the RB assignment for CUs $\{\eta_j^r\}$ and the D2D admission control $\{\lambda_d^j\}$ constitute high-dimensional discrete variables. As the number of users increases, the search space expands exponentially, making it difficult for traditional optimization methods to achieve optimal solutions through iterative search within the coherence time. To address this challenge, this section proposes a resource allocation mechanism based on CMAD3QN. We model the RB assignment process as a Markov decision process (MDP) and leverage the unique dueling architecture and double Q-learning mechanism of D3QN to intelligently determine the optimal RB assignment in a complex interference environment.

4.2.1. MADRL Model

In the system under consideration, each network slice is regarded as an intelligent agent, thereby transforming the RB assignment problem into a distributed multi-agent problem. Its MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{F}, \sigma)$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ represents the finite action space, $\mathcal{P}$ represents the state transition probabilities, $\mathcal{F}$ represents the reward function, and $\sigma \in [0, 1]$ is the discount factor for future rewards. The elements of the MDP model tuple are defined as follows:

State: The state $s_j(t)$ of agent j at time slot t is composed of the channel gains between the current base station and all CUs in the jth network slice $\boldsymbol{h}_{BS,\mathcal{U}_j}(t)$, the channel gains between the base station and all D2D receivers $\boldsymbol{h}_{BS,d_{re}}(t)$, the channel gains between D2D pairs and all CUs in the jth network slice $\boldsymbol{h}_{d_{tr},\mathcal{U}_j}(t)$, and the RB assignment in the previous time slot $\boldsymbol{\eta}_j(t-1)$. The state space of the jth agent in time slot t is defined as:

$$s_j(t) = \left\{ \boldsymbol{h}_{BS,\mathcal{U}_j}(t), \boldsymbol{h}_{BS,d_{re}}(t), \boldsymbol{h}_{d_{tr},\mathcal{U}_j}(t), \boldsymbol{\eta}_j(t-1) \right\} \tag{30}$$

where $\boldsymbol{h}_{BS,\mathcal{U}_j}(t) = \left\{ h^r_{BS,u_{i,j}}(t) \right\}_{i\in\mathcal{U}_j, r\in\mathcal{R}}$, $\boldsymbol{h}_{BS,d_{re}}(t) = \left\{ h^r_{BS,d_{re}}(t) \right\}_{d\in\mathcal{D}, r\in\mathcal{R}}$, $\boldsymbol{h}_{d_{tr},\mathcal{U}_j}(t) = \left\{ h^r_{d_{tr},u_{i,j}} \right\}_{i\in\mathcal{U}_j, r\in\mathcal{R}, d\in\mathcal{D}}$, and $\boldsymbol{\eta}_j(t-1) = \left\{ \eta^1_j(t-1), \eta^2_j(t-1), \cdots, \eta^R_j(t-1) \right\}$ is a vector of dimension $R \times 1$.

Action: We define the action executed by agent j at time slot t as $a_j(t)$. The action $a_j(t)$ consists of two components: selecting an RB and determining which D2D pairs are permitted to reuse the RB of slice j. The action space of the jth agent in time slot t is defined as:

$$a_j(t) \in \mathcal{A}_j = \left\{ \left(\eta^1_j, \eta^2_j, \cdots, \eta^R_j \right) \times \left(\lambda^j_1, \lambda^j_2, \cdots, \lambda^j_D \right) \right\} \tag{31}$$

It can be observed that the size of the action space for agent j is $|\mathcal{A}_j| = R \times D$.

Reward: The reward function in the MADRL framework controls the training process, which can adopt either a centralized or a decentralized reward structure. Using a decentralized reward structure may lead to selfish behaviors among agents, where they compete with each other to maximize their own rewards, thereby degrading the global performance. The objective of this paper is to optimize the network SE while satisfying the QoS requirements of all users. Therefore, we adopt a centralized reward structure to provide a common reward for all agents. Specifically, we use the SE to define the common reward for all agents. Thus, the reward function in time slot t, denoted by $f(t)$, is defined as:

$$f(t) = \begin{cases} SE, & \text{if constrains satisfied} \\ f_{neg}, & \text{otherwise} \end{cases} \tag{32}$$

The equation indicates that when the agent satisfies all constraints, the reward is positive, i.e., equal to the SE; otherwise, it is negative, where f_{neg} is a large constant penalty term. To achieve favorable long-term performance, it is necessary to consider both immediate rewards and future rewards. Therefore, the primary goal of reinforcement learning is to find a policy that maximizes the long-term discounted reward. The long-term discounted reward is defined as:

$$F(t) = f(t) + \sigma f(t+1) + \sigma^2 f(t+2) + \cdots = \sum_{n=0}^{\infty} \sigma^n f(t+n) \tag{33}$$

Obviously, by maximizing the long-term discounted reward, an effective policy for maximizing the network SE can be obtained. In this way, the MDP transforms the original RB assignment problem into a reward maximization problem.

4.2.2. CMAD3QN Algorithm

To solve this transformed problem, we propose the CMAD3QN algorithm to address the RB assignment problem. The detailed flowchart of the algorithm is illustrated in Figure 2. Specifically, at each time slot t, each agent observes the current state $s_j(t)$ and selects an action $a_j(t)$ from the action space according to a specific policy π. Upon executing the action, the agent receives a reward from the environment and then transitions to a new

state. Subsequently, the DRL framework updates its action selection policy π to maximize the cumulative discounted reward derived from the interactions.

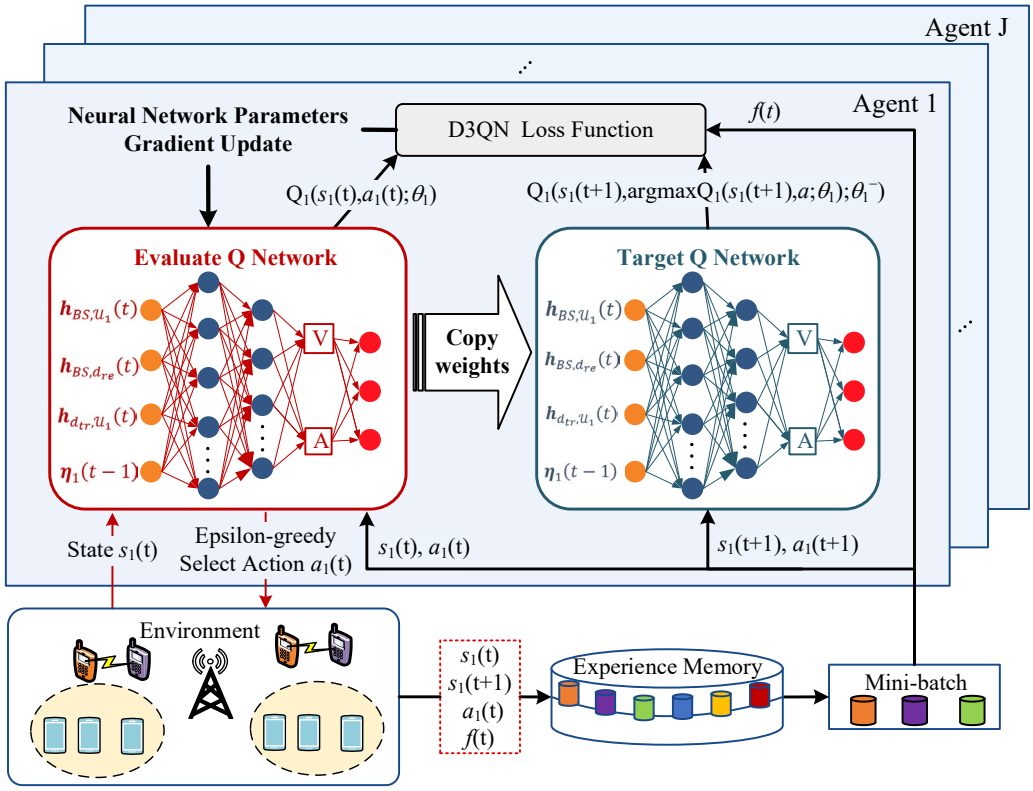

Figure 2. Flowchart of CMAD3QN algorithm.

Q-learning algorithm is a Q-value-based RL algorithm that can be used to obtain the optimal policy π^*. Under policy π, the Q-value $Q_j(s_j(t), a_j(t))$ for a given "state-action" pair $s_j(t), a_j(t)$ is defined as the expectation of the expected long-term discounted reward obtained when the agent slice j executes action $a_j(t)$ and subsequently follows policy π. That is:

$$Q_j(s_j(t), a_j(t)) = \mathbb{E}_\pi(F(t)|s_j(t), a_j(t)) \tag{34}$$

Conversely, once the value $Q_j(s_j(t), a_j(t))$ corresponding to each "state-action" pair is known, the optimal policy π^* can be derived by selecting the action from $\mathcal{A}_j$ that maximizes $Q_j(s_j(t), a_j(t))$ for a given state $s_j(t)$.

$$\pi^* \leftarrow \underset{a_j(t)}{\text{argmax}}\, Q_j(s_j(t), a_j(t)) \tag{35}$$

According to the Bellman criterion [42], the optimal Q-value is estimated as:

$$Q_j^*(s_j(t), a_j(t)) = f(t) + \sigma \max_{a_j(t+1)} Q_j^*(s_j(t+1), a_j(t+1)) \tag{36}$$

According to the Bellman equation in Equation (36), the Q-value can be iteratively updated to approximate the optimal Q-value. Its expression is:

$$Q_j^{new}(s_j(t), a_j(t)) = Q_j^{old}(s_j(t), a_j(t)) + \chi\left[f(t) + \sigma \max_{a_j(t+1)} Q_j(s_j(t+1), a_j(t+1)) - Q_j^{old}(s_j(t), a_j(t))\right] \tag{37}$$

where $\chi \in [0, 1]$ is the learning rate, $f(t)$ is the immediate reward obtained after executing action a (denoted as $a_j(t)$ in the context), $s_j(t+1)$ is the new state observed after executing

action $a_j(t)$, and $a_j(t+1)$ is the next action selected in the new state $s_j(t+1)$. With an appropriate learning rate, $Q_j\left(s_j(t), a_j(t)\right)$ will converge to $Q_j^*\left(s_j(t), a_j(t)\right)$ with probability 1, thereby allowing the optimal policy π^* to be found.

However, in the cellular network system investigated in this paper, due to the dynamic characteristics of the channels, it is impractical to determine π^* iteratively within an exploding state space. Deep neural networks (DNNs) are capable of updating θ_j based on extensive training data, thereby handling the complex mapping relationship between channel information and expected outputs. This mapping is then utilized to estimate the corresponding $Q_j\left(s_j(t), a_j(t)\right)$ for various "state-action" combinations. Therefore, we employ DNNs with interconnected weighted neurons, denoted as $Q_j\left(s_j(t), a_j(t); \theta_j\right)$, to approximate the Q-value of agent j, where θ_j represents the weights of the neural network.

The training and testing data of DNNs are generated through the interaction between the environment simulator and the agents. Each training sample used to optimize the Deep Q-Network (DQN) comprises the state $s_j(t)$, the next state $s_j(t+1)$, the action $a_j(t)$, and the reward $f(t)$. To eliminate the temporal correlation of the generated data, we employ the Deep Q-learning algorithm combined with experience replay. The generated training data is stored in a memory buffer $\mathcal{O}_j$. In each iteration, the DQN algorithm samples a mini-batch $\hat{\mathcal{O}}_j$ from the experience memory as the dataset, and trains the DQN via backpropagation and updates the values of θ_j to reduce the loss function, which is formulated as:

$$L\left(\theta_j\right) = \sum_{\left(s_j(t), a_j(t)\right) \in \hat{O}_j} \left(y_i(t) - Q_j^{eval}\left(s_j(t), a_j(t); \theta_j\right)\right)^2 \tag{38}$$

Here

$$y_i(t) = f(t) + \chi Q_j^{tar}\left(s_j(t+1), \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q_j^{eval}\left(s_j(t+1), a; \theta_j^-\right); \theta_j^-\right) \tag{39}$$

Here, $Q_j^{eval}\left(s_j(t), a_j(t); \theta_j\right)$ is generated by the Q-evaluation network, and $Q_j^{tar}\left(s_j(t), a_j(t); \theta_j^-\right)$ is generated by the Q-target network. Both networks employ the same network architecture. The updated parameters θ_j of the Q-evaluation network are copied to the Q-target network parameters θ_j^- every W time slots. During this interval, the target Q-value remains constant, which to some extent reduces the correlation between the current Q-value and the target Q-value, thereby improving the algorithm's stability. In the D3QN model, action selection (e.g., $\underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q_j^{eval}\left(s_j(t+1), a; \theta_j\right)$) and action evaluation (e.g., $Q_j^{tar}\left(s_j(t+1), \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q_j^{eval}\left(s_j(t+1), a; \theta_j^-\right); \theta_j^-\right)$) are decoupled by using two distinct Q-value functions, thereby avoiding the overestimation inherent in traditional methods.

Furthermore, since a large volume of samples must be utilized during the training process, estimating the value for every action selection in a specific state would reduce the learning efficiency of the network. Therefore, we adopt the dueling architecture to evaluate states, decomposing the output Q-value into two components: the value function and the advantage function, namely:

$$Q_j\left(s_j(t), a_j(t); \theta_j\right) = V_j\left(s_j(t); \theta_j\right) + A_j\left(s_j(t), a_j(t); \theta_j\right) \tag{40}$$

Here, $V_j\left(s_j(t); \theta_j\right)$ estimates the expected value of the state, and $A_j\left(s_j(t), a_j(t); \theta_j\right)$ estimates the advantage value of each action minus the average of all action advantages in that state.

To achieve a balance between exploration and exploitation behaviors at different stages of training, we employ the ε-greedy policy. Specifically, at time slot t, given the current state $s_j(t)$, the action executed by agent j is defined as:

$$a_j(t) = \begin{cases} \text{random action } a_j(t), & \text{with probability } \varepsilon \\ \underset{a_j(t)\in\mathcal{A}_j}{\text{argmax}}\, Q_j\big(s_j(t), a_j(t)\big), & \text{with probability } 1-\varepsilon \end{cases} \tag{41}$$

where $\varepsilon \in [0, 1]$ represents the exploration rate. The agent selects the action with the highest current Q-value with probability $1 - \varepsilon$, and randomly selects an action with probability ε. This strategy allows the agent to exploit the known optimal actions most of the time, while retaining a certain probability of exploring new actions to discover potentially superior policies.

4.3. Joint Optimization and CMAD3QN Algorithm

Based on the preceding discussions, we propose a JOCMAD3QN method to solve problem **P**. Specifically, the proposed method is a combination of the power allocation for a given RB assignment presented in Section 4.1 and the RB assignment algorithm based on CMAD3QN presented in Section 4.2. Once the RB assignment is determined in the CMAD3QN algorithm, Algorithm 1 is invoked to perform power allocation. This approach effectively addresses the limitation of the D3QN algorithm, which is not suitable for continuous action spaces.

The overall procedure is outlined in Algorithm 2. After initializing the network environment, each agent j observes the initial state space. At each time slot t, each agent j selects an action using the ε-greedy strategy. Subsequently, the discrete RB assignment action information from all agents is input into Algorithm 1 as known parameters. Algorithm 1 then iteratively calculates the current optimal continuous power allocation matrix $\boldsymbol{P}^{(*)}$ for all CUs. Based on the determined RB assignment and the optimal power allocation, the environment calculates the reward using Equation (32) and feeds it back to the agents. The agent stores the generated experience tuple $\big(s_j(t), a_j(t), f(t), s_j(t+1)\big)$ into the experience replay buffer $\mathcal{O}_j$. When the samples in the replay buffer are sufficient, a mini-batch of data is randomly sampled to train the D3QN network.

4.4. Complexity Analysis of the Proposed Method

The time complexity of JOCMAD3QN consists of two components: the power allocation and the CMAD3QN network training. Assuming the maximum number of iterations for the SCA outer loop, the Lagrange dual middle loop, and the power update inner loop are T_s, T_d, and T_p, respectively, the maximum complexity of Algorithm 1 is $O\big(T_s \cdot T_d \cdot T_p \cdot U \cdot R\big)$. In the CMAD3QN algorithm, the state and action spaces of a single agent are significantly reduced due to the adoption of the multi-agent distributed architecture. Because of the double Q-network and dueling architecture, both the evaluation DNN and the target DNN are required for computation. Consequently, its time complexity per iteration is slightly higher than that of the basic DQN, given by $O((2N + M) \cdot K)$, where N denotes the computational load of forward propagation, M denotes the computational load of backward propagation, and K is the mini-batch size sampled for training. The JOCMAD3QN invokes Algorithm 1 after each RB assignment decision, the total complexity per step is $O\big((2N + M) \cdot K + T_s \cdot T_d \cdot T_p \cdot U \cdot R\big)$.

Algorithm 2. JOCMAD3QN Algorithm for Maximizing SE

Initialize the weight matrices of the online and target networks, i.e., θ_j and θ_j^-.
for episode $e = 1, \cdots, E_{max}$ **do**
  Initialize environment topology, channel gains, and observe initial state $s_j(1)$;
  for step $t = 1, \cdots, T$ **do**
    All agents select action $a_j(t)$ following the ε-greedy policy in (41);
    Run Algorithm 1 to determine the optimal power allocation matrix $\boldsymbol{P}^{(*)}$;
    The BS broadcasts the resource allocation;
    All agents observe reward $f(t)$ in (32) and next state $s_j(t+1)$;
    for agent $j = 1, \cdots, J$ **do**
      Store experience tuple $\left(s_j(t), a_j(t), f(t), s_j(t+1)\right)$ into $\mathcal{O}_j$;
      if store enough experience **then**
        Randomly sample a minibatch $\hat{\mathcal{O}}_j$ from $\mathcal{O}_j$;
        Calculate target $y_i(t)$ via (39);
        Calculate loss value $L\left(\theta_j\right)$ via (38);
        Update θ_j using gradient descent;
        Copy θ_j to θ_j^- every W steps;
      end if
    end for
    Update state $s_j(t) \leftarrow s_j(t+1)$;
  end for
  Update exploration rate ε;
end for

Regarding the single-step execution time, the execution time per step of JOCMAD3QN is slightly higher than that of pure learning algorithms due to the introduction of precise iterative calculations in Algorithm 1. However, this hybrid architecture decouples the complex resource allocation problem and isolates the continuous power variables, which effectively prevents inefficient trial-and-error explorations by multiple agents within a massive action space.

5. Experimental Results and Discussion

In this section, we evaluate the performance of our proposed algorithm from multiple dimensions. Before delving into the simulation results, we first describe the parameter settings.

5.1. Numerical Settings

The default experimental configuration consists of 1 BS, 3 network slices, 3 CUs per slice and 3 D2D pairs. The CUs and D2D pairs are randomly distributed within a range of 30 to 300 m from the BS. The D3QN training model is built upon a fully connected neural network. The network adopts a dueling architecture. First, feature extraction is performed through three shared hidden layers containing 512, 512, and 256 neurons, respectively. Subsequently, the signal splits into two independent streams: a state-value stream containing 128 neurons and an action-advantage stream containing 256 neurons. The model utilizes ReLU as the activation function and Adam as the optimizer. The simulation was implemented using torch 2.8.0 on Python 3.11.14. Detailed information regarding other parameters is presented in Table 2.

Table 2. Simulation Parameters.

Parameters	Value
Cellular coverage radius	300 m
Maximum BS transmission power P_{BS}^{max}	46 dBm
D2D pairs transmission power P_{d}^{max}	15 dBm
Minimum communication distance for D2D pairs	30 m
Carrier frequency	2 GHz
Total network communication bandwidth	5 MHz
Number of RBs R	6
Noise's power spectral density N_0	-174 dBm/Hz
Standard deviation of lognormal shadowing ξ	8 dB
CU link path loss model	128.1 + 37.6log (l/l_0)
D2D link path loss model	148 + 40 log (l/l_0)
QoS constraint for each slice $R_{slice\ j}^{min}$	1 Mbps, 2 Mbps, 3 Mbps
QoS constraint for D2D pairs R_0	5 Mbps
Maximum number of iterations t_s and threshold ϵ_s	100, 0.0001
Toleration of dual Lagrange and power method ϵ_d, ϵ_p.	0.01
Episodes	3000
Learning rate χ	0.0005
Mini-batch size	256
Discount factor σ	0.95
Exploration rate ϵ	0.05–1
Memory buffer size $\mathcal{O}_j$	100,000
Target network update frequency W	1000

5.2. Result Analysis

To evaluate the performance of the proposed scheme, we compare it with the following five benchmark algorithms:

1. JOCMADQN: This algorithm modifies the JOCMAD3QN framework by removing the dueling architecture, employing a standard double DQN. It mitigates the overestimation bias in Q-value estimation by introducing a target network.
2. Centralized D3QN (CD3QN) [43]: This approach aggregates the state space and action space of all agents. A global super-agent BS makes decisions centrally and outputs the RB assignment vector for all users. Internally, it still invokes Algorithm 1 to achieve optimal power allocation.
3. Full MAD3QN (FMAD3QN) [44]: This method employs the standard D3QN framework to perform both power allocation and RB assignment simultaneously. It discretizes the continuous power allocation into three levels. Each slice acts as an agent to select the power index for its three CUs, while also determining sub-channel selection and D2D access control.
4. Fixed PA (FPA): In this scheme, RB assignment is performed by the proposed CMAD3QN agents, whereas power allocation adopts the fixed fractional transmit power control (FTPC) scheme. Specifically, users reusing the same RB are ranked based on their channel gains. Users with poorer channel conditions are assigned higher fixed power coefficients to ensure successful decoding via SIC.
5. Random Method: In this scheme, each user randomly selects RB and power values without any learning process.

Given that hyperparameters have a significant impact on the learning process of DRL algorithms, we first evaluate the influence of different learning rates $\chi \in \{0.00001,\ 0.0005,\ 0.005,\ 0.05\}$ on the convergence performance of the JOCMAD3QN algorithm in terms of rewards, as shown in Figure 3a. Figure 3a shows that a large learning rate, e.g., $\chi = 0.05$, prevents the model from converging to the optimal solution and

causes the reward curve to exhibit significant fluctuations. Conversely, when the learning rate is too small, e.g., $\chi = 0.00001$, although the training process is relatively stable, the convergence speed is extremely slow, and the performance has not yet reached its saturation point even after 2000 rounds. Experimental results indicate that setting $\chi = 0.0005$ achieves the optimal balance between training efficiency and policy stability; thus, it is selected as the default parameter for subsequent experiments. On this basis, Figure 3b further compares the convergence of the system SE between the proposed algorithm and the benchmark CD3QN algorithm under $\chi = 0.0005$ and $\chi = 0.05$. The results show that the under $\chi = 0.0005$, the algorithm stabilizes after approximately 1000 iterations, whereas the CD3QN requires over 3000 iterations. Moreover, the proposed algorithm eventually converges to a higher steady-state SE, verifying its efficiency and superiority in solving high-dimensional complex resource allocation problems. Under χ = 0.05, the SE of the CD3QN system experiences a cliff-like collapse and continues to fluctuate in a low trough. In contrast, the proposed algorithm still successfully converges. This strongly proves that, compared to the benchmark, the proposed model exhibits significantly lower sensitivity to hyperparameter perturbations.

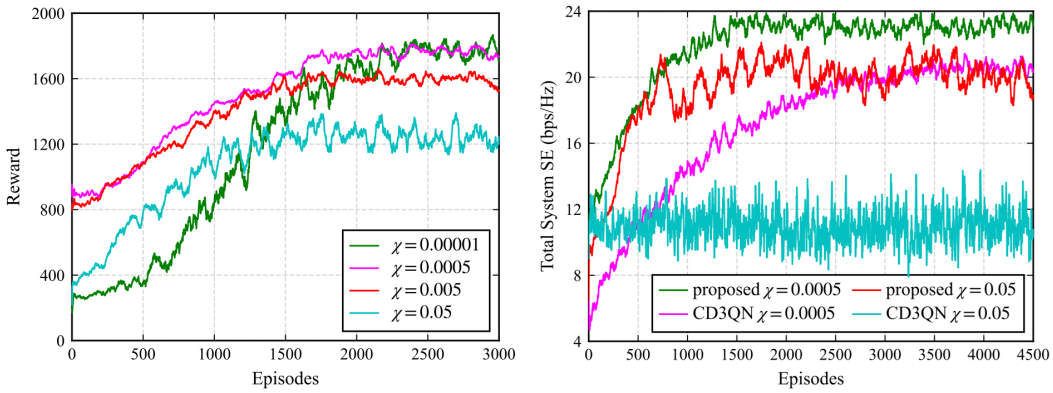

Figure 3. Convergence performance of the proposed algorithm (**a**) with different learning rate values. (**b**) compared with CD3QN.

Figure 4a provides the SE performance validation of the JOCMAD3QN algorithm under different network scales. In this context, the number of slices (i.e., agents) is set to $J = \{2, 3, 4, 5\}$, the number of D2D pairs matches the number of slices, and the number of RBs is $R = \{2, 6, 12, 20\}$, corresponding to RB density values of $RD = \frac{R}{J} = \{1, 2, 3, 4\}$. Figure 4a shows that as RD increases, the total system SE improves significantly. This is because a larger RD implies that R is much greater than J, thereby reducing the interference among users sharing the same RB. Although the state and action spaces grow exponentially as J and R increase, the JOCMAD3QN algorithm maintains favorable convergence characteristics. Even in the largest configuration (i.e., $J = 5$, $R = 20$), the algorithm reaches convergence around 2000 rounds. This is primarily attributed to the multi-agent distributed architecture adopted in this paper, where each slice makes decisions as an independent agent. This approach effectively mitigates the curse of dimensionality problem faced by single agents in large-scale networks. Figure 4b further evaluates the algorithm's performance under various numbers of CUs per slice ($U_j \in \{2, 3, 4, 5\}$). As U_j increases, the system total SE steadily improves, demonstrating the algorithm's ability to effectively exploit multi-user diversity. Despite the intensified co-channel interference and the heightened difficulty of satisfying strict QoS constraints in such dense scenarios, JOCMAD3QN consistently maintains robust convergence. This further validates the effectiveness of the proposed hybrid architecture in high-load environments.

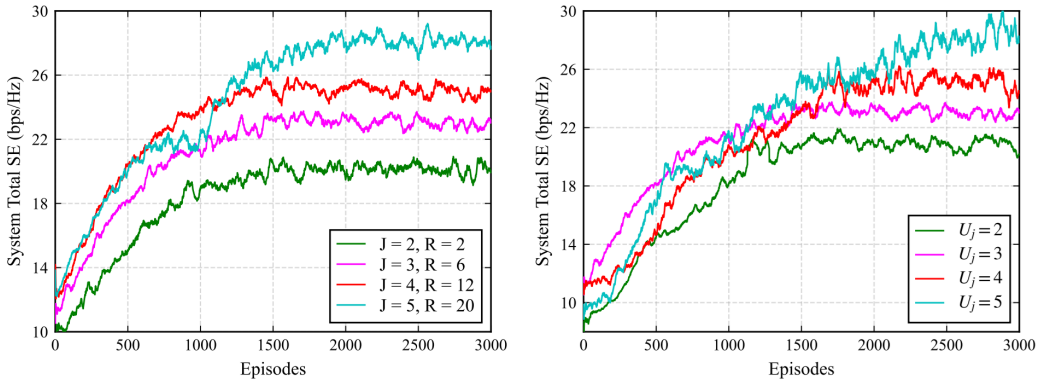

Figure 4. JOCMAD3QN performance under different (**a**) network scales (**b**) number of CUs per slice.

Figure 5 depicts the impact of the maximum BS transmit power P_{BS}^{max} and the D2D transmit power P_d^r on the SE performance achieved by different methods. As observed in Figure 5a, with the increase of P_{BS}^{max}, the total system SE exhibits a significant upward trend, benefiting from the improvement in the downlink SNR. However, this improvement comes at a cost. As shown in Figure 5c,e, although the SE of CUs increases substantially with increasing power, the increase in BS transmit power causes severe co-channel interference at the D2D receivers. This is because the underlay D2D users reuse cellular resources, leading to a gradual decline in the SE of D2D users. Despite this physical constraint, the proposed algorithm effectively balances the conflict between increasing CU rates and suppressing D2D interference by leveraging SCA-based continuous power control and multi-agent cooperative resource allocation. Simulation results demonstrate that the total SE of the proposed scheme consistently outperforms that of the benchmark schemes across the entire power range. Notably, the proposed scheme has the smallest gap compared to the JOCMADQN across all power values. This validates the advantage of the dueling network architecture in value evaluation, as it can more accurately identify differences in state values caused by changes in channel conditions, thereby making more robust resource allocation decisions. In contrast, the CD3QN struggles to converge to the global optimum due to the excessively large state space.

Figure 5b shows that the system's total SE exhibits a steady upward trend as P_d^r increases. This gain primarily stems from the significant improvement in D2D link performance shown in Figure 5e, where higher transmit power directly improves the SNR at the D2D receiver. However, this improvement comes at the expense of CU performance. As illustrated in Figure 5d, since D2D users share the spectrum in underlay mode, as P_d^r increases, the co-channel interference generated at the CU receivers intensifies significantly, leading to a decline in the SE of CUs across all schemes. Nevertheless, compared with other benchmark algorithms, the proposed algorithm demonstrates superior interference management capabilities. Figure 5d shows that even at high D2D power levels (e.g., 35 dBm), the CU spectral efficiency of the proposed scheme remains approximately 2bps/Hz higher than that of FMAD3QN. This indicates that the JOCMAD3QN algorithm can intelligently allocate high-power D2D pairs to RBs that cause minimal interference to CUs.

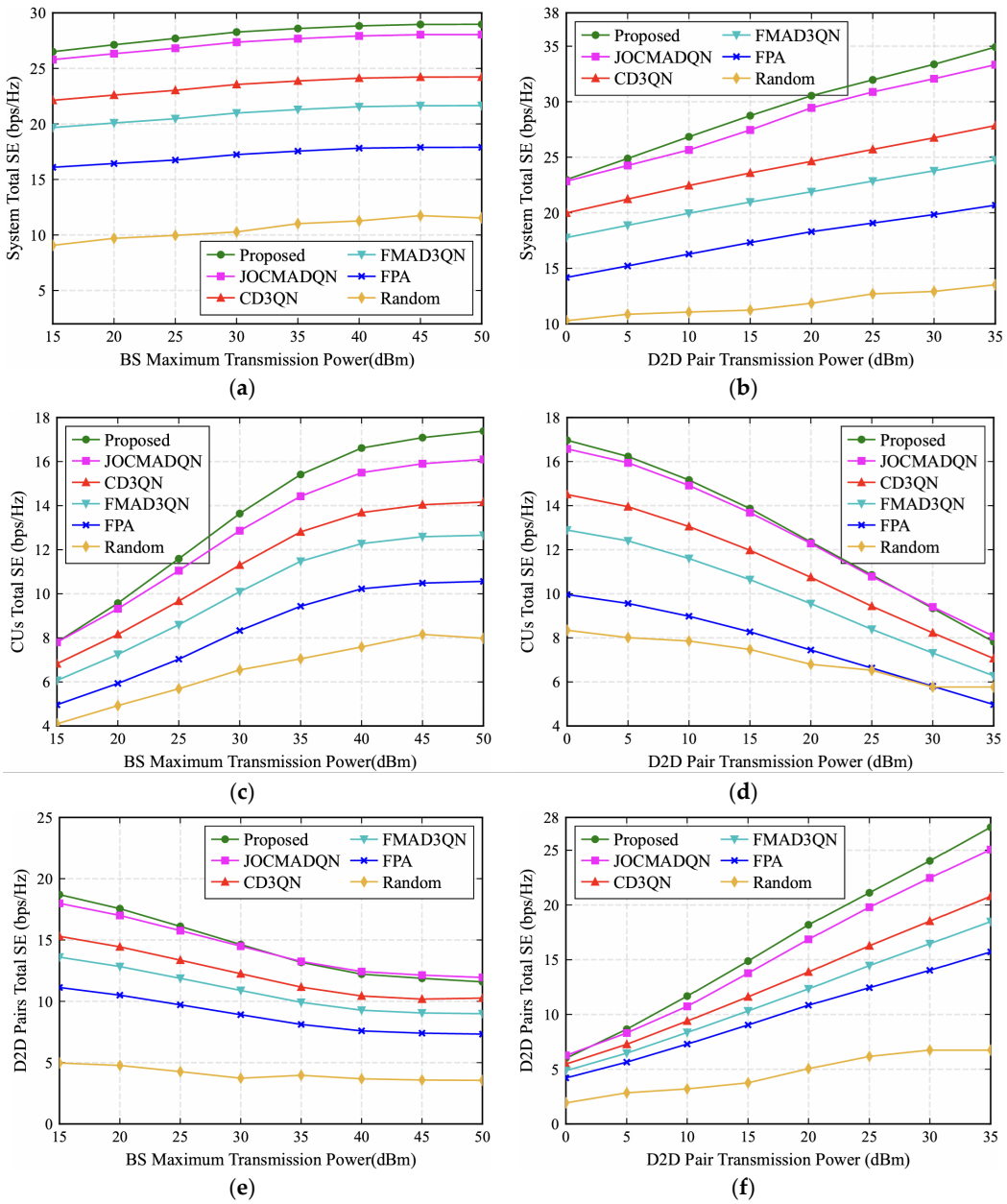

Figure 5. Comparative analysis. (**a**) System total SE versus BS maximum transmission power. (**b**) System total SE versus D2D pairs transmission power. (**c**) CU total SE versus BS maximum transmission power. (**d**) CU total SE versus D2D pairs transmission power. (**e**) D2D pairs total SE versus BS maximum transmission power. (**f**) D2D pairs total SE versus D2D pairs transmission power.

Figure 6 reveals the impact of different network slice QoS requirements on the system's total SE. The scaling factor τ is used to adjust the QoS requirements of each slice via multiplication. Specifically, the QoS requirements for each slice are defined as $\tau \times [1\text{ Mbps}, 2\text{ Mbps}, 3\text{ Mbps}]$. It can be observed from Figure 6 that as the scaling factor increases from 0.5 to 2.5, the system's total SE of all optimization algorithms exhibits a downward trend. This is a manifestation of the typical "efficiency–fairness" trade-off in resource allocation: as QoS constraints tighten, the BS is forced to allocate more power resources to users with poorer channel conditions to meet their minimum QoS requirements. Nevertheless, the proposed algorithm still demonstrates significant superiority across the entire tested range. Particularly in the high-load scenario where the scaling factor is 2.5, the SE of the proposed scheme remains above 26 bps/Hz, significantly outperforming the FMAD3QN and FPA schemes. This demonstrates that the JOCMAD3QN algorithm

can maintain system performance robustness under strict QoS constraints through precise power control and intelligent interference management.

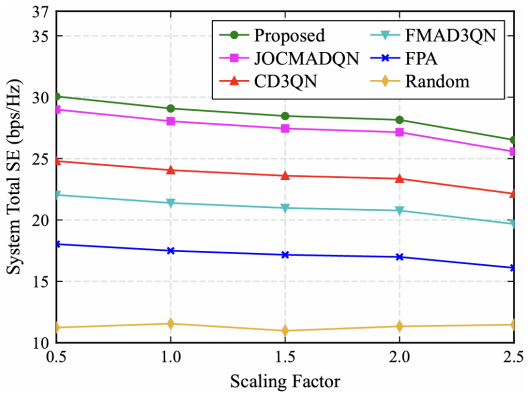

Figure 6. The SE performance versus slice QoS scaling factor.

6. Conclusions

In this study, we investigated the resource allocation for multi-slice cellular networks with underlay D2D communications based on NOMA, ensuring differentiated QoS requirements for users in different slices. We established a non-convex optimization problem to maximize the total system SE under constraints including QoS requirements, power budgets for CUs and D2D users, and RBs assignment. Given the complexity of MINLP inherent in this problem, we designed a JOCMAD3QN resource allocation mechanism. This mechanism breaks the limitations of traditional single optimization methods or pure learning approaches. By constructing a hierarchical optimization framework, we first derived the analytical iterative solution for optimal power allocation using SCA and the Lagrange dual method, effectively avoiding quantization errors caused by action space discretization. Building on this, we further designed a CMAD3QN, which efficiently solved the discrete RBs assignment problems through distributed agent interactions. Simulation results demonstrate that, compared to benchmark schemes such as CD3QN, FMAD3QN, and FPA, the proposed algorithm significantly reduces the state space dimension through a distributed architecture. It achieves faster convergence speeds and consistently maintains superior spectral efficiency and robust interference management capabilities under high transmit power interference environments and increasingly stringent QoS constraints. Future work will focus on extending this framework to multi-cell cooperative interference management scenarios and further investigating real-time resource scheduling for integrated ultra-reliable low-latency communication slices under high user mobility.

Author Contributions: Conceptualization, L.D. and J.W.; software, L.D. and Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Antoine, H.; Wang, J.-T.; Tang, C.C. From 5G to beyond 5G: From 2G to 6G: A Comprehensive Survey on the Evolution and Enhancement of AKA Protocols for Next-generation Wireless Networks. *ICT Express* **2026**, *3*, 1–11.
2. Elhattab, M.; Arfaoui, M.A.; Assi, C. Joint Clustering and Power Allocation in Coordinated Multipoint Assisted C-NOMA Cellular Networks. *IEEE Trans. Commun.* **2022**, *70*, 3483–3498. [CrossRef]

3. Messou, F.J.A.; Chen, J.; Zhang, S.; Liu, T.; Tolba, A.; Alfarraj, O.; Yu, K. TAPformer: A transformer with adaptive attention and temporal projection encoding for long-term electricity load forecasting in IoT systems. *Internet Things* **2026**, *37*, 101895. [CrossRef]
4. Zhang, Y.; Cheng, Z.; Guo, D.; Yuan, S.; Ma, T.; Zhang, Z. Downlink resource allocation for NOMA-based hybrid spectrum access in cognitive network. *China Commun.* **2023**, *20*, 171–184. [CrossRef]
5. Wang, C.X.; You, X.; Gao, X.; Zhu, X.; Li, Z.; Zhang, C.; Wang, H.; Huang, Y.; Chen, Y.; Haas, H.; et al. On the Road to 6G: Visions, Requirements, Key Technologies, and Testbeds. *IEEE Commun. Surv. Tutor.* **2023**, *25*, 905–974. [CrossRef]
6. Jayawardena, C.; Katsaros, G.N.; Nikitopoulos, K. NL-COMM: Enabling High-Performing Next-Generation Networks via Advanced Non-Linear Processing. *Future Internet* **2025**, *17*, 447. [CrossRef]
7. Niu, J.; Zhang, S.; Chi, K.; Shen, G.; Gao, W. Deep Learning for Online Computation Offloading and Resource Allocation in NOMA. *Comput. Netw.* **2022**, *216*, 109238. [CrossRef]
8. Ren, J.; Guo, C. A Game Theoretic Approach for D2D Assisted Uncoded Caching in IoT Networks. *Future Internet* **2025**, *17*, 423. [CrossRef]
9. Chung, Y.L. Transmission-Path Selection with Joint Computation and Communication Resource Allocation in 6G MEC Networks with RIS and D2D Support. *Future Internet* **2025**, *17*, 565. [CrossRef]
10. Shen, K.Z.; So, D.K.C.; Tang, J.; Ding, Z. Power Allocation for NOMA with Cache-Aided D2D Communication. *IEEE Trans. Wirel. Commun.* **2024**, *23*, 529–542. [CrossRef]
11. Huang, W.; Song, T.; An, J. QA2: QoS-Guaranteed Access Assistance for Space–Air–Ground Internet of Vehicle Networks. *IEEE Internet Things J.* **2022**, *9*, 5684–5695. [CrossRef]
12. Debbabi, F.; Jmal, R.; Fourati, L.C.; Aguiar, R.L. An Overview of Interslice and Intraslice Resource Allocation in B5G Telecommunication Networks. *IEEE Trans. Netw. Serv. Manag.* **2022**, *19*, 5120–5132. [CrossRef]
13. Rafique, W.; Hafid, A.S.; Cherkaoui, S. Complementing IoT Services Using Software-Defined Information Centric Networks: A Comprehensive Survey. *IEEE Internet Things J.* **2022**, *9*, 23545–23569. [CrossRef]
14. Lim, H.; Ullah, I.; Han, Y.; Kim, S. Reinforcement Learning-based Virtual Network Embedding: A Comprehensive Survey. *ICT Express* **2023**, *9*, 983–994. [CrossRef]
15. Azimi, Y.; Yousefi, S.; Kalbkhani, H.; Kunz, T. Energy-Efficient Deep Reinforcement Learning Assisted Resource Allocation for 5G-RAN Slicing. *IEEE Trans. Veh. Technol.* **2022**, *71*, 856–871. [CrossRef]
16. Jha, R.K.; Pedhadiya, M.K.; Dogra, A.; Kour, H.; Puja. Joint Resource and Power Allocation for 5G Enabled D2D Networking with NOMA. *Comput. Netw.* **2023**, *222*, 109536. [CrossRef]
17. Zhang, C.; Zhang, W.; Wu, Q.; Fan, P.; Fan, Q.; Wang, J.; Letaief, K.B. Distributed Deep Reinforcement Learning-Based Gradient Quantization for Federated Learning Enabled Vehicle Edge Computing. *IEEE Internet Things J.* **2025**, *12*, 4899–4913. [CrossRef]
18. Ding, Z. A Study on the Optimality of Downlink Hybrid NOMA. *IEEE Signal Process. Lett.* **2025**, *32*, 511–515. [CrossRef]
19. Abuajwa, O.; Mitani, S. Dynamic Resource Allocation for Energy-efficient Downlink NOMA Systems in 5G Networks. *Heliyon* **2024**, *10*, e29956. [CrossRef]
20. He, X.; Huang, Z.; Wang, H.; Song, R. Sum Rate Analysis for Massive MIMO-NOMA Uplink System with Group-Level Successive Interference Cancellation. *IEEE Wirel. Commun. Lett.* **2023**, *12*, 1194–1198. [CrossRef]
21. Wu, G.; Chen, G. Task Offloading and Resource Allocation in Cellular Heterogeneous Networks for NOMA-based Mobile Edge Computing. *Ad Hoc Netw.* **2025**, *169*, 103742. [CrossRef]
22. Li, N.; Wu, P.; Ning, B.; Zhu, L. Sum Rate Maximization for Movable Antenna Enabled Uplink NOMA. *IEEE Wirel. Commun. Lett.* **2024**, *13*, 2140–2144. [CrossRef]
23. Meng, S.; Wang, X.; Zhang, H.; Qian, Z.; Zou, Y.; Liu, Y. Capacity Enhancement for D2D-Assisted Cooperative NOMA Systems. *IEEE Trans. Commun.* **2025**, *73*, 6546–6560. [CrossRef]
24. Wei, L.; Guan, Z. Energy-efficient Joint Power Control and Channel Allocation for D2D Communication Underlaying Cellular Network. *Phys. Commun.* **2025**, *68*, 102552. [CrossRef]
25. Jose, J.; Agarwal, A.; Shaik, P.; Goyal, V.; Choi, K.; Bhatia, V. Performance Analysis and Learning-Assisted Power Control for NOMAEnabledD2D-Cellular Network. *IEEE Syst. J.* **2024**, *18*, 278–281. [CrossRef]
26. Moussa, S.; Benslimane, A.; Darazi, R.; Jiang, C. Power Allocation-Based NOMA and Underlay D2D Communication for Public Safety Users in the 5G Cellular Network. *IEEE Syst. J.* **2023**, *17*, 3572–3583. [CrossRef]
27. Tran, D.D.; Ha, V.N.; Sharma, S.K.; Nguyen, T.T.; Chatzinotas, S.; Popovski, P. Energy-Efficient NOMA for 5G Heterogeneous Services: A Joint Optimization and Deep Reinforcement Learning Approach. *IEEE Trans. Commun.* **2025**, *73*, 2448–2465. [CrossRef]
28. Amer, A.; Hoteit, S.; Othman, J.B. Throughput Maximization in Multi-slice Cooperative NOMA-based System with Underlay D2D Communications. *Comput. Commun.* **2024**, *217*, 134–151. [CrossRef]
29. Taskou, S.K.; Rasti, M.; Hossain, E. End-to-End Resource Slicing for Coexistence of eMBB and URLLC Services in 5G Advanced/6G Networks. *IEEE Trans. Mob. Comput.* **2024**, *23*, 8015–8032. [CrossRef]
30. Liu, W.; Hossain, M.A.; Ansari, N.; Kiani, A.; Saboorian, T. Reinforcement Learning-Based Network Slicing Scheme for Optimized UE-QoS in Future Networks. *IEEE Trans. Netw. Serv. Manag.* **2024**, *21*, 3454–3464. [CrossRef]

31. Lotfi, F.; Rajoli, H.; Afghah, F. LLM-Augmented Deep Reinforcement Learning for Dynamic O-RAN Network Slicing. In Proceedings of the IEEE International Conference on Communications-ICC, Montreal, QC, Canada, 11 June 2025.
32. Hossain, M.A.; Ansari, N. Network Slicing for NOMA-Enabled Edge Computing. *IEEE Trans. Cloud Comput.* **2023**, *11*, 811–821. [CrossRef]
33. Jeong, Y.J.; Yu, S.; Lee, J.W. DRL-Based Resource Allocation for NOMA-Enabled D2D Communications Underlay Cellular Networks. *IEEE Access* **2023**, *11*, 140270–140286. [CrossRef]
34. Khan, M.A.A.; Kaidi, H.M.; Ahmad, N.; Rehman, M.U. Sum Throughput Maximization Scheme for NOMA-Enabled D2D Groups Using Deep Reinforcement Learning in 5G and Beyond Networks. *IEEE Sens. J.* **2023**, *23*, 15046–15057. [CrossRef]
35. Syed, M.H.; Jawwad, N.C.; Muhammad, B. Towards Intelligent User Clustering Techniques for Non-orthogonal Multiple Access: A Survey. *EURASIP J. Wirel. Commun. Netw.* **2024**, *2024*, 7. [CrossRef]
36. Ghafouri, N.; Vardakas, J.S.; Ramantas, K.; Verikoukis, C. A Multi-Level Deep RL-Based Network Slicing and Resource Management for O-RAN-Based 6G Cell-Free Networks. *IEEE Trans. Veh. Technol.* **2024**, *73*, 17472–17484. [CrossRef]
37. Wu, X.; Farooq, J.; Chen, J. Multi-Agent Resource Orchestration Based on D3QN for Network Slicing in 5G Edge-Cloud Networks. *IEEE Trans. Netw. Serv. Manag.* **2026**, *23*, 1766–1781. [CrossRef]
38. Xu, C.; Zhang, P.; Yu, H.; Li, Y. D3QN-Based Multi-Priority Computation Offloading for Time-Sensitive and Interference-Limited Industrial Wireless Networks. *IEEE Trans. Veh. Technol.* **2024**, *73*, 13682–13693. [CrossRef]
39. Tran, D.D.; Sharma, S.K.; Ha, V.N.; Chatzinotas, S.; Woungang, I. Multi-Agent DRL Approach for Energy-Efficient Resource Allocation in URLLC-Enabled Grant-Free NOMA Systems. *IEEE Open J. Commun. Soc.* **2023**, *4*, 1470–1486. [CrossRef]
40. Nauman, A.; Alshahrani, H.M.; Nemri, N.; Othman, K.M.; Aljehane, N.O.; Maashi, M.; Dutta, A.K.; Assiri, M.; Khan, W.U. Dynamic Resource Management in Integrated NOMA Terrestrial-satellite Networks Using Multi-agent Reinforcement Learning. *J. Netw. Comput. Appl.* **2024**, *221*, 103770. [CrossRef]
41. Chen, J.; Jia, J.; Liu, Y.; Wang, X.; Aghvami, A.H. Optimal Resource Block Assignment and Power Allocation for D2D-Enabled NOMA Communication. *IEEE Access* **2019**, *7*, 90023–90035. [CrossRef]
42. Luong, N.C.; Hoang, D.T.; Gong, S.; Niyato, D.; Wang, P.; Liang, Y.C.; Kim, D.I. Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 3133–3174. [CrossRef]
43. Vishnoi, V.; Budhiraja, I.; Gupta, S.; Kumar, N. A Deep Reinforcement Learning Scheme for Sum Rate and Fairness Maximization Among D2D Pairs Underlaying Cellular Network with NOMA. *IEEE Trans. Veh. Technol.* **2023**, *72*, 13506–13522. [CrossRef]
44. Yu, S.; Lee, J.W. Deep Reinforcement Learning Based Resource Allocation for D2D Communications Underlay Cellular Networks. *Sensors* **2022**, *22*, 9459. [CrossRef] [PubMed]