Adaptive and Efficient Resource Allocation for Cognitive Radio Networks

by

Nagwa Shaghluf

B.Sc., University of Tripoli, 2009

M.Sc., University of Victoria, 2018

A Dissertation Submitted in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Electrical and Computer Engineering

University of Victoria

We acknowledge and respect the Lək̓ʷəŋən (Songhees and Xʷsepsəm/Esquimalt) Peoples on whose territory the university stands, and the Lək̓ʷəŋən and W̱SÁNEĆ Peoples whose historical relationships with the land continue to this day.

Adaptive and Efficient Resource Allocation for Cognitive Radio Networks

by

Nagwa Shaghluf
B.Sc., University of Tripoli, 2009
M.Sc., University of Victoria, 2018

Supervisory Committee

Dr. T. Aaron Gulliver, Supervisor
(Department of Electrical and Computer Engineering)

Dr. Xiaodai Dong, Departmental Member

(Department of Electrical and Computer Engineering)

Dr. Andrew Rowe, Outside Member

(Department of Mechanical Engineering)

ABSTRACT

The increasing demand for wireless connectivity, massive device access, and data-intensive applications has intensified the need for efficient and intelligent spectrum utilization in modern wireless networks. Although radio spectrum is a limited resource, many licensed bands remain underutilized due to static allocation policies. Cognitive Radio Networks (CRNs) address this challenge by enabling Secondary Users (SUs) to dynamically access underutilized spectrum while protecting Primary Users (PUs) from harmful interference. However, efficient resource allocation in CRNs remains challenging due to sensing uncertainty, dynamic channel conditions, interference coupling, limited Channel State Information (CSI), and the high computational complexity of conventional optimization methods.

This dissertation develops an intelligent and adaptive resource allocation framework for CRNs by improving spectrum awareness through predictive spectrum sensing and progressively integrating Multi-Agent Deep Reinforcement Learning (MADRL) and Non-Orthogonal Multiple Access (NOMA). The first contribution is a Predictive-Cooperative Spectrum Sensing (PCSS) framework that combines spectrum prediction with Cooperative Spectrum Sensing (CSS). Unlike conventional sensing schemes that rely only on instantaneous sensing decisions, the proposed PCSS approach uses historical cooperative sensing decisions to predict future PU channel availability before sensing is performed. This reduces unnecessary sensing of busy channels and improves spectrum awareness. The sensing time and fusion decision threshold are jointly optimized to characterize the trade-off between Spectrum Efficiency (SE) and Energy Efficiency (EE). Simulation results show that PCSS improves SE and achieves a better SE–EE trade-off compared with conventional sensing, local prediction, and cooperative spectrum prediction schemes.

The second contribution is a MADRL-based resource allocation framework for Cognitive Device-to-Device (C-D2D) communication underlaying cellular networks. The joint Resource Block (RB) assignment and power control problem is formulated as a Mixed-Integer Nonlinear Programming (MINLP) problem due to binary RB allocation, nonlinear SINR expressions, and interference-coupled constraints. To address this complexity, a Proximal Policy Optimization (PPO)-based MADRL approach is developed, where each C-D2D pair acts as an autonomous agent. Both Centralized Training with Decentralized Execution (CTDE) and Decentralized Training with Decentralized Execution (DTDE) architectures are investigated under perfect and

imperfect CSI. Simulation results demonstrate that the proposed PPO-based MADRL framework improves C-D2D sum-rate, scalability, and robustness compared with random allocation, random search, and other DRL baselines.

The third contribution extends the framework to NOMA-enabled C-D2D networks. In this model, each C-D2D transmitter serves two receivers using power-domain NOMA while reusing cellular uplink RBs under CR protection constraints. This introduces additional challenges due to inter-group interference, intra-group NOMA interference, SIC decoding requirements, and fairness among C-D2D groups. A PPO-based MADRL framework is proposed for joint RB reuse and power allocation under constrained sum-rate and fairness-aware objectives. Different observation models, including local, partial, and full CSI, are systematically evaluated to study the impact of information availability on learning, coordination, and performance. The results show that structured partial CSI can approach full-CSI performance while reducing signaling overhead. Fairness-aware reward shaping further improves rate distribution, coordination stability, and EE when the fairness coefficient is properly selected.

Overall, this dissertation demonstrates that intelligent CRNs can be developed through a progressive framework that begins with spectrum awareness and extends toward decentralized learning-based resource allocation. The proposed approaches improve SE, EE, throughput, fairness, and constraint satisfaction compared with conventional optimization and learning-based baselines. The work provides a scalable foundation for future 5G, beyond-5G, and 6G networks requiring adaptive, distributed, and autonomous spectrum management under practical constraints.

Contents

List of Figures

List of Acronyms

AWGN	Additive White Gaussian Noise
AI	Artificial Intelligence
BS	Base Station
B5G	Beyond-5G
CN	Cellular Network
CU	Cellular User
CSI	Channel State Information
CDMA	Code Division Multiple Access
CD-NOMA	Code-domain NOMA
C-D2D	Cognitive Device-to-Device
CR	Cognitive Radio
CRN	Cognitive Radio Network
CSP	Cooperative Spectrum Prediction
CSS	Cooperative Spectrum Sensing
CTDE	Centralized Training with Decentralized Execution
DTDE	Decentralized Training with Decentralized Execution
DDPG	Deep Deterministic Policy Gradient
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
D2D	Device-to-Device
DDQN	Double Deep Q-Network
DSA	Dynamic Spectrum Access
ED	Energy Detection
EE	Energy Efficiency
EH	Energy Harvesting
EMA	Exponential Moving Average
FCC	Federal Communications Commission
FDMA	Frequency Division Multiple Access

FC	Fusion Center
GAE	Generalized Advantage Estimate
GA	Genetic Algorithm
HMM	Hidden Markov Model
IoT	Internet of Things
IMPALA	Importance Weighted Actor-Learner Architecture
IRS	Intelligent Reflecting Surface
ML	Machine Learning
M2M	Machine-to-Machine
MDP	Markov Decision Process
MINLP	Mixed-Integer Nonlinear Programming
MADRL	Multi-Agent Deep Reinforcement Learning
NOMA	Non-Orthogonal Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
OMA	Orthogonal Multiple Access
PCSS	Predictive-Cooperative Spectrum Sensing
POMDP	Partially Observable Markov Decision Process
PD-NOMA	Power-Domain NOMA
PSS	Predictive Spectrum Sensing
PU	Primary User
PPO	Proximal Policy Optimization
QoS	Quality of Service
RRA	Random Resource Allocation
RS	Random Search
RL	Reinforcement Learning
RB	Resource Block
SU	Secondary User
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SA-DRL	Single-Agent Deep Reinforcement Learning
SAC	Soft Actor Critic

SE	Spectrum Efficiency
SIC	Successive Interference Cancellation
SP	Spectrum Prediction
SS	Spectrum Sensing
TDMA	Time Division Multiple Access
V2V	Vehicle-to-Vehicle

List of Tables

ACKNOWLEDGEMENTS

This journey has been both challenging and rewarding, and it would not have been possible without the support and encouragement of many individuals.

I would like to express my sincere gratitude to my supervisor, Prof. T. Aaron Gulliver, for his continuous guidance, support, and patience throughout my Ph.D. degree program. His deep expertise, insightful feedback, and encouragement have been instrumental in shaping this research. I am truly grateful for his mentorship and for inspiring my academic and professional growth.

I would also like to thank my supervisory committee members, Prof. Xiaodai Dong and Prof. Andrew Rowe, for their valuable feedback and thoughtful suggestions, which have significantly improved the quality of this work. I am equally thankful to my external examiner, Prof. Brent Petersen, for his time, careful review, and constructive comments.

My deepest appreciation goes to my family. Your unconditional love and unwavering support have been my foundation throughout this journey and in all aspects of my life. You have been by my side through both challenging and joyful moments, and for that, I am forever grateful.

To my dear daughters, Yasmine, Ayat, and Maysam, you are the light of my life and my greatest source of happiness and motivation. Everything I do is inspired by you.

Finally, to my husband, Mohamed, words cannot fully express my gratitude. Your constant support and unwavering belief in me have carried me through this journey. You have stood by me through every challenge and every success. Without your encouragement, this achievement would not have been possible. This accomplishment is as much yours as it is mine.

DEDICATION

To the memory of my parents, Abdulbaset and Shukria, whose love, guidance, and values continue to inspire me every day. Though they are no longer with me, their belief in the power of knowledge and perseverance remains the foundation of all that I have achieved.

To my beloved husband, Mohamed, for his unwavering love, patience, and support throughout this journey.

And to my precious children, Yasmine, Ayat, and Maysam, who fill my life with joy, hope, and purpose, you are my greatest inspiration.

Chapter 1

Introduction

The rapid proliferation of wireless devices, Internet of Things (IoT) applications, and data-intensive services has led to an unprecedented increase in demand for radio spectrum resources. Modern wireless networks must support massive connectivity, high data rates, and low latency, placing significant pressure on the limited available spectrum. Despite this apparent scarcity, empirical studies have shown that large portions of licensed spectrum remain underutilized due to static allocation policies. This inefficiency highlights the need for more flexible and intelligent spectrum access mechanisms. Cognitive Radio Networks (CRNs) have emerged as a promising solution to address this challenge by enabling dynamic spectrum access. In CRNs, unlicensed users, referred to as Secondary Users (SUs), are allowed to opportunistically access underutilized spectrum bands while ensuring minimal interference to licensed Primary Users (PUs). This paradigm introduces adaptability into wireless systems, allowing devices to sense, learn, and respond to their surrounding environment.

A fundamental requirement for CRNs is accurate spectrum awareness, typically achieved through spectrum sensing. However, conventional sensing techniques, such as energy detection, are limited by noise uncertainty, fading effects, and sensing delays. These limitations can lead to inaccurate detection of spectrum availability, resulting in reduced Spectrum Efficiency (SE) and increased energy consumption. To address these challenges, predictive approaches have been introduced, enabling SUs to anticipate future spectrum availability based on historical observations. While Predictive Spectrum Sensing (PSS) improves awareness, efficient spectrum utilization also requires intelligent resource allocation. In this context, Device-to-Device (D2D) communication has been proposed as a key technology to enhance spectrum reuse and reduce communication latency. Cognitive Device-to-Device (C-D2D) communication allows devices to directly communicate while reusing cellular spectrum under interference constraints. However, this introduces significant challenges related to interference management, resource allocation, and system scalability, particularly in dynamic and decentralized environments.

Recent advances in Artificial Intelligence (AI), particularly Deep Reinforcement Learning (DRL), provide powerful tools for addressing these challenges. DRL enables agents to learn optimal decision-making policies through interaction with the environment, without requiring explicit system models. In multi-user wireless systems, Multi-Agent Deep Reinforcement Learning (MADRL) enables distributed and scalable resource allocation, making it well suited for C-D2D networks operating under dynamic conditions. In parallel, Non-Orthogonal Multiple Access (NOMA) has emerged as a key technology for improving SE by allowing multiple users to share the same time–frequency resources through power-domain multiplexing. The integration of NOMA with C-D2D communication further enhances spectrum utilization but introduces additional complexity due to interference coupling and Successive Interference Cancellation (SIC) constraints.

This chapter introduces the background and motivation for the dissertation and presents the key enabling technologies in CRNs. It discusses spectrum sensing and prediction, C-D2D communication, DRL, and NOMA-based spectrum sharing. The chapter also identifies key research gaps that motivate the proposed contributions, which are developed in the subsequent chapters.

1.1 Cognitive Radio Fundamentals

Wireless communication has experienced rapid growth over the past two decades, resulting in a significant increase in spectrum demand. Traditional static spectrum allocation assigns fixed frequency bands to specific services, leading to inefficient spectrum utilization. While some bands become congested, others remain underutilized. This inefficiency is primarily due to poor spectrum usage rather than actual scarcity [1]. Studies conducted by the Federal Communications Commission (FCC) indicate that spectrum utilization varies significantly, ranging from 15% to 85% depending on time and location. To address this issue, CR has been introduced as a promising solution for improving spectrum efficiency. CR enables unlicensed users, known as SUs, to access licensed spectrum either by opportunistically utilizing unused bands (spectrum holes) or by coexisting with PUs under controlled interference conditions [2]. A CR system continuously senses and analyzes its surrounding environment to detect spectrum availability. Based on this information, it dynamically adapts its transmission parameters, such as frequency, power, and

modulation, to optimize performance while maintaining Quality of Service (QoS) requirements. This adaptability makes CR a key enabler for future wireless communication systems [1].

Spectrum sensing plays a critical role in identifying spectrum holes, which represent frequency bands that are temporarily unoccupied by PUs. When a PU becomes active, the CR system must promptly detect its presence and vacate the channel to avoid interference. This dynamic behavior is illustrated in Figure 1.1. Once spectrum availability is identified, spectrum management and spectrum handoff mechanisms allow SUs to select optimal channels and switch frequencies when necessary. These decisions are based on factors such as interference levels, noise, channel conditions, and QoS requirements. CR systems are characterized by two main capabilities. The first is cognitive capability, which enables the system to sense and analyze environmental conditions, including spectrum availability and interference levels. The second is reconfigurability, which allows the system to dynamically adjust its operational parameters to maintain efficient communication under varying network conditions [1].

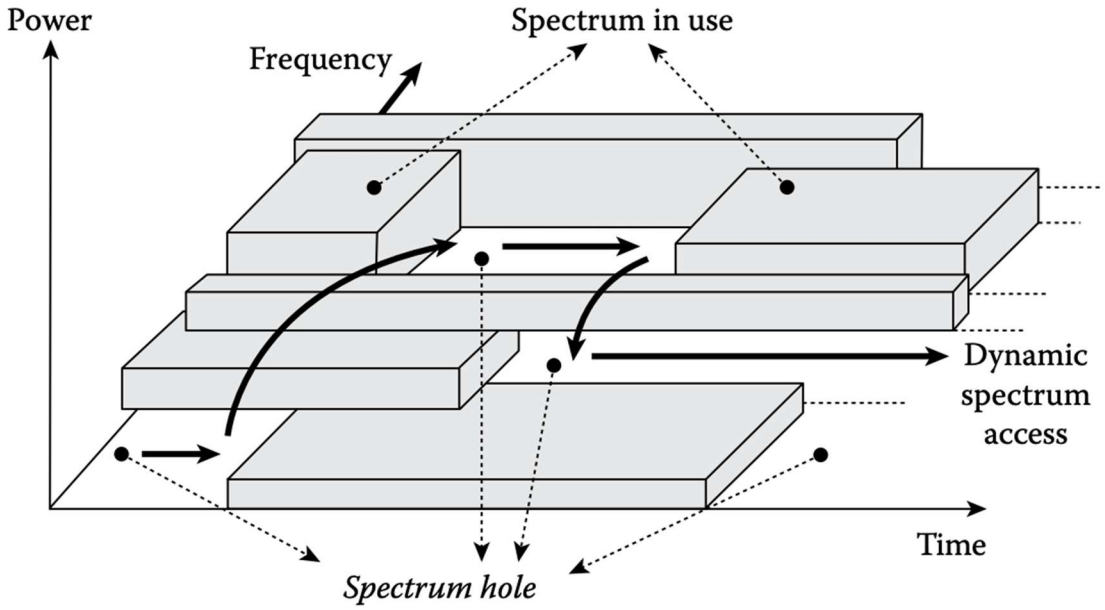


Figure 1.1: Concept of spectrum hole or white space [1].

1.1.1 Cognitive Radio Paradigms

CR systems operate under three primary paradigms: interweave, underlay, and overlay, which differ in how SUs access the licensed spectrum. In the interweave paradigm, SUs opportunistically access spectrum only when it is not occupied by PUs. This approach, commonly referred to as opportunistic spectrum access, relies heavily on accurate spectrum sensing. SUs are prohibited from transmitting when PUs are active. Therefore, they must vacate the channel immediately upon detecting primary user activity. In the underlay paradigm, SUs are allowed to transmit concurrently with PUs, provided that the interference introduced remains below a predefined threshold. This is typically achieved by limiting the transmit power of SUs. While this approach enables continuous spectrum access, it restricts transmission range and data rates. In the overlay paradigm, SUs assist primary transmissions, for example by relaying signals, in exchange for spectrum access. This approach requires prior knowledge of the PU signal, which increases system complexity and coordination requirements. Although overlay communication can improve SE, its practical implementation remains challenging [2]-[4].

1.1.2 Features of Cognitive Radio Networks

CRNs operate through a cognitive cycle consisting of SS, spectrum management, spectrum sharing, and spectrum mobility. This cognitive cycle relies on several key functions, as shown in Figure 1.2. SS is responsible for detecting PU activity and identifying available spectrum. Accurate sensing is essential, as subsequent operations depend on reliable detection results. The primary objective is to determine whether a channel is occupied or idle while minimizing interference to PUs. Spectrum management involves selecting the most suitable frequency bands based on channel conditions, such as Signal-to-Noise Ratio (SNR), interference levels, and channel availability. This process includes spectrum analysis and spectrum decisions, where sensing data is evaluated and appropriate transmission strategies are determined.

Spectrum sharing enables multiple users to access available spectrum dynamically. Efficient sharing mechanisms are required to coordinate access among users and avoid collisions and excessive interference. Spectrum mobility, also known as spectrum handoff, ensures that SUs can switch to alternative frequency bands when a primary user becomes active. This enables seamless

communication while maintaining QoS requirements [1], [3], [5]. Despite these capabilities, CRNs face several challenges, including dynamic spectrum availability, decentralized operation, and the need to balance QoS requirements with spectrum efficiency. Addressing these challenges requires advanced adaptive and intelligent resource management strategies.

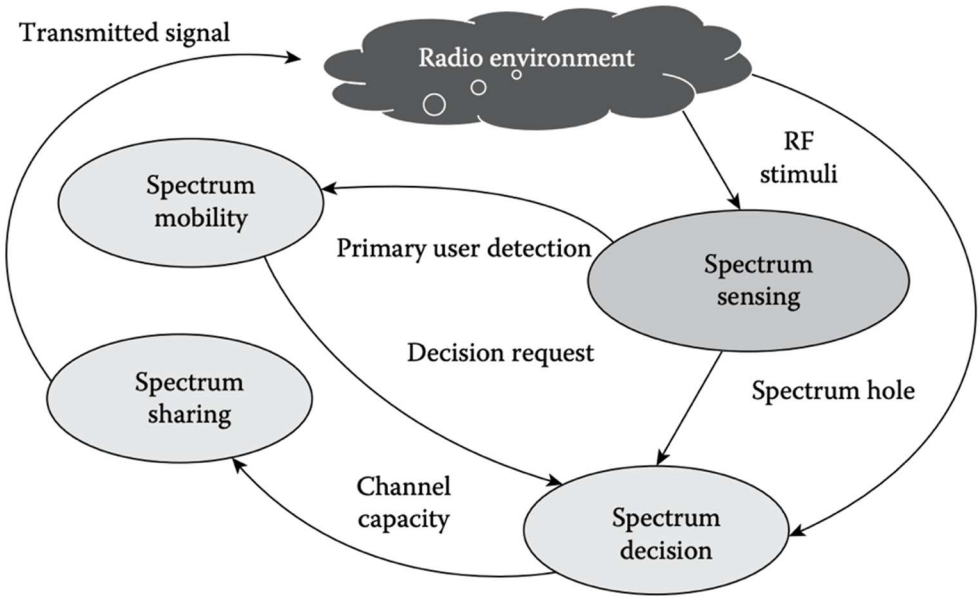


Figure 1.2: The cognitive cycle [1].

1.1.3 Cognitive Radio Networks Components

A CRN consists of two main components: the primary network and the cognitive (secondary) network. The primary network represents the licensed communication system, where PUs have exclusive access to specific spectrum bands. Primary Base Stations (BSs) manage communication and ensure reliable service for licensed users. The CRN operates without licensed spectrum access and opportunistically utilizes available frequency bands. SUs employ spectrum sensing and dynamic access techniques to utilize both licensed and unlicensed spectrum without interfering with PUs. This coexistence requires careful coordination and intelligent decision-making to ensure efficient spectrum utilization while protecting primary user communications [5].

1.2 Spectrum Sensing in CRNs

CRs are designed to be aware of their surroundings, making SS a critical function in CRNs. SS allows CR users to adaptively identify and utilize unused spectrum based on the radio environment.

SS techniques, focused mainly on detecting primary transmitters through local SU measurements, are categorized based on their complexity, accuracy, and reliance on prior information.

This work focuses on the energy detector as it is a widely used spectrum-sensing method due to its simplicity and low computational complexity. It does not require prior knowledge of the PU signal and determines spectrum occupancy by comparing the computed energy of the received signal with a threshold [1].

1.2.1 Energy Detection

Energy Detection (ED) is one of the most widely used SS techniques in CRNs due to its simplicity and low implementation complexity. It is a semi-blind method that requires only knowledge of the noise power and does not depend on prior information about the PU signal. The method determines the presence of a PU by comparing the energy of the received signal over a given observation interval to a predefined threshold.

Let $S(n)$ represent the PU signal and $W(n)$ denote the Additive White Gaussian Noise (AWGN) for the n-th signal. The received signal $y(n)$ can be expressed as [7]

$$y(n) = S(n) + W(n) \tag{1.1}$$

If $S(n) = 0$, it indicates no transmission by the PU. The two hypotheses for signal detection are H_0**:** No primary signal present $y(n) = W(n)$ and H_1**:** Primary signal present $y(n) = S(n) + W(n)$. The detector computes the energy of the received signal over N samples and compares it to a threshold λ. As illustrated in Figure 2.3, the decision made based on the test statistic is given by $T(y) = \frac{1}{\sigma^2}\sum y^2$, where σ^2 denotes the noise variance and N is the number of samples [8]. If $T(y)$ exceeds the threshold λ, the presence of a primary signal is declared; otherwise, the channel is considered idle. The performance of ED is characterized by two key metrics. The probability of detection (P_d) represents the likelihood of correctly detecting the presence of a PU, $P_d = P_r(T >$

$\lambda | H_1)$. The probability of false alarm (P_f) represents the likelihood of incorrectly detecting a primary signal when the channel is idle, $P_f = P_r(T > \lambda | H_0)$. A high P_d is critical to avoid interference with PUs, while a low P_f ensures efficient spectrum utilization. However, achieving perfect detection is impossible due to the inherent statistical nature of the problem. Signal detectors are therefore designed to minimize errors within acceptable levels [6]. A high detection probability is essential to protect PUs from interference, while a low false alarm probability ensures efficient spectrum utilization. In practice, a tradeoff exists between these two metrics due to the stochastic nature of noise and channel conditions [6].

Despite its advantages, ED has several limitations. Its performance is highly sensitive to noise uncertainty, which can significantly degrade detection accuracy, especially in low SNR environments. In addition, ED cannot distinguish between different signal types and may falsely detect interference or noise as a primary signal. Furthermore, achieving reliable detection often requires longer observation times compared to more complex techniques such as matched filtering [1], [6].

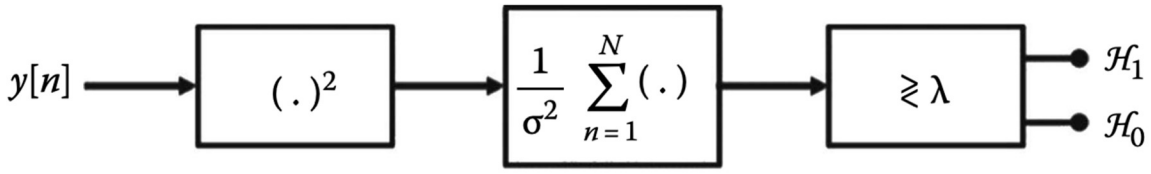


Figure 1.3: The energy detector [4].

1.2.2 Cooperative Spectrum Sensing

Spectrum sensing in CRNs is significantly affected by practical impairments such as fading, shadowing, noise uncertainty, and the hidden PU problem. In particular, a SU may fail to detect a primary transmission when the received signal is severely attenuated, which may lead to harmful interference with the primary user. As illustrated in Fig. 1.4(a), this hidden terminal problem occurs when a primary user is outside the sensing range of an SU due to obstacles or channel impairments. Cooperative Spectrum Sensing (CSS) has been proposed to mitigate these limitations

by allowing multiple spatially distributed SUs to share sensing information and make a more reliable spectrum occupancy decision [4], [6].

In CSS, each secondary user performs local sensing and reports its observation or decision to other users or to a fusion center. The shared information is then combined to determine whether the primary user is present. By exploiting spatial diversity, CSS improves detection reliability and reduces the sensitivity of spectrum sensing to local channel impairments such as shadowing and deep fading. As shown in Fig. 1.4(b), cooperation among users enables correct detection even when individual users experience poor channel conditions [5].

CSS can be broadly classified into centralized and distributed approaches. In centralized sensing, users forward their local sensing results to a fusion center, which makes the final decision. In distributed sensing, users exchange sensing information directly and collaboratively reach a decision without a central coordinator. Depending on the type of shared information, CSS may also be categorized into hard combining and soft combining. In hard combining, each user sends a one-bit local decision, which reduces reporting overhead. In soft combining, users send test statistics or more detailed sensing information, which generally improves detection performance at the cost of higher communication overhead [4], [6].

The main advantage of CSS is its ability to improve sensing accuracy and mitigate the hidden terminal problem. However, these gains come with additional sensing, reporting, and coordination overhead. Therefore, the design of CSS schemes requires balancing sensing reliability against complexity, delay, and energy consumption [4], [6].

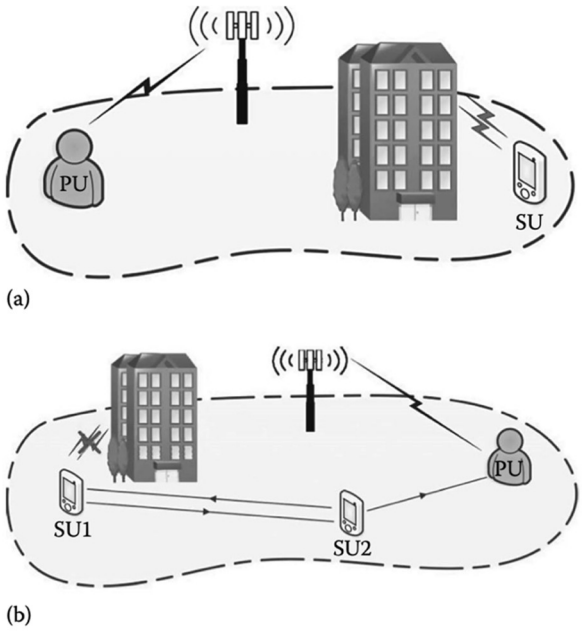


Figure 1.4: (a) Illustration of the hidden terminal problem. (b) Cooperative sensing mitigating the hidden terminal problem [4].

1.3 Spectrum Prediction in CRNs

Spectrum prediction has emerged as an effective approach to enhance spectrum awareness in CRNs by exploiting temporal correlations in channel occupancy. Unlike conventional spectrum sensing, which relies on instantaneous observations, prediction techniques utilize historical sensing data to estimate future channel states. This enables SUs to make proactive spectrum access decisions, thereby improving spectral efficiency and reducing the collision probability with PUs [7].

Various prediction methods have been proposed, including Machine Learning (ML) techniques and statistical models such as Hidden Markov Models (HMMs). In these approaches, the channel occupancy is typically modeled as a stochastic process with two states, namely idle and busy. Since the true channel state cannot be directly observed, it is treated as a hidden variable, while the sensing outcomes represent observable data. By learning the temporal dynamics of channel usage, the SU can predict the likelihood of future channel availability [10].

Among these methods, HMM-based prediction has been widely adopted due to its ability to capture temporal dependencies and model uncertainty in spectrum sensing. In this framework, the sequence of sensing observations is used to estimate the underlying channel states and predict future states based on learned transition behavior. The overall prediction process typically involves model training, state estimation, and future state prediction.

Spectrum prediction improves sensing performance by reducing uncertainty and enabling more informed channel selection. However, its effectiveness depends on the accuracy of the learned model and the quality of historical sensing data. Inaccurate training data or rapidly changing environments may lead to unreliable predictions [10]. Overall, spectrum prediction complements conventional sensing techniques and provides a foundation for more intelligent and adaptive spectrum access strategies in CRNs.

1.4 Emerging Technologies and Applications

The evolution of cognitive networks will be driven by advances in emerging technologies and applications, enabling new use cases and improving network capabilities. Integration with 5G infrastructure, including features like massive multiple-input multiple-output, edge computing, devices in proximity, and network slicing, will allow cognitive networks to support a wide range of applications that demand high throughput, low latency, and high reliability. AI and ML will play a crucial role in optimizing resource allocation, predicting network behavior, and adapting to dynamic user demands, thereby improving both efficiency and user experience [8].

Additionally, incorporating D2D communication will enable cognitive networks to support devices in proximity, such as IoT devices, Vehicle-to-Vehicle (V2V), and Machine-to-Machine (M2M) deployments with diverse connectivity requirements, such as low-power, wide-area, and reliable communication. Moreover, the rapid evolution of wireless communication networks has brought the need for innovative technologies to meet the increasing demands for higher data rates, improved SE, massive connectivity, and ultra-reliable low-latency communications. Among the emerging paradigms, NOMA and DRL have garnered significant attention due to their potential to revolutionize resource management and performance optimization in next-generation networks such as 5G, 6G, and beyond.

The integration of CR techniques with D2D communication aims to combine the strengths of both approaches, enhancing the efficiency and flexibility of wireless networks. C-D2D communication can operate in an underlay mode, where C-D2D pairs, functioning as SUs, share the same spectrum as PUs. To maintain the service quality of PUs, C-D2D pairs must comply with strict interference constraints, enabling Dynamic Spectrum Access (DSA) with minimal disruption. This approach offers significant benefits, including direct D2D connectivity, efficient spectrum utilization, and adaptive spectrum access, all of which enhance overall network performance. By improving SE, C-D2D pairs can access resources effectively while safeguarding PU services.

However, these networks face challenges in managing interference and allocating resources, as C-D2D devices must make decentralized, real-time decisions to minimize disruptions to PUs. The C-D2D communication system combines CR and D2D communication, leveraging the advantages of opportunistic spectrum access for proximity-based communication within 5G/ B5G cellular networks. As illustrated in Figure 1.5, C-D2D communication enables C-D2D users to establish direct communication links while utilizing spectrum gaps identified through CR. This process occurs seamlessly, ensuring no interference with the parent cellular network. If a licensed user requires a channel currently occupied by C-D2D users, the C-D2D users will perform a spectrum handoff, vacating the channel to prioritize the legitimate user, who holds higher rights over the spectrum [11].

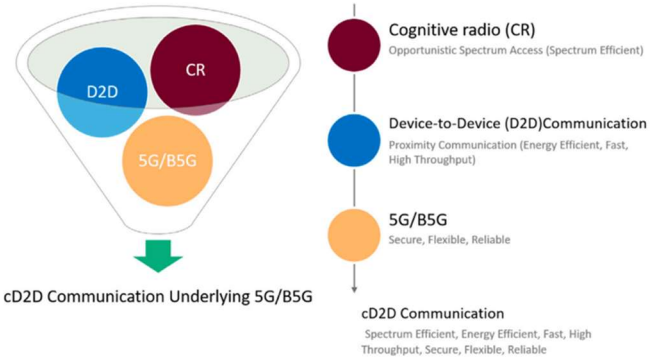


Figure 1.5: A C-D2D communication system underlying a cellular network [11].

In the C-D2D communication system, a C-D2D enabled user scans its environment to detect nearby users. Upon finding potential peers, the user opts for direct communication. Next, SS is performed to evaluate channel conditions. Once an available channel is identified, communication begins on that channel. During data transfer, the C-D2D users focus on optimizing power consumption to improve energy efficiency. Security is maintained through advanced algorithms to safeguard communication and prevent unauthorized data sharing, ensuring privacy. If supported, Energy Harvesting (EH) technologies can further extend the battery life of C-D2D users. Simultaneously, spectrum sharing is employed to facilitate cooperative communication among users. The session concludes once data transfer is complete [11].

1.5 Deep Reinforcement Learning

Reinforcement Learning (RL) is a significant branch of machine learning. RL techniques do not rely on explicit knowledge of optimal outputs. Unlike supervised learning, an agent learns through interaction with its environment. This process is referred to as model-free learning. The agent interacts with the environment by transitioning through states, taking actions, and receiving feedback in the form of rewards. These rewards, which can be positive or negative, inform the agent about the quality of its actions, where the impact often extends beyond the immediate reward, influencing future rewards as well. The primary goal is to discover an optimal policy that is the best mapping between states and actions that maximizes agent long-term rewards. Among various RL methods, Q-learning is particularly well-known [3], [17], 18].

A key aspect of RL is managing the tradeoff between exploration and exploitation. Exploration allows the agent to try new actions to discover their effectiveness, while exploitation leverages known actions to maximize rewards. Striking a balance is critical, as excessive focus on either can lead to suboptimal outcomes. RL techniques are widely used for decision-making in CR networks, addressing challenges like spectrum allocation, power control, and interference management [3], [18].

In wireless networks, RL can be employed to enable devices and sensors to make decisions and derive inferences under uncertain and dynamically changing network conditions. For example, RL is useful in spectrum sharing for channel access in CRNs, where channel availability is unknown, and in small cell networks, where resource conditions are uncertain. Additionally, RL

can assist with BS association in scenarios with unknown energy levels and load statuses. Although RL offers significant potential for solving complex decision-making problems, it may encounter challenges such as slow convergence rates when learning optimal policies for complex system states, as well as scalability issues arising from the need to process and explore large state and action spaces [17].

DRL combines Deep Neural Networks (DNNs) with RL, where a DNN serves as the agent in the RL framework. Traditional RL methods can be inefficient for problems involving large state spaces, high-dimensional data, or non-stationary environments due to slower convergence rates. By integrating DNNs with RL, agents can autonomously learn and develop policies that maximize long-term rewards. The experience gained through RL helps the DNN determine the best action for a given state. DRL methods are generally categorized into two approaches, value-based methods and policy-based methods. A value-based method such as Deep Q-Networks (DQN), where the Q-value, representing the expected future reward for each action, is calculated for all actions in the action space. The action with the highest Q-value is selected. In this approach, Q-learning is combined with a DNN, which approximates the state-action value function.

Policy-based reinforcement learning methods learn a parameterized policy that maps the observed network state to a probability distribution over possible actions. Instead of selecting actions only by estimating Q-values, these methods directly optimize the policy to maximize the expected cumulative reward. In actor–critic implementations, such as PPO, a value function is also used to estimate the expected return and improve the stability of policy updates. Deep reinforcement learning has therefore become a promising approach for intelligent wireless networks, especially for resource-allocation problems involving large state spaces, discrete or continuous actions, and multiple coupled optimization variables. It has also been proposed to support predictive analysis, autonomous decision-making, and network automation [17].

1.6 Non-Orthogonal Multiple Access

Traditional multiple access schemes in cellular networks, such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), Code Division Multiple Access (CDMA), and Orthogonal Frequency Division Multiple Access (OFDMA), rely on orthogonal resource allocation. In these schemes, users are assigned distinct time, frequency, or

code resources to avoid interference. While this approach simplifies receiver design, it limits the number of users that can simultaneously access the network and may lead to inefficient spectrum utilization. NOMA has been proposed as an alternative multiple access technique to improve spectral efficiency and support the increasing demand for connectivity in 5G and B5G networks. Unlike orthogonal schemes, NOMA allows multiple users to share the same time-frequency Resource Block (RB) by separating them in other domains, such as power or code. This enables more efficient utilization of available spectrum resources at the cost of increased receiver complexity [4].

NOMA techniques are broadly categorized into Power-Domain NOMA (PD-NOMA) and Code-Domain NOMA (CD-NOMA). In PD-NOMA, users are multiplexed based on different power levels, whereas CD-NOMA employs non-orthogonal spreading codes for user separation. In this dissertation, the focus is on PD-NOMA due to its practical relevance and compatibility with existing cellular systems [6].

In PD-NOMA, signals from multiple users are superimposed at the transmitter using different power allocation coefficients. Typically, users with weaker channel conditions are allocated higher transmission power, while users with stronger channel conditions are assigned lower power. At the receiver, signals are separated using SIC. In this process, users with stronger channel conditions first decode and remove the signals of users with weaker channels before decoding their own signals. In contrast, users with weaker channel conditions decode their signals directly while treating other signals as interference [2].

This decoding strategy enables efficient multi-user detection and improves SE compared to Orthogonal Multiple Access (OMA) schemes. In addition, NOMA enhances user fairness by allowing flexible power allocation among users sharing the same RB. It also improves system performance in scenarios with heterogeneous channel conditions, particularly for users located at the cell edge.

Despite these advantages, NOMA introduces several challenges, including interference management, power allocation, and the requirement for accurate SIC decoding. In particular, the performance of SIC depends on the Signal-to-Interference-plus-Noise Ratio (SINR), and decoding errors may propagate if interference is not properly removed. Therefore, ensuring reliable SIC operation is a critical design requirement in NOMA-based systems.

Due to its ability to improve spectral efficiency, support massive connectivity, and enhance fairness, NOMA has been widely studied for next-generation wireless networks. In this dissertation, NOMA is integrated with C-D2D communication, where resource allocation and power control must be carefully designed to satisfy both interference constraints and SIC decoding requirements.

1.7 From Cognitive to Intelligent Radio

The term *cognitive* implies awareness, perception, thinking, and decision-making. For a CR node to derive reasoning and judgment from its observations, it must incorporate learning. Learning enables present actions to be informed by past and current observations of the environment. In this context, ML empowers CR systems to recognize patterns and behavior across diverse network configurations, enhancing performance in novel or dynamic settings without requiring prior knowledge. This adaptability is a defining trait of intelligence.

Traditional optimization methods in CR lack the inherent flexibility to adapt to rapidly changing environments or unexpected scenarios. These methods typically rely on predefined models and mathematical approximations that may not capture system complexity accurately. In contrast, ML enables CR systems to learn directly from historical data, avoiding the limitations of approximations. This allows for the simultaneous optimization of multiple network parameters. However, ML approaches like RL do not always guarantee convergence to optimal performance. Factors such as available data, model selection, and algorithm complexity must be carefully considered to ensure effective deployment. The choice of a learning algorithm depends on the nature of the problem. In data-limited scenarios, real-time adaptable approaches like RL or online learning are preferable to offline-trained ML or deep learning models. Furthermore, the algorithm complexity should align with the capabilities of resource-constrained devices to ensure efficient performance [8].

1.8 Scope of the Work

Despite significant advances in spectrum sensing, resource allocation, and wireless communication technologies, several challenges remain in the design of efficient CR systems.

Existing approaches often address individual components of the problem in isolation, such as spectrum sensing, D2D communication, or multiple access techniques, without considering their joint interaction in realistic and dynamic environments. Based on the reviewed literature, several key limitations can be identified. First, there is a lack of integration between spectrum sensing, learning, and resource allocation mechanisms. Second, predictive intelligence is not fully exploited in dynamic spectrum access, limiting the ability of systems to anticipate channel availability. Third, many existing approaches rely on centralized optimization methods that incur high signaling overhead and lack scalability. In addition, there is limited investigation of decentralized multi-agent learning under realistic Channel State Information (CSI) constraints. Finally, the joint optimization of NOMA and C-D2D communication remains insufficiently explored. These limitations motivate the central problem addressed in this dissertation. Designing an intelligent, adaptive, and scalable resource allocation framework for CRNs that integrates predictive spectrum sensing, distributed multi-agent learning, and advanced spectrum sharing techniques under practical constraints such as limited CSI, dynamic environments, and interference coupling.

To address this problem, the main objectives of this dissertation are as follows.

- To develop Predictive-Cooperative Spectrum Sensing (PCSS) techniques that enhance spectrum awareness and reduce sensing overhead.
- To design distributed resource allocation frameworks using MADRL for C-D2D communication systems.
- To integrate NOMA with C-D2D networks and develop learning-based approaches for joint optimization of spectrum reuse and power allocation.
- To analyze system performance in terms of spectral efficiency, energy efficiency, fairness, and scalability.
- To provide a unified perspective on intelligent resource allocation in CRNs.

1.9 Dissertation Contributions and Organization

This dissertation presents a comprehensive framework for intelligent and adaptive resource allocation in CRNs by integrating predictive spectrum sensing, MADRL, and NOMA. The main contributions of this dissertation are summarized as follows.

Chapter 2 presents a novel PCSS framework to enhance spectrum awareness in CRNs. By integrating spectrum prediction with cooperative spectrum sensing, the proposed approach improves the accuracy of channel state estimation and enables more efficient spectrum utilization. In the proposed architecture, spectrum prediction is performed prior to sensing, allowing SUs to selectively sense channels that are more likely to be idle, thereby reducing unnecessary sensing operations. The framework leverages centralized prediction at the BS based on historical CSS decisions, which further enhances prediction reliability. Analytical and simulation results demonstrate that the proposed PCSS scheme significantly outperforms conventional sensing and single-user prediction methods, as well as CSP-based approaches, in terms of SE. The results also show that majority fusion achieves a favorable balance between detection accuracy and robustness. Furthermore, the chapter explicitly investigates the SE–EE tradeoff in high-traffic CRNs, highlighting the impact of cooperation overhead, sensing duration, and decision thresholds on system performance. To this end, a joint optimization of sensing time and final decision threshold is formulated to characterize the achievable SE–EE tradeoff. The analysis reveals that while cooperation improves sensing and prediction accuracy, it introduces additional energy consumption, leading to a fundamental tradeoff between SE and EE. However, the proposed framework assumes centralized fusion and coordination via a BS, which may limit scalability and adaptability in decentralized environments, thereby motivating the need for distributed and learning-based approaches in subsequent chapters.

Chapter 3 introduces a MADRL framework for distributed resource allocation in C-D2D networks. The joint power allocation and RB assignment problem is formulated as a Markov Decision Process (MDP) to maximize the sum rate of C-D2D links while satisfying CU rate constraints. Due to the non-convex MINLP nature of the problem, a model-free PPO-based MADRL approach is developed. Both Centralized Training with Decentralized Execution (CTDE) and Decentralized Training with Decentralized Execution (DTDE) frameworks are investigated under perfect and imperfect CSI. Each C-D2D pair acts as an agent that autonomously selects transmission power and spectrum resources based on locally observed energy levels, enabling distributed decision-making without global network knowledge. Simulation results demonstrate that the proposed MADRL-PPO framework significantly outperforms conventional optimization and existing DRL-based methods in terms of sum rate, scalability, and robustness, particularly under dynamic and interference-limited conditions. The study further examines the impact of

policy design (shared versus individual policies) and highlights the effectiveness of learning-based approaches for scalable resource allocation. Despite these gains, the framework reveals challenges related to training complexity, reward design, and coordination under partial observability. Additionally, it does not exploit advanced spectrum sharing techniques such as NOMA, which motivates the extension presented in Chapter 4.

Chapter 4 extends the proposed framework by integrating NOMA into C-D2D networks and investigates joint RB reuse and power allocation under CR protection and SIC decoding constraints. The resulting problem is highly non-convex due to coupled cross-tier and intra-group interference as well as mixed discrete–continuous decision variables. To address this, a multi-agent PPO-based framework is developed to enable distributed and scalable resource allocation. Two optimization objectives are considered: constrained sum-rate maximization and fairness-regularized sum-rate maximization. A structured observation-state design is introduced to support learning under partial CSI, and both CTDE and DTDE architectures are evaluated to examine the tradeoffs between coordination capability, scalability, and signaling overhead. Simulation results demonstrate that the proposed approach significantly outperforms conventional optimization and DRL baselines in terms of sum rate, fairness, energy efficiency, and reliability, while approaching centralized full-CSI performance with reduced signaling. The results further reveal a non-monotonic tradeoff between throughput and fairness, where moderate fairness regularization improves coordination and performance, whereas excessive weighting degrades overall efficiency.

Chapter 5 concludes the dissertation. The chapter also outlines future research directions, including extending the framework to mobile environments, incorporating energy-aware mechanisms, and integrating emerging technologies such as 6G and edge intelligence.

Chapter 2

Predictive-Cooperative Spectrum Sensing in Cognitive Radio Networks

CR has been proposed as an effective approach for sharing the available radio spectrum through dynamic spectrum allocation. The main objective is to improve the SE of CRNs and allow more devices to operate simultaneously. The SE of a CRN depends strongly on the employed SS technique. However, achieving reliable SS remains challenging because the sensing process must accurately determine the instantaneous state of the primary channel.

In local SS, each SU independently decides whether a PU signal is present using a sensing technique such as ED. The accuracy of local sensing is affected by the SNR of the corresponding PU signal [20]. Therefore, local sensing may suffer from unreliable detection, especially under low SNR, fading, and shadowing conditions. To mitigate the effects of channel variability, CSS has been proposed to improve sensing performance by allowing multiple SUs to collaborate during the sensing process [21]. In CSS, the local sensing decisions of multiple SUs are combined using fusion rules to obtain a final sensing decision [22]. Although CSS improves sensing reliability, it also introduces additional reporting overhead, energy consumption, and computational complexity.

Recently, PU channel statistics have been used to further improve CRN performance. Historical PU activity information can be exploited for spectrum prediction, ED threshold selection, or PU channel selection for sensing and access [20]. Spectrum prediction-based SS has been proposed to increase SE, reduce the collision rate, and mitigate sensing uncertainty, particularly in high-traffic environments [23]. In local spectrum prediction-based SS, prediction can be performed using ML techniques or HMMs. In this approach, an SU predicts the future states of the primary channels using historical information and then randomly selects a channel predicted to be idle for local SS [24].

However, local spectrum prediction may be inaccurate due to learning errors, imperfect sensing observations, and data inaccuracies in historical local sensing datasets. To address this issue, CSP-based SS has been introduced to improve prediction accuracy and SE. In CSP, each

SU performs local prediction, and the individual prediction results are combined to predict future channel availability [25]–[27]. Although CSP improves prediction accuracy, increasing the number of cooperative SUs increases overhead and consequently leads to higher network energy consumption.

Energy-efficient communication has become increasingly important with the growing interest in green communications. Reducing operating costs and supporting battery-operated devices are key challenges in the design of CRNs [28], [29]. EE, defined as the average SE per Joule of energy, is an important network performance metric [21]. However, EE and SE cannot generally be maximized simultaneously under the same constraints. Therefore, the joint optimization of EE and SE represents an important trade-off in the design of wireless communication networks [30]. In green CRNs, EE should be optimized while satisfying SE requirements to maintain the QoS of SUs.

This chapter presents the first contribution of this dissertation by proposing a PCSS framework to enhance spectrum awareness in CRNs. The proposed approach combines spectrum prediction with CSS to improve channel selection, reduce unnecessary sensing, and improve the SE–EE trade-off in high-traffic CRNs.

2.1 Related Work

The effectiveness of channel prediction for opportunistic channel access has recently been investigated. The importance of implementing intelligent spectrum sensing schemes has also been discussed in [31] to reduce sensing time and energy consumption. In [32], the SE performance of local spectrum prediction was examined by considering the impact of PU channel occupancy, prediction error, and the number of PU channels. A hierarchical deep learning model was designed in [31] for local spectrum prediction. In [24], a neural network-based multi-slot prediction and adaptive mode selection scheme was employed to improve SE in full-duplex CRNs. In [33], a prediction-based spectrum management strategy was proposed to enhance SE by jointly considering spectrum prediction, user mobility prediction, and channel selection.

In addition to SE improvement, other performance metrics have been considered to reduce spectrum handoff time and improve link stability and connection reliability. In [23], a hybrid spectrum access scheme based on local spectrum prediction was proposed for high-traffic CRNs

to improve SE and reduce waiting time. Hybrid spectrum access improves SE, while spectrum prediction mitigates sensing uncertainty in high-traffic environments. Estimation of PU activity statistics based on periodic channel observations with a finite sensing period was considered in [34]. In [20], the impact of sensing errors on the estimation of PU activity was studied.

Cooperative Spectrum Prediction (CSP) has also been investigated to improve prediction accuracy. In [25], CSP was proposed, and its prediction accuracy was examined under different PU channel occupancies. An optimal cooperative forecasting algorithm was further proposed to minimize cooperative prediction error. It was shown that prediction accuracy is influenced by the PU traffic pattern and that CSP improves prediction accuracy under most traffic conditions, but at the cost of increased computational complexity. In [26], the SE and EE of CSP in a CRN were investigated. The results showed a significant SE improvement using the majority fusion rule, but with a slight EE degradation compared to local spectrum prediction and local SS. The impact of busy-state and idle-state prediction errors on SE was also examined. However, the SE–EE trade-off in predictive CRNs was not considered in that study. In [27], CSP was proposed to improve detection performance, and SE was optimized by considering minimum rate requirements for each SU with beamforming.

Energy-constrained CRNs have become an important consideration for future wireless communication networks. Parameters such as sensing duration, transmit power, detection threshold, and the number of SUs have been considered in the design of energy-constrained CRNs. In [21], EE maximization with cluster-based CSS was formulated under collision and false alarm probability constraints. In [30], the EE–SE trade-off in a CRN-based cellular network with local SS was examined. The energy penalty of EH CRNs was investigated in [35] for joint CSP and CSS with an OR fusion rule.

Several studies have also optimized sensing-related parameters to improve SE and EE. In [36], the joint optimization of sensing time, detection threshold, and the selection of sensing and data-transmitting SUs for CSS was proposed to improve SE and minimize energy consumption in a CRN under detection and false alarm probability constraints. An optimal CSS strategy was developed in [22] for EH CRNs using a final decision threshold that maximizes SE subject to collision and energy constraints. In [37], the EE–SE trade-off with CSS was studied. EE was maximized under an SE constraint, and SE was maximized under an EE constraint, by jointly optimizing sensing duration and the final decision threshold. It was shown that the optimal sensing

duration and final decision threshold depend on the SE or EE requirements. However, spectrum prediction was not considered in this study. The EE–SE trade-off in an EH CSS CRN was investigated in [38].

Thus, the SE–EE trade-off in CRNs has been studied for CSS and local SS. However, the SE–EE trade-off in predictive CRNs has not been fully examined. In particular, existing works have not jointly considered spectrum prediction, cooperative sensing, sensing-time optimization, and final decision threshold optimization for balancing SE and EE in high-traffic CRNs. This motivates the PCSS framework proposed in this chapter.

2.2 Contributions of This Chapter

To improve the performance of CSS and exploit the benefits of both spectrum prediction and cooperative sensing, this chapter jointly employs these two techniques in the proposed PCSS framework. In the proposed scheme, spectrum prediction is performed first, followed by CSS based on the prediction results. This allows SUs to sense only PU channels that are likely to be available, rather than randomly sensing any PU channel.

The proposed PCSS scheme is coordinated through a central unit, such as a BS. The BS predicts future PU spectrum availability using previous CSS decisions. Since the prediction is based on cooperative sensing outcomes rather than individual local sensing results, the prediction accuracy can be improved, leading to better SE. In addition, integrating CSS within the prediction framework helps reduce sensing uncertainty caused by fading, shadowing, and local observation errors.

In high-traffic CRNs, random channel selection for SS often results in poor SE and EE because the probability of selecting a busy PU channel is high. This increases sensing energy consumption and leads to longer transmission waiting times [26]. Therefore, incorporating prediction with sensing is particularly useful in high-traffic scenarios. It allows SUs to focus sensing on channels that are more likely to be idle, which improves channel reuse while still protecting PU QoS.

The performance of the proposed PCSS scheme is investigated in high-traffic CRNs in terms of SE and EE. The proposed scheme is compared with well-known approaches in the literature, including CSP in [26]. Although spectrum prediction and CSP have been studied previously, the

SE–EE trade-off has not been fully considered in predictive CRNs. Motivated by prior work on the SE–EE trade-off with CSS [38]–[40], this chapter investigates the EE–SE trade-off in predictive CRNs with PCSS.

The main contributions of this chapter are summarized as follows.

- A novel PCSS scheme is proposed to enhance SE and EE by exploiting the benefits of spectrum prediction and CSS.
- The SE and EE of the proposed PCSS technique are analyzed and compared with well-known approaches in the literature, such as CSP in [26]. The results demonstrate that the proposed scheme significantly improves SE.
- The joint optimization of sensing time and final decision threshold is investigated for the SE–EE trade-off with PCSS in a high-traffic CRN. The achievable SE–EE trade-off performance of the proposed PCSS scheme is compared with CSS as in [37], showing that PCSS outperforms CSS.
- The impact of balancing SE and EE with PCSS is examined based on different network requirements.

2.3 System Model

A centralized architecture is considered where a BS controls the communications and spectrum management of M SUs through a dedicated control channel. A flowchart of the proposed PCSS scheme is given in Figure 2.1. The BS is responsible for predicting the spectrum availability of N PU channels, coordinating the SUs for local SS, collecting the local SS results from the SUs, combining these results (fusion), and broadcasting the decision to the SUs.

In the proposed PCSS framework, spectrum availability prediction is performed at the BS using the history of previous fusion decision results. When an SU requests information about future PU spectrum availability, the BS predicts the states of the PU channels and recommends the channels that are most likely to be idle. The SU then randomly selects one channel from the set of recommended channels for sensing.

After the channel is selected, the SUs participating in CSS sense the selected PU channel and report their local sensing results to the BS. The BS combines these results using a fusion rule to determine the final channel state. If the channel is declared idle, one SU is selected to access the PU channel according to the spectrum sharing algorithm.

By directing SUs to sense only channels that are predicted to be idle with high probability, the proposed scheme reduces unnecessary sensing operations. This saves both sensing time and SU energy, particularly in high-traffic CRNs where many PU channels are likely to be busy. However, the detailed design of spectrum management and spectrum sharing algorithms is beyond the scope of this chapter.

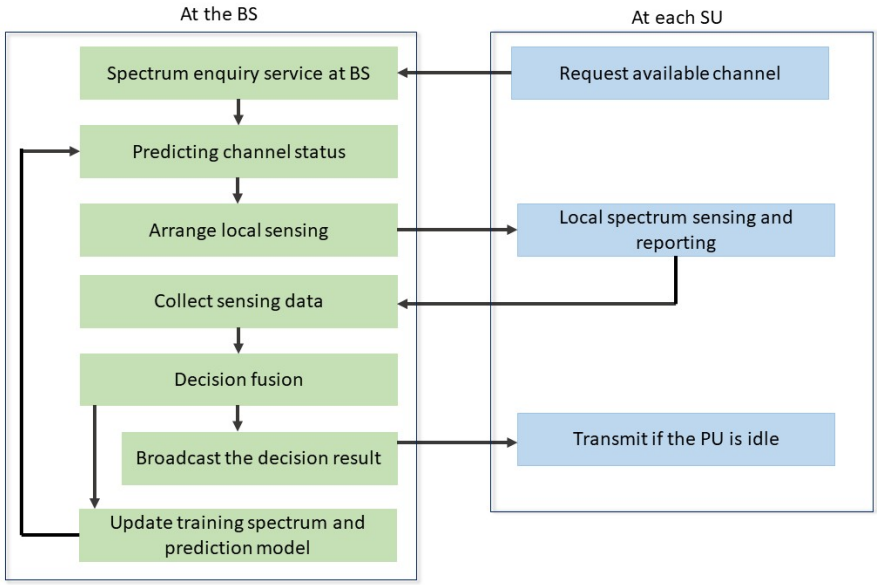


Figure 2.1: A flowchart of the PCSS scheme.

2.3.1 Local Spectrum Sensing with Energy Detection

In this chapter, PU transmissions are modeled as binary stochastic processes having a Poisson distributed arrival time with parameter λ and a binomial distributed holding time with parameter μ. Therefore, the PU channel is busy with probability $P(H_1)$ and idle with probability $P(H_0)$ where $P(H_1) = \mu/\lambda$ and $P(H_0) = 1 - \mu/\lambda$ [27]. The PUs are assumed to employ binary phase-shifted keying modulation, and the received signals are corrupted by AWGN with zero mean and variance σ^2. ED is considered for spectrum sensing due to its simplicity and ability to identify spectrum white spaces without requiring a priori PU information. ED is employed by each SU for local SS with sampling frequency f_s and sensing duration τ_s.

The SS is modeled as a binary hypothesis testing problem with two hypotheses H_0 and H_1 corresponding to the absence or presence of a PU signal, respectively [37, 40]. The local SS

performance is measured by the probability of false alarm P_f and probability of detection P_d. P_f is the probability of wrongly sensing an idle PU channel while P_d is the probability of correctly sensing a busy PU channel. The false alarm and detection probabilities are [37]

$$P_f = Q\left(\left(\frac{\epsilon}{\sigma^2} - 1\right)\sqrt{\frac{\tau_s f_s}{2}}\right) \tag{2.1}$$

$$P_d = Q\left(\left(\frac{\epsilon}{\sigma^2} - \gamma - 1\right)\sqrt{\frac{\tau_s f_s}{2(2\gamma+1)}}\right) \tag{2.2}$$

where ϵ is the ED threshold and γ is the SNR of the received PU signal when a PU is present. It is assumed that the SUs are located such that they experience similar path loss, so their SS performance in terms of P_f and P_d are the same. For a target probability of detection $\overline{P_d}$, the detection threshold ϵ can be expressed as $Q^{-1}(\overline{P_d}) = \left(\frac{\epsilon}{\sigma^2} - \gamma - 1\right)\sqrt{\frac{\tau_s f_s}{2(2\gamma+1)}}$. On the other hand, this threshold is related to the probability of false alarm according to $Q^{-1}(P_f) = \left(\frac{\epsilon}{\sigma^2} - 1\right)\sqrt{\frac{\tau_s f_s}{2}}$. Therefore, for a target probability of detection $\overline{P_d}$, the probability of false alarm is

$$P_f = Q\left(\sqrt{2\gamma + 1}\, Q^{-1}(\overline{P_d}) + \gamma\sqrt{\frac{\tau_s f_s}{2}}\right) \tag{2.3}$$

2.3.2 Cooperative Spectrum Sensing

In this chapter, periodic CSS is employed with frame duration T consisting of an SS duration τ_s, reporting duration τ_r for each SU, and data transmission duration $T - \tau_s - M\tau_r$. The M cooperating SUs report their one-bit binary decisions, where 0 represents the idle state and 1 represents the busy state.

In this case, the BS acts as a Fusion Center (FC). It is assumed that each SU transmits their one-bit decision over a dedicated, error-free channel to employ the hard fusion. In hard fusion, each SU first makes a local SS decision and then transmits this decision to the FC. After collecting all local SS decisions from the SUs, the final decision on the presence/absence of primary signals is made by the FC [40].

In this chapter, a majority fusion rule is used to evaluate the performance of the proposed PCSS scheme. This rule provides a tradeoff between the performance of the OR and AND fusion

rules. With the majority fusion rule, the FC makes the decision that the PU is present if $M/2$ or more SUs reported 1 [27]. With the OR fusion rule, the FC makes the decision that the PU is present if at least one SU reported 1. With the AND fusion rule, the FC makes the decision that the PU is present if all SUs reported 1 [36].

Later in the chapter, a *K* out of *M* fusion rule is used when the SE-EE tradeoff is considered. The BS makes the final decision and broadcasts this result to the cooperating SUs. The final detection and false alarm probabilities with CSS are then [37]

$$Q_d = I_{P_d}(K, M - K + 1) \tag{2.4}$$

$$Q_f = 1 - I_{1-P_f}(K, M - K + 1) \tag{2.5}$$

where K is the final decision threshold in the FC and $y = I_x(a, b)$ is the regularized incomplete beta function [38]. The joint probabilities of true channel state and cooperative sensing result are given in Table 2.1.

Table 2.1: Joint probabilities of true and cooperative channel states

True Channel State	Cooperative Sensing Result	Probability
Idle	Idle	$P(H_0)(1 - Q_f)$
Idle	Busy	$P(H_0)\, Q_f$
Busy	Idle	$P(H_1)(1 - Q_d)$
Busy	Busy	$P(H_1)\, Q_d$

2.3.3 Spectrum Prediction

The FC results are a binary sequence for each sensed PU channel. With a sufficiently long history of PU channel usage, PU activity patterns can be characterized and then prediction can be performed [42]. HMM and neural networks are the most widely used prediction methods [27].
In this chapter, HMM-based spectrum prediction is employed to predict the future availability of the N PU channels at the BS after combining the sensing results from the M SUs as shown in Figure 2.2. In the proposed PCSS-based HMM framework, the true PU channel state is treated as a hidden state, while the FC fusion decisions are used as the observable sequence.

The objective of the HMM is to predict future PU channel states by learning the state transition probability matrix and the emission probability matrix, as in [26], [39]. The transition probability matrix captures the temporal behavior of the PU channel states, while the emission probability matrix describes the relationship between the hidden true channel states and the observed fusion decisions.

In this case, the observation sequence is the last L fusion decision results from the M cooperating SUs at the FC. The emission matrix e is a function of the final detection and false alarm probabilities of the CSS and is given by

$$e = \begin{bmatrix} 1 - Q_f & Q_f \\ 1 - Q_d & Q_d \end{bmatrix} \tag{2.6}$$

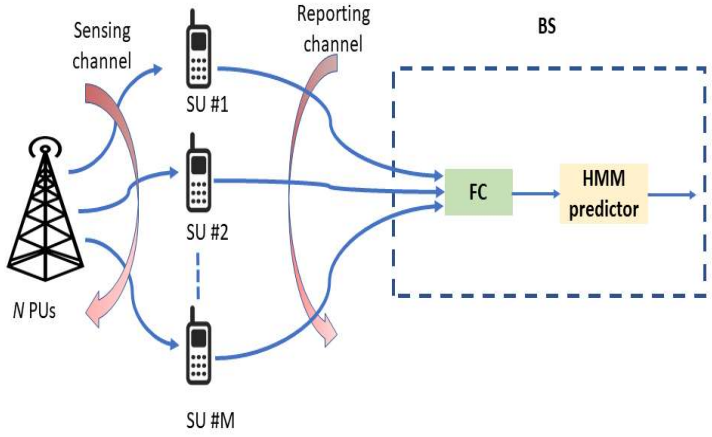


Figure 2.1: The PCSS model.

The HMM prediction performance is evaluated by the total probability of prediction error P_e^p, which is the probability of wrongly predicting the PU channel state. The joint probabilities of true channel state and prediction result are given in Table 2.2. From Table 2.2, the overall probabilities of predicting the PU channel to be idle P_0^p or busy P_1^p are

$$P_0^p = P(H_0)\left(1 - P_e^p\right) + P(H_1)P_e^p \tag{2.7}$$

$$P_1^p = P(H_0)\, P_e^p + P(H_1)(1 - P_e^p) \tag{2.8}$$

Table 2.2: The joint probabilities of true and prediction channel states

True Channel State	Prediction Result	Probability
Idle	Idle	$P(H_0)(1 - P_e^p)$
Idle	Busy	$P(H_0)\, P_e^p$
Busy	Idle	$P(H_1)P_e^p$
Busy	Busy	$P(H_1)(1 - P_e^p)$

In this chapter, the BS performs spectrum prediction on the N PU channels based on past activity and SU sensing of randomly selected PU channels among those predicted to be idle. Thus, the event that v PUs are predicted to be idle can be seen as v repeated Bernoulli trials, which has probability $C_N^v\ (P_0^p)^v\ (P_1^p)^{N-v}$ where $v \leq N$.

Therefore, the probability of predicting at least one PU channel as idle is $P_0^N = \sum_{v=1}^{N} C_N^v\ (P_0^p)^v\ (P_1^p)^{N-v}$ and the probability of predicting that every channel is busy is $P_1^N = (P_1^p)^N$ [4, 5, 7]. The joint probabilities of true channel state, and the prediction and cooperative sensing results are given in Table 2.3.

Table 2.3: The joint probabilities of true channel state, and the prediction and cooperative sensing results.

True Channel State	Prediction Result	Cooperative Sensing Result	Probability
Idle	Idle	Idle	$P_1 = \frac{P(H_0)(1-Q_f)\left(1-P_e^p\right)}{P_0^p/P_0^N}$
Idle	Idle	Busy	$P_2 = \frac{P(H_0)Q_f\left(1-P_e^p\right)}{P_0^p/P_0^N}$
Idle	Busy	Idle	$P_3 = \frac{P(H_0)(1-Q_f)P_e^p}{P_1^p/P_1^N}$
Idle	Busy	Busy	$P_4 = \frac{P(H_0)Q_fP_e^p}{P_1^p/P_1^N}$
Busy	Idle	Idle	$P_5 = \frac{P(H_1)(1-Q_d)P_e^p}{P_0^p/P_0^N}$
Busy	Idle	Busy	$P_6 = \frac{P(H_1)Q_dP_e^p}{P_0^p/P_0^N}$
Busy	Busy	Idle	$P_7 = \frac{P(H_1)(1-Q_d)\left(1-P_e^p\right)}{P_1^p/P_1^N}$
Busy	Busy	Busy	$P_8 = \frac{P(H_1)Q_d\left(1-P_e^p\right)}{P_1^p/P_1^N}$

2.3.4 SE and EE Based Predictive-Cooperative Spectrum Sensing

In the proposed scheme, an SU only senses one channel from the channels that are predicted to be idle and transmits if the cooperative sensing result is idle. The SU then transmits data in the following two cases: (1) when the PU is absent and it is correctly predicted and sensed to be absent; and (2) when the PU is present and it is falsely predicted and sensed to be absent. Therefore, the average SE of the PCSS scheme can be expressed as

$$\eta_{SE} = \widetilde{\eta_{SE}} + \overline{\overline{\eta_{SE}}} \tag{2.9}$$

where $\widetilde{\eta_{SE}} = P_1 C_0$ and $\overline{\overline{\eta_{SE}}} = P_5 C_1$, and $C_0 = \frac{T-\tau_s-M\tau_r}{T} \log_2(1+\gamma_s)$ and $C_1 = \frac{T-\tau_s-M\tau_r}{T} \log_2\left(1+\frac{\gamma_s}{1+\gamma}\right)$ are the throughput of the SUs when they transmit in the absence and presence of PUs, respectively, where γ_s is the SU SNR. It is clear that $C_0 > C_1$ as the SU transmissions will be corrupted by interference when a PU is present [32].

The average energy consumption E for the PCSS scheme can be expressed as a function of time and power consumption. Let P_s, P_r , P_c and P_t be the SU sensing power consumption, SU reporting power consumption, SU electronics power consumption, and the SU transmit power, respectively. Then the average SU energy consumption is

$$E = M\tau_s P_s(P_1 + P_2 + P_5 + P_6) + M\tau_r P_r + MP_c(\tau_s + \tau_r) + (T - \tau_s - M\tau_r)(P_t + P_c)(P_1 + P_5) \quad (3.10)$$

The EE (measured in bits/Joule/Hz) for the PCSS scheme can be expressed as

$$\eta_{EE} = \frac{\eta_{SE}\, T}{E} = \widetilde{\eta_{EE}} + \overline{\overline{\eta_{EE}}} \quad (2.11)$$

where $\widetilde{\eta_{EE}} = \frac{\widetilde{\eta_{SE}}\, T}{E}$ and $\overline{\overline{\eta_{EE}}} = \frac{\overline{\overline{\eta_{SE}}}\ T}{E}$.

2.4 Problem Formulation and Solution

In this section, we consider the optimization of EE and SE with PCSS to obtain an EE-SE tradeoff while providing an acceptable PU QoS. Hence, the general optimization problem is formulated as

$$\max_{\tau_s,\ K} \ \rho\, \eta_{SE} + (1-\rho)\eta_{EE} \quad (2.12)$$

$$s.t. \qquad Q_d \geq Q_d^{th} \quad (2.12a)$$

$$0 < \tau_s < T - M\tau_r \quad (2.12b)$$

$$1 \leq K \leq M \quad (2.12c)$$

where ρ is the balancing factor. When $\rho = 1$, the problem becomes an SE maximization problem $\max_{\tau_s,\ K} \eta_{SE}$, and when $\rho = 0$, the problem becomes an EE maximization problem $\max_{\tau_s,\ K} \eta_{EE}$. When $0 < \rho < 1$, it is a joint EE and SE maximization problem. According to the IEEE802.22 WRAN standard, in order to maintain PU QoS, the target detection probability should be $Q_d^{th} = 0.9$ [32].

The objective is to design the sensing duration and final decision threshold to provide a balance between SE and EE. In a high traffic CRN, selecting channels randomly to perform SS results in low SE and high energy consumption. It was observed in [42] that SUs have a low SE when the PU traffic load is high. However, the SE can be improved, and the power consumption reduced, by having SUs only sense channels that are predicted to be idle in the next time slot [26]. To evaluate the performance of the proposed PCSS scheme, the probability of a busy PU channel is set to $P(H_1) = 0.7$ with a target detection probability $Q_d = Q_d^{th} = 0.9$. In this case, $\overline{\overline{\eta_{SE}}}$ can be ignored as $P_5 < P_1$ from Table 2.3 and $C_0 > C_1$. Therefore, the average SE can be approximated as $\eta_{SE} = \widetilde{\eta_{SE}}$ and the average EE as $\eta_{EE} = \widetilde{\eta_{EE}}$.

In the HMM algorithm, P_e^p is a function of τ_s as the emission matrix e is a function of Q_f and Q_d and the transition matrix is a function of $P(H_0)$ and $P(H_1)$ [23]. Thus, a closed-form expression of P_e^p cannot be obtained.

Therefore, P_e^p is assumed to be a constant as in [23, 32, 35]. Thus, the general optimization problem can be approximated as

$$\max_{\tau_s,\ K} \quad \rho\, \widetilde{\eta_{SE}} + (1-\rho)\, \widetilde{\eta_{EE}} \tag{2.13}$$

$$s.t. \qquad Q_d \geq Q_d^{th} \tag{2.13a}$$

$$0 < \tau_s < T - M\tau_r \tag{2.13b}$$

$$1 \leq K \leq M \tag{2.13c}$$

where $\widetilde{\eta_{SE}} = P_1 C_0$ and $\widetilde{\eta_{EE}} = \frac{\widetilde{\eta_{SE}}\, T}{E}$. When $\rho = 1$, the problem becomes an SE maximization problem $\max_{\tau_s,\ K} \quad \widetilde{\eta_{SE}}$, and when $\rho = 0$, the problem becomes an EE maximization problem $\max_{\tau_s,\ K} \widetilde{\eta_{EE}}$, subject to $0 < \tau_s < T - M\tau_r$ and $1 \leq K \leq M$.

Two single variable subproblems are considered to solve these two variable optimization problems. Then the steps for solving these subproblems are combined to obtain Algorithm 3.2. The following sections present the optimal solution of these subproblems.

2.4.1 SE Optimization Solution

For given values of ρ and $K = \widetilde{K}$, the value of τ_s that maximizes $\rho\, \eta_{\mathrm{SE}} + (1-\rho)\eta_{\mathrm{EE}}$ and meets the target probability of detection Q_d^{th} is given by

Subproblem 1

$$\max_{\tau_s} \quad \rho\, \widetilde{\eta_{SE}}\ (\tau_s, K) + (1-\rho)\, \widetilde{\eta_{EE}}\ (\tau_s, K)\,|_{K=\widetilde{K}} \tag{2.14}$$

$$s.t. \qquad 0 < \tau_s < T - M\tau_r \tag{2.14a}$$

where $\widetilde{\eta_{SE}} = \frac{P(H_0)(1-Q_f(\tau_s))\left(1-P_e^p\right)}{P_0^p/P_0^N}\ \frac{T-\tau_s-M\tau_r}{T}\ \log_2(1+\gamma_s)$ and $\widetilde{\eta_{EE}} = \frac{\widetilde{\eta_{SE}}\,T}{E}$. For $K = \widetilde{K}$, the partial derivative of (16) with respect to τ_s is $\rho\frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s} + (1-\rho)\ \frac{\partial\widetilde{\eta_{EE}}}{\partial\tau_s}$ and $\frac{\partial\widetilde{\eta_{EE}}}{\partial\tau_s} = \frac{\partial(\frac{\widetilde{\eta_{SE}}\,T}{E})}{\partial\tau_s} = \frac{T}{E^2}\left(E\ \frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s} - \widetilde{\eta_{SE}}\frac{\partial E}{\partial\tau_s}\right)$. Let U be the partial derivative of (14) with respect to τ_s

$$U = \frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s}\left(\rho + (1-\rho)\frac{T}{E}\right) - \frac{\partial E}{\partial\tau_s}\widetilde{\eta_{SE}}\frac{T}{E^2}(1-\rho) \tag{2.15}$$

Then, the partial derivative of $\widetilde{\eta_{SE}}$ with respect to τ_s is

$$\frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s} = \frac{\partial(P_1C_0)}{\partial\tau_s} = C_0\frac{\partial P_1}{\partial\tau_s} + P_1\frac{\partial C_0}{\partial\tau_s} \tag{2.16}$$

$$\frac{\partial P_1}{\partial\tau_s} = -\ \frac{P(H_0)\left(1-P_e^p\right)}{P_0^p/P_0^N}\ \frac{\partial Q_f}{\partial\tau_s} \tag{2.17}$$

$$\frac{\partial C_0}{\partial\tau_s} = \frac{-1}{T}\ \log_2(1+\gamma_s) \tag{2.18}$$

$$\frac{\partial Q_f}{\partial\tau_s} = \frac{\partial Q_f}{\partial P_f}\ \frac{\partial P_f}{\partial\tau_s} = M\binom{M-1}{K-1}{P_f}^{K-1}(1-P_f)^{M-K}\frac{\partial P_f}{\partial\tau_s} \tag{2.19}$$

$$\frac{\partial P_f}{\partial\tau_s} = -\frac{\gamma}{4}\sqrt{\frac{f_s}{4\pi\tau_s}}\ \ e^{-\frac{1}{2}\left(\sqrt{2\gamma+1}\,Q^{-1}(\overline{P_d})+\gamma\sqrt{\frac{\tau_s f_s}{2}}\right)^2} \tag{2.20}$$

The partial derivative of E with respect to τ_s is

$$\frac{\partial E}{\partial\tau_s} = MP_s(P_1+P_2+P_5+P_6) + M\tau_sP_s\left(\frac{\partial P_1}{\partial\tau_s}+\frac{\partial P_2}{\partial\tau_s}\right) + MP_c - (P_t+P_c)(P_1+P_5) + (T-\tau_s-M\tau_r)(P_t+P_c)\frac{\partial P_1}{\partial\tau_s} \tag{2.21}$$

where $\frac{\partial P_2}{\partial\tau_s} = \frac{P(H_0)\left(1-P_e^p\right)}{P_0^p/P_0^N}\ \frac{\partial Q_f}{\partial\tau_s}$. It can be shown following [17] that $\lim_{\tau_s\to 0}\frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s} = +\infty$, $\lim_{\tau_s\to T-M\tau_r}\frac{\partial\widetilde{\eta_{SE}}}{\partial\tau_s} < 0$, $\lim_{\tau_s\to 0}\frac{\partial\widetilde{\eta_{EE}}}{\partial\tau_s} = +\infty$ and $\lim_{\tau_s\to T-M\tau_r}\frac{\partial\widetilde{\eta_{EE}}}{\partial\tau_s} < 0$ [37]. Thus, $\lim_{\tau_s\to 0}\frac{\partial(\rho\,\widetilde{\eta_{SE}}+(1-\rho)\,\widetilde{\eta_{EE}})}{\partial\tau_s} = +\infty$ and $\lim_{\tau_s\to T-M\tau_r}\frac{\partial(\rho\,\widetilde{\eta_{SE}}+(1-\rho)\,\widetilde{\eta_{EE}})}{\partial\tau_s} < 0$, so a unique value of τ_s which

maximizes $\rho\,\widetilde{\eta_{SE}} + (1-\rho)\,\widetilde{\eta_{EE}}$ exists in the range $0 < \tau_s < T - M\tau_r$. Therefore, the bisection method is used in Algorithm 2.1 to obtain the optimal τ_s that satisfies $\frac{\partial\,(\,\rho\,\widetilde{\eta_{SE}}+(1-\rho)\,\widetilde{\eta_{EE}}\,)}{\partial\tau_s} = 0$.

2.4.2 EE Optimization Solution

For given values of ρ and $\tau_s = \widetilde{\tau_s}$, the optimal value of K that maximizes $\rho\,\eta_{\mathrm{SE}} + (1-\rho)\eta_{\mathrm{EE}}$ and meets the target probability of detection Q_d^{th} is given by

Subproblem 2

$$\max_{K} \quad \rho\,\widetilde{\eta_{SE}}\;(\tau_s, K) + (1-\rho)\,\widetilde{\eta_{EE}}\;(\tau_s, K)\,|_{\tau_s=\widetilde{\tau_s}} \tag{2.22}$$

$$s.t. \quad 1 \le K \le M \tag{2.22a}$$

For $\tau_s = \widetilde{\tau_s}$, the partial derivative of $\widetilde{\eta_{\mathrm{SE}}}$ with respect to K is

$$\frac{\partial\widetilde{\eta_{\mathrm{SE}}}}{\partial K} = \frac{\partial(P_1 C_0)}{\partial K} = C_0\frac{\partial \mathrm{P_1}}{\partial \mathrm{K}} \tag{2.23}$$

$$\frac{\partial P_1}{\partial K} = -\frac{P(H_0)\left(1-P_e^p\right)}{P_0^p/P_0^N}\frac{\partial Q_f}{\partial K} \tag{2.24}$$

and the partial derivative of $\widetilde{\eta_{\mathrm{EE}}}$ with respect to K is

$$\frac{\partial\widetilde{\eta_{EE}}}{\partial K} = = \frac{T}{E^2}\left(E\,\frac{\partial\widetilde{\eta_{SE}}}{\partial K} - \widetilde{\eta_{SE}}\,\frac{\partial E}{\partial K}\right) \tag{2.25}$$

where $\frac{\partial E}{\partial K} = -(T-\tau_s-M\tau_r)(P_t+P_c)\,\frac{P(H_0)\left(1-P_e^p\right)}{P_0^p/P_0^N}\,\frac{\partial Q_f}{\partial K}$. Thus, $\frac{\partial\,(\,\rho\,\widetilde{\eta_{SE}}+(1-\rho)\,\widetilde{\eta_{EE}}\,)}{\partial K} = \frac{\partial Q_f}{\partial K}\,C_0\,\frac{P(H_0)\left(1-P_e^p\right)}{P_0^p/P_0^N}\left(-\rho - \frac{T}{E} + \frac{T}{E^2}P_1(T-\tau_s-M\tau_r)(P_t+P_c)(1-\rho)\right)$. The optimal value of K satisfies $\frac{\partial\,(\,\rho\,\widetilde{\eta_{SE}}+(1-\rho)\,\widetilde{\eta_{EE}}\,)}{\partial K} = 0$, and results in $\frac{\partial Q_f}{\partial K} = 0$. Therefore, Subproblem 2 reduces to

$$\text{Subproblem 2} \quad \min_{K} \quad Q_f\;(\tau_s, K)\big|_{\tau_s=\widetilde{\tau_s}} \tag{2.26}$$

$$s.t. \quad 1 \le K \le M \tag{2.26a}$$

where $Q_f(K) = 1 - I_{1-P_f(K)}(K, M-K+1)$ and $P_f(K) = Q\left(\sqrt{2\gamma+1} + Q^{-1}(P_d(K)) + \gamma\sqrt{\frac{t_s f_s}{2}}\right)$.

For a given $Q_d = Q_d^{th}$, $P_d(K) = I^{-1}\left(Q_d^{th}, K, M-K+1\right)$ where $x = I^{-1}(y,a,b)$ is the inverse regularized incomplete beta function. There is no closed-form solution for the optimum value K, but an exhaustive search over all values can be employed as K is an integer in the range 1 to M. Combining the steps for solving the two subproblems gives the iterative algorithm in Algorithm 2.2.

2.4.3 Maximizing the EE while Satisfying SE Requirements

To guarantee the QoS of SUs that have limited bandwidth and transmit power, it is reasonable to maximize the EE while satisfying minimum SE requirements. In this case, the general optimization problem is

$$\max_{\tau_s,\ K} \quad \widetilde{\eta_{EE}} \tag{2.27}$$

$$S.t \quad Q_d \geq Q_d^{th} \tag{2.27a}$$

$$\widetilde{\eta_{SE}} \geq \widetilde{\eta_{SE}}^{th} \tag{2.27b}$$

$$0 < \tau_s < T - M\tau_r \tag{2.27c}$$

$$1 \leq K \leq M \tag{2.27d}$$

where $\widetilde{\eta_{SE}} = \frac{P(H_0)(1-Q_f)\left(1-P_e^p\right)}{P_0^p/P_0^N} \frac{T-\tau_s-M\tau_r}{T} \log_2(1+\gamma_s)$ and $\widetilde{\eta_{SE}}^{th}$ is the SE threshold. When $\widetilde{\eta_{SE}} = \widetilde{\eta_{SE}}^{th}$, there will be two roots ${\tau_s}^{\text{right}}{}_{SE}$ and ${\tau_s}^{\text{left}}{}_{SE}$ for $\widetilde{\eta_{SE}} - \widetilde{\eta_{SE}}^{th} = 0$, where ${\tau_s}^{\text{right}}{}_{SE}$ has range $0 < {\tau_s}^{\text{right}}{}_{SE} < {\tau_s}^{*}{}_{SE}$ and ${\tau_s}^{\text{left}}{}_{SE}$ has range ${\tau_s}^{*}{}_{SE} < {\tau_s}^{\text{left}}{}_{SE} < T - M\tau_r$. ${\tau_s}^{*}{}_{SE}$ is the optimal sensing time that maximizes the SE when $\rho = 1$.

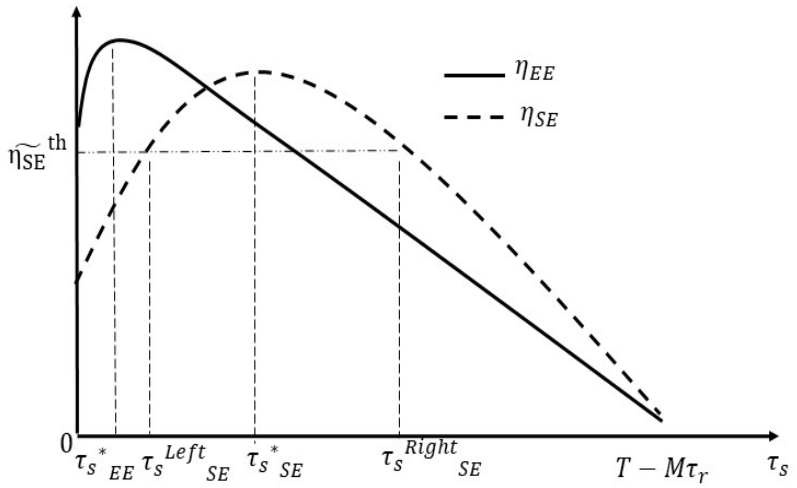


Figure 2.3: SE and EE for a CRN with high SE requirements $\widetilde{\eta_{SE}} \geq \widetilde{\eta_{SE}}^{th}$.

Figure 2.3 shows the SE and EE versus τ_s for a CRN with high SE requirements. In this case, $\widetilde{\eta_{SE}}(\tau_s^*{}_{EE}) \leq \widetilde{\eta_{SE}}^{th}$, and the two roots of $\widetilde{\eta_{SE}} - \widetilde{\eta_{SE}}^{th} = 0$ are $\tau_s{}^{\text{right}}{}_{SE}$ and $\tau_s{}^{\text{left}}{}_{SE}$ where $\tau_s^*{}_{EE}$ is the optimal sensing time that maximizes the EE when $\rho = 0$. To maximize the EE while satisfying the SE requirements, the optimal sensing time can be determined by comparing $\widetilde{\eta_{EE}}(\tau_s{}^{\text{right}}{}_{SE})$ and $\widetilde{\eta_{EE}}(\tau_s{}^{\text{left}}{}_{SE})$. When $\widetilde{\eta_{EE}}(\tau_s{}^{\text{left}}{}_{SE}) > \widetilde{\eta_{EE}}(\tau_s{}^{\text{right}}{}_{SE})$, $\tau_s{}^{\text{left}}{}_{SE}$ is chosen as the optimal sensing time and when $\widetilde{\eta_{SE}}(\tau_s^*{}_{EE}) > \widetilde{\eta_{SE}}^{th}$, $\tau_s^*{}_{EE}$ is chosen. The algorithm that maximizes the EE while satisfying the SE requirements is given in Algorithm 2.3.

Algorithm 2.1: Optimal Sensing Duration

1: Set $\tau_s(0) = 0$, $\tau_s(f) = T - M\tau_r$, $\tau_s(m) = \frac{\tau_s(f)-\tau_s(0)}{2}$ and the tolerance level Δ

2: Determine the value of U in (15) at $\tau_s(0)$ and at $\tau_s(f)$.

3: Repeat

4: if $U\big(\tau_s(m)\big) < 0$ **then**

5: Set $\tau_s(f) = \tau_s(m)$

6: End

7: if $U\big(\tau_s(m)\big) > 0$ **then**

8: Set $\tau_s(0) = \tau_s(m)$

9: End

10: Determine $\tau_s(m) = \frac{\tau_s(f)-\tau_s(0)}{2}$

11: Until $\big|U\big(\tau_s(m)\big)\big| < \Delta$

12: Output $\tau_s(m) = {\tau_s}^*{}_{SE\_EE}$

$*$ denotes optimal value.

Algorithm 2.2: Optimizing the SE and EE

1: Set ρ where $0 < \rho < 1$ (ρ = 0 for EE maximization and ρ = 1 for SE maximization)

2: For $i = 1$ to M

3: Set $K = i$

4: Find ${\tau_s}^*{}_{SE\_EE}(i)$ from (15) using Algorithm 1

5: Determine $V(i) = \rho\, \widetilde{\eta_{SE}}(i) + (1-\rho)\, \widetilde{\eta_{EE}}(i)$ with ${\tau_s}^*{}_{SE\_EE}$.

6: End

7: Choose $K^* = i$, such that $V(K^*)$ is the maximum of $[V(i)]$ by exhaustive search for i = 1 to M.

8: Determine $\widetilde{\eta_{SE}}$ and $\widetilde{\eta_{EE}}$ with ${\tau_s}^*{}_{SE\_EE}$ and K^*.

9: Output ${\tau_s}^*{}_{SE-EE}(K^*)$, K^*, $\widetilde{\eta_{SE}}\big({\tau_s}^*{}_{SE-EE}, K^*\big)$, $\widetilde{\eta_{EE}}({\tau_s}^*{}_{SE-EE}, K^*)$

$*$ denotes optimal value.

Algorithm 2.3: Maximizing the EE while satisfying the SE requirements

1: For $K = 1$ to M

2: Find ${\tau_s}^*{}_{SE}(K)$ from (15) with $\rho = 1$ that maximizes $\widetilde{\eta_{SE}}$ using the bisection method

3: Find the roots ${\tau_s}^{\text{right}}{}_{SE}(K)$ and ${\tau_s}^{\text{left}}{}_{SE}(K)$ of $\widetilde{\eta_{SE}} - \widetilde{\eta_{SE}}^{th} = 0$ using the bisection method in $\big(0, {\tau_s}^*{}_{SE}\big)$ and $\big({\tau_s}^*{}_{SE}, T - M\tau_r\big)$, respectively

4: Find ${\tau_s}^*{}_{EE}(K)$ from (15) with $\rho = 0$ that maximizes $\widetilde{\eta_{EE}}$ using the bisection method

5: If ${\tau_s}^*_{EE}(K) < {\tau_s}^{\text{left}}_{SE}(K)$ **then**

6: If $\widetilde{\eta_{EE}}({\tau_s}^{\text{left}}_{SE}) > \widetilde{\eta_{EE}}({\tau_s}^{\text{right}}_{SE})$ **then**

7: The optimal sensing time is ${\tau_s}^{\text{left}}_{SE}$

8: Else

9: The optimal sensing time is ${\tau_s}^{\text{right}}_{SE}$

10: End

11: The optimal sensing time is ${\tau_s}^*_{EE}$

12: End

13: Calculate $\widetilde{\eta_{EE}}(K)$ for the optimal sensing time

14: End

15: Choose K^* such that $\widetilde{\eta_{EE}}(K^*)$ is the maximum of $[\widetilde{\eta_{EE}}(K)]$ by exhaustive search for K= 1 to M

16: Output $\widetilde{\eta_{EE}}(K^*)$

$*$ denotes optimal value.

2.5 Performance Evaluation

In this section, the SE-EE tradeoff with the PCSS scheme in a high traffic CRN is evaluated. The simulation parameters are summarized in Table 2.4. To evaluate the prediction error for the local spectrum prediction, PCSS, and CSP schemes, two HMMs were trained using L=100 time frames and tested on 30,000 frames for $P(H_0) = 0.3$.

The first HMM is used with the local spectrum prediction and CSP schemes, while the second is used with the PCSS scheme. For the first HMM, the input is the last L SS results from local spectrum prediction with $P_d = 0.9$ in the emission matrix e.

For the second HMM, the input is the last L CSS results using the majority fusion rule with $Q_d = 0.9$ in matrix e. The prediction error of the CSP scheme is the fusion of M local spectrum prediction results using the majority fusion rule as in [26]. Q_f and P_f in matrix e change as τ_s varies in the range $0 < \tau_s < T - M\tau_r$.

Table 2.4: Simulation parameters

Parameter	Value	Parameter	Value
f_s	10 kHz	γ	-12 dB
M	8	γ_s	5 dB

$P(H_0)$	0.3	P_s	40 mW
Q_d^{th}	0.9	P_r	10 mW
T	100 ms	P_p	20 mW
τ_r	0.1 ms	P_c	1.8 W
τ_p	5 ms	N	10

Figure 2.4 shows the prediction error P_e^p with the three schemes versus the sensing time. The prediction error of the CSP scheme is lowest when $\tau_s < 30$ ms. For $Q_d = Q_d^{th} = 0.9$ and $P(H_0) = 0.3$, the probability of being in a specific state given the previous observations becomes a function of Q_f according to (3.11) in [43]. Therefore, P_e^p is monotonically decreasing from 0.3 to 0.07 as τ_s increases. As mentioned previously, it is intractable to find a closed-form solution for P_e^p in the HMM algorithm as a function of τ_s. In [35], P_e^p was assumed to be a constant versus the sensing time for all SUs and two values $P_e^p = 0.1$ and $P_e^p = 0.01$ were considered. In [23, 32], P_e^p was assumed to be a constant versus the PU channel occupancy, and $P_e^p = 0.05$ and 0.2 were considered in [32]. Different values of P_e^p between 0.05 and 0.9 were considered in [23]. In this chapter, $P_e^p = 0.2$ is used to investigate the SE-EE tradeoff with the PCSS scheme.

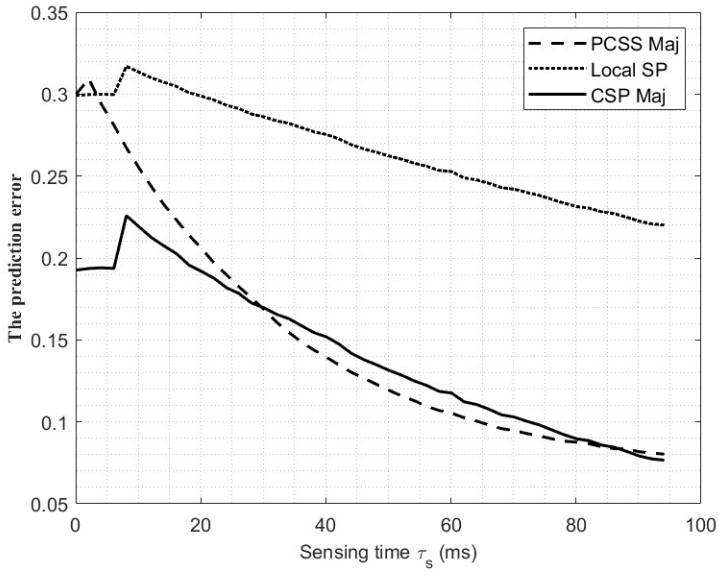


Figure 2.4: The prediction error P_e^p versus the sensing time.

Figure 2.5 shows the SE performance versus the sensing time τ_s with PCSS and majority fusion rule, local SS, local spectrum prediction, CSS with majority fusion rule, and CSP with majority rule. The prediction error values in Figure 2.4 were used to obtain these results. Figure 2.5 shows that there is a significant improvement in the SE with PCSS compared to the other schemes. The benefits of combining prediction with CSS in the PCSS scheme can be seen by comparing the results with those for CSS with majority fusion rule. The improvement in SE with PCSS over CSP [24] is almost a factor of 2. This is because of the improved SS performance in terms of Q_f with PCSS compared to P_f with CSP. The SE with local spectrum prediction is better than with local SS. Further, the SE with CSS and majority fusion rule is better than with CSP and majority fusion rule for $\tau_s > 5$ ms.

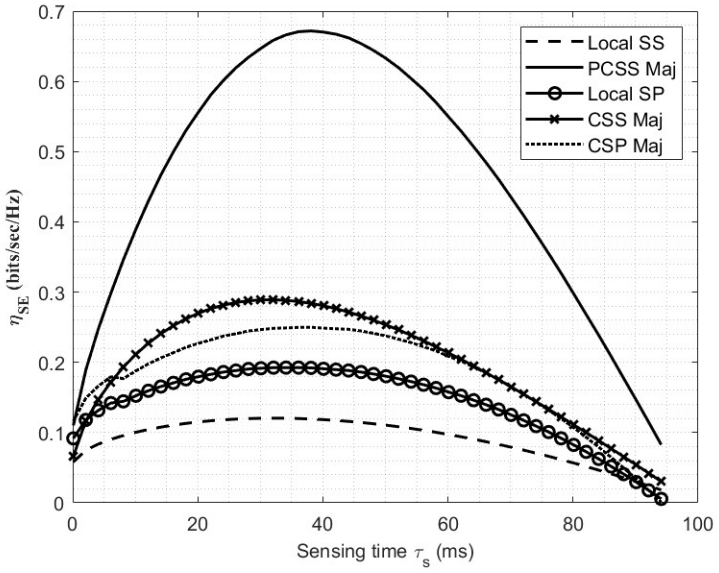


Figure 2.5: SE versus the sensing time for different schemes.

Figure 2.6 shows the EE performance with PCSS and majority rule, local SS, local spectrum prediction, CSS with majority rule, and CSP versus the sensing time τ_s. The prediction error values in Figure 2.4 were used to obtain these results. For $\tau_s < 15$ ms, which is the usual operational range of the sensing time, the EE with PCSS and majority fusion rule is better than that with local spectrum prediction, CSS with majority rule and CSP. However, the EE with local SS is the best among these schemes as increasing the number of cooperative SUs results in higher network energy consumption.

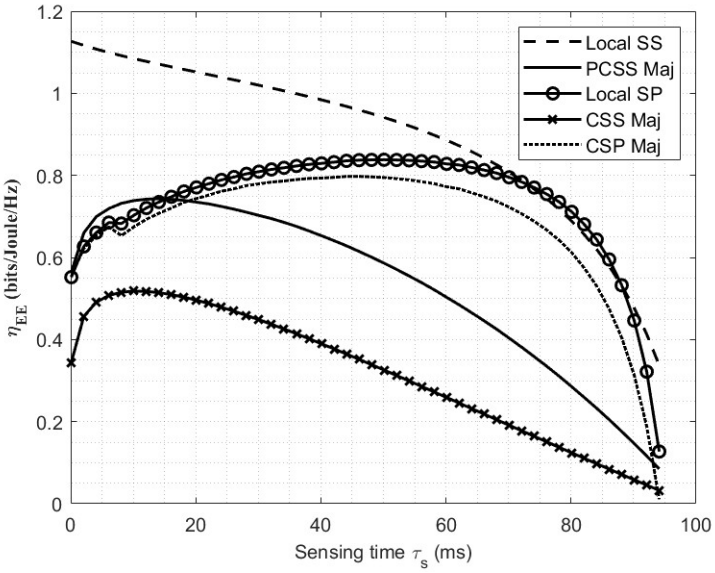


Figure 2.6: EE versus the sensing time for different schemes.

Figure 2.7 presents the SE and EE with PCSS and different fusion rules versus the sensing time τ_s for $P_e^p = 0.2$. This shows that increasing τ_s improves the SE and EE until the optimal τ_s is reached. Then, the SE and EE decrease as τ_s is increased beyond this point. For any fusion rule, an optimal ${\tau_s}^*{}_{SE}$that maximizes the SE exists in the range $0 < {\tau_s}^*{}_{SE} < T - M\tau_r$ and an optimal ${\tau_s}^*{}_{EE}$ that maximizes the EE exists in the range $0 < {\tau_s}^*{}_{EE} < T - M\tau_r$. However, ${\tau_s}^*{}_{SE} \neq {\tau_s}^*{}_{EE}$, so the EE and SE cannot both be maximized considering τ_s and K. Moreover, majority fusion rule at the fusion centre provides almost the same SE and EE performance as the optimal final decision threshold K. Since K is an integer in the range 1 to M, it is feasible to perform an exhaustive search to find the optimal K for a given τ_s. The AND fusion rule provides the worst SE and EE while the OR fusion rule has better performance than the majority fusion rule when $\tau_s < 4$ ms. Thus, both the final decision threshold K and the sensing time τ_s are important parameters that should be optimized in order to maximize the SE and EE.

Figure 2.8 shows the SE and EE as a function of the final decision threshold K with the optimum sensing times. The SE is maximized when $K = 4$ with a value of 0.604 bits/sec/Hz and the EE is maximized when $K = 2$ with a value of 0.807 bits/Joule/Hz. For $K = 1$ (OR fusion rule), the SE is 0.506 bits/sec/Hz and the EE is 0.799 bits/Joule/Hz. For $K = 8$ (AND fusion rule), the SE is 0.340 bits/sec/Hz and the EE is 0.664 bits/Joule/Hz.

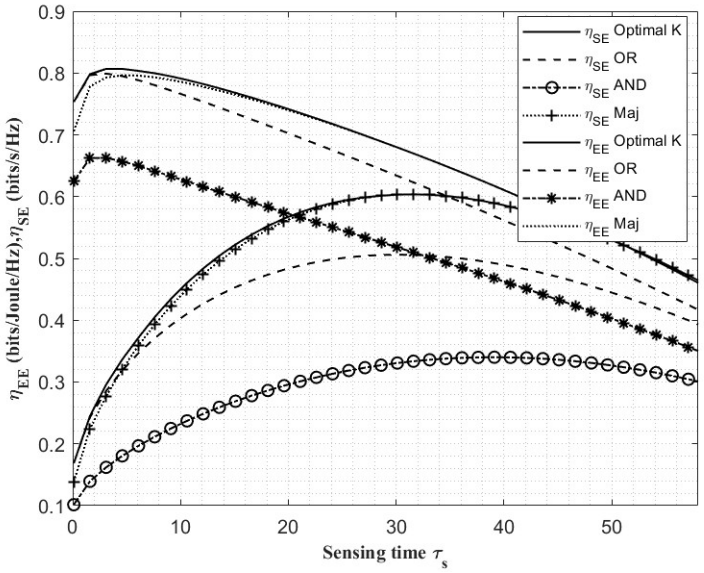


Figure 2.7: SE and EE versus the sensing time τ_s for different fusion rules.

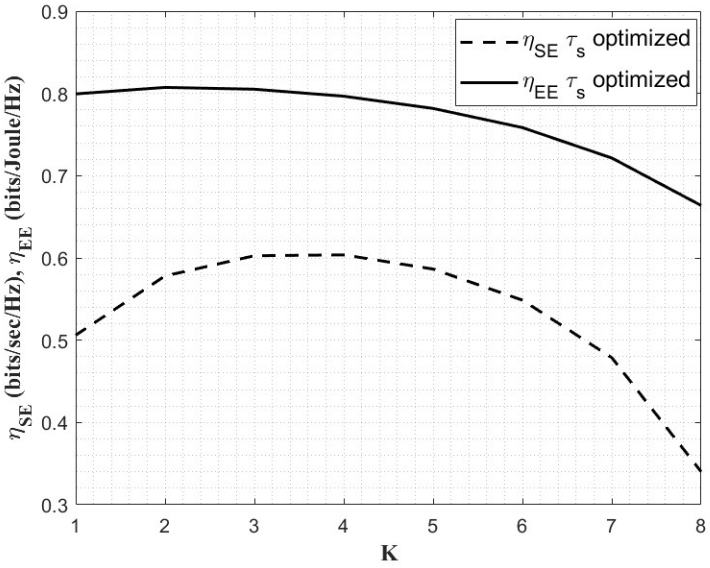


Figure 2.8: SE and EE versus the final decision threshold K with the optimum sensing times τ_s.

Figure 2.9 presents the SE and EE for the PCSS scheme with the majority fusion rule for different M. As the number of SUs who are cooperating in the PCSS scheme increased, the SE increased and the optimal τ_s that maximizes the SE decreased. As the number of SUs who are cooperating in the PCSS scheme increased, the EE decreased and the optimal τ_s that maximizes the EE decreased.

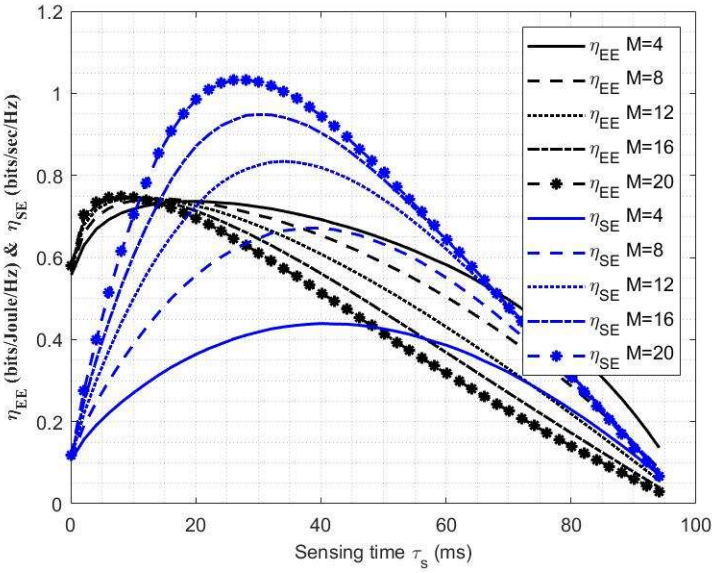


Figure 2.9: SE and EE versus the sensing times τ_s for the PCSS with the majority fusion rule for different M.

Figure 2.10 presents the optimal SE from the general optimization problem and the corresponding SE when ρ=0 versus γ, while Figure 2.11 gives the optimal EE when ρ=0 and the corresponding EE when $\rho = 1$ versus γ. At $\gamma = -16$ dB, the optimal SE when ρ =1 is 0.098 bits/sec/Hz higher than that of the corresponding SE when ρ=0. This gap decreases as $\gamma > -12$ dB until at $\gamma = 0$ dB the solutions are approximately the same. The optimal EE when ρ=0 and the corresponding EE when ρ =1 are shown in Figure 2.11. At $\gamma = -16$ dB, the optimal EE when ρ =0 is 0.269 bits/Joule/Hz higher than that of the corresponding EE when ρ =1. This gap decreases as γ increases until at $\gamma = 0$ dB, the solutions are approximately the same.

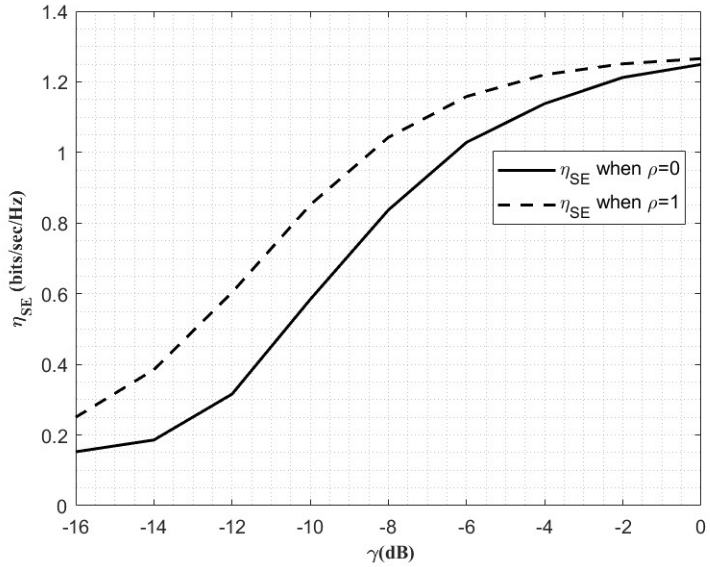


Figure 2.10: The optimized SE when $\rho = 1$ and the corresponding SE when $\rho = 0$ versus the SNR γ.

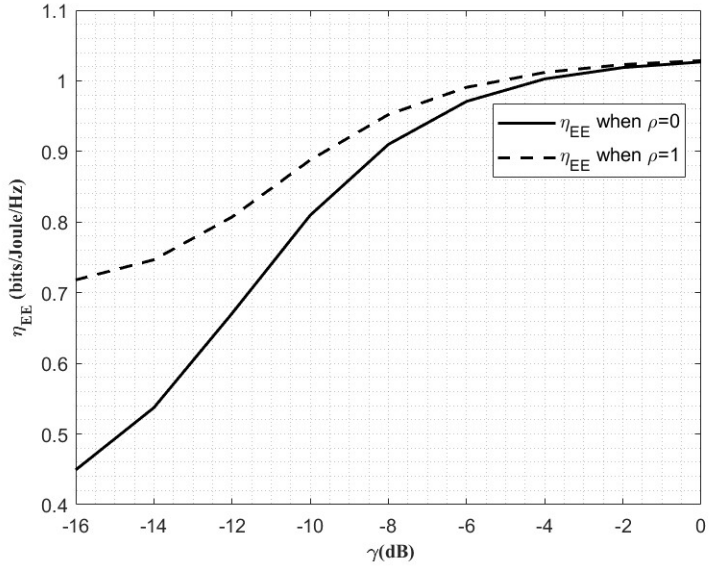


Figure 2.11: The optimized EE when $\rho = 0$ and the corresponding EE when $\rho = 1$ versus the SNR γ.

Figure 2.12 shows $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$ versus ρ when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$ as well as the corresponding values of SE and EE. When $\rho = 0$, the optimization problem becomes maximizing η_{EE} subject to $0 < \tau_s < T - M\tau_r$ and $1 \leq K \leq$

M. When K and τ_s are jointly optimized, the optimal EE is $\eta_{EE} = 0.807$ bits/Joule/Hz. When $\rho = 1$, the optimization problem becomes maximizing η_{SE} subject to $0 < \tau_s < T - M\tau_r$ and $1 \leq K \leq M$. When K and τ_s are jointly optimized, the optimal SE is $\eta_{SE} = 0.604$ bits/sec/Hz. These results show that ρ should be chosen carefully to balance the EE and the SE depending on the network goals. For an energy-constrained CRN, the goal is to maximize the EE while satisfying the SE requirements. If the goal is for the EE to be greater than 0.732 bits/Joule/Hz, then $\rho \leq 0.5$. In addition, the η_{SE} and η_{EE} curves show that the EE and SE cannot both be maximized under the same constraints.

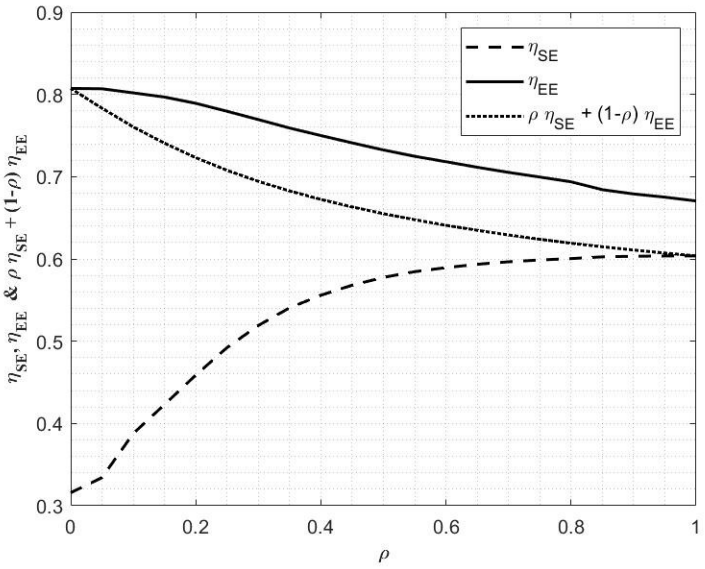


Figure 2.12: $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$, SE and EE versus ρ when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$.

Figure 2.13 shows the EE versus SE when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$ for the PCSS and CSS schemes. The ends of the curves represent the cases $\rho = 0$ (left) and $\rho = 1$ (right). These results indicate that the EE and SE cannot be optimized simultaneously. Moreover, the benefits of combining the prediction with CSS in the PCSS scheme can be seen by comparing the result of PCSS with CSS. There is a significant improvement in the optimized SE and EE with the PCSS scheme compared to the CSS scheme. When K and τ_s are jointly optimized, the optimal EE with PCSS is $\eta_{EE} = 0.807$ bits/Joule/Hz, while the optimal EE with CSS is only $\eta_{EE} = 0.525$ bits/Joule/Hz. When K and τ_s are jointly optimized, the optimal SE with PCSS is $\eta_{SE} = 0.604$ bits/sec/Hz while the optimal SE with CSS is $\eta_{SE} = 0.289$ bits/sec/Hz.

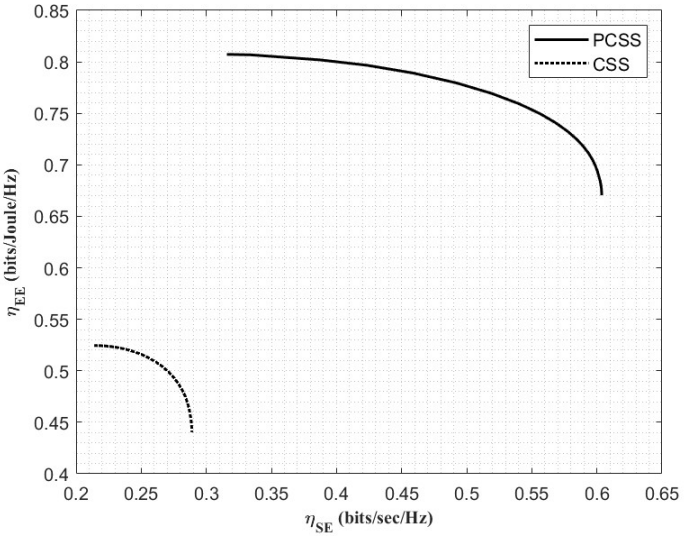


Figure 2.13: EE versus SE when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$.

Figure 2.14 shows the EE versus SE when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$ for the PCSS scheme for different M. As the number of SUs who are cooperating in the PCSS scheme increased, there is a significant improvement in the optimal SE compared to the optimal EE when K and τ_s are jointly optimized. This result is expected from Figure 2.9.

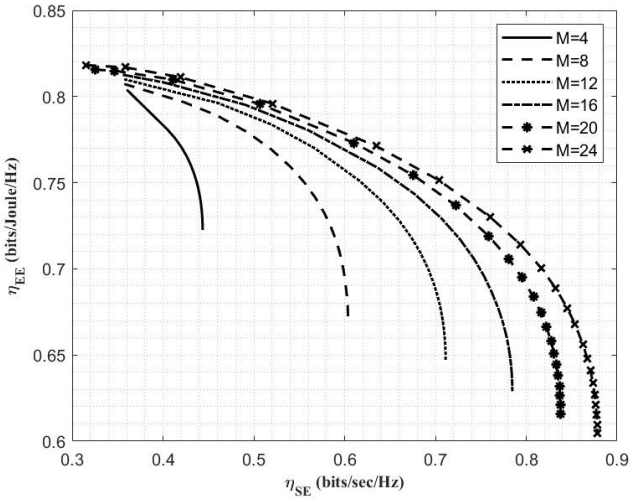


Figure 2.14 EE versus SE with PCSS when K and τ_s are jointly optimized to maximize $\rho\,\eta_{SE} + (1-\rho)\eta_{EE}$ for different M.

Figure 2.15 shows the EE versus the SE threshold $\widetilde{\eta_{SE}}^{th}$. A CRN with high SE requirements is considered so that $\widetilde{\eta_{SE}}\left(\tau_s{}^*{}_{EE}\right) \leq \widetilde{\eta_{SE}}^{th}$. In this case, using the optimal value of K from Algorithm 4 results in $\widetilde{\eta_{EE}} = 0.807$ bits/Joule/Hz for $\widetilde{\eta_{SE}}^{th} < 0.330$ bits/sec/Hz. Then, the EE decreases as $\widetilde{\eta_{SE}}^{th}$ increases. For the majority fusion rule, $\widetilde{\eta_{EE}} = 0.796$ bits/Joule/Hz for $\widetilde{\eta_{SE}}^{th} < 0.350$ bits/sec/Hz. Then, the EE decreases as $\widetilde{\eta_{SE}}^{th}$ increases. The performance with the OR fusion rule is better than that with the majority rule for $\widetilde{\eta_{SE}}^{th} < 0.310$, as was also shown in Figure 3.8. For the OR fusion rule, $\widetilde{\eta_{EE}} = 0.799$ bits/Joule/Hz for $\widetilde{\eta_{SE}}^{th} < 0.280$ bits/sec/Hz. Then, the EE decreases as $\widetilde{\eta_{SE}}^{th}$ increases. Note that $\widetilde{\eta_{SE}}^{th} > 0.5$ cannot be met with the OR fusion rule. Thus, designing an optimal final decision threshold K is important to improve the EE for a CRN. Note that results for the AND fusion rule are not given as this rule provides the worst performance as shown in Figure 2.8.

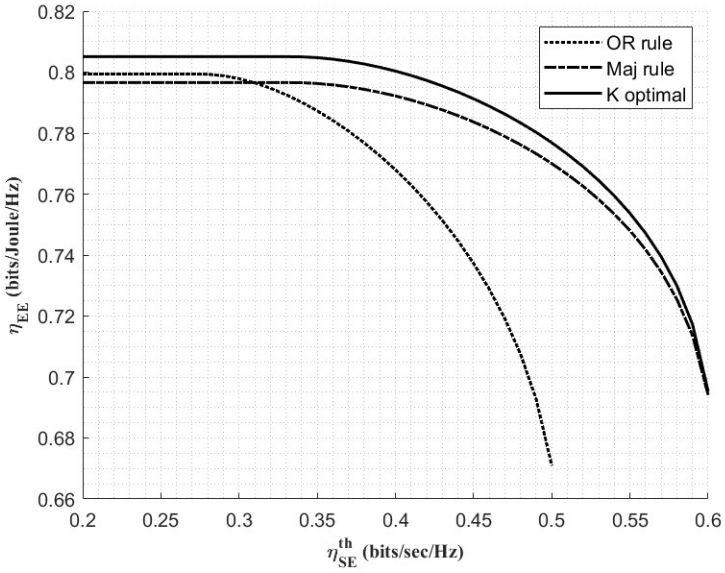


Figure 2.15: The optimal EE versus η_{SE}^{th}.

2.6 Conclusion

This chapter investigated PCSS as an approach to enhance spectrum awareness in CRNs. By integrating spectrum prediction with cooperative sensing, the proposed framework improves the accuracy of channel state estimation and enables more efficient spectrum utilization. Simulation

results demonstrated that CSP significantly outperforms single spectrum prediction and traditional spectrum sensing. In particular, majority fusion achieves the best overall performance by effectively balancing prediction accuracy and robustness. The results also show that improved prediction accuracy leads to enhanced SE through better channel selection, while EE is influenced by the tradeoff between reduced sensing operations and increased cooperation overhead. Several important insights can be drawn from this work. First, prediction improves sensing efficiency but is not sufficient on its own for optimal system performance. Second, cooperation enhances robustness and accuracy, although it introduces additional communication and energy costs. Third, a fundamental tradeoff exists between SE and EE, which must be carefully managed in system design.

Despite these advantages, this work is limited to spectrum sensing and prediction and does not address dynamic resource allocation in decentralized environments. In addition, the framework assumes a centralized fusion mechanism, which may not be scalable in large networks. These limitations motivate the next chapter of this research. In Chapter 3, learning-based approaches are introduced to enable distributed and adaptive resource allocation in C-D2D networks, addressing the challenges of dynamic environments and decentralized decision-making.

.

Chapter 3

Multi-Agent Deep Reinforcement Learning for Resource Allocation in Cognitive D2D Networks

The explosive growth in wireless devices and the IoT, along with machine-type communications, has created tremendous demand for high mobile traffic data rates. In 2022, there were over 7.2 billion mobile phone users globally, and projections indicate that this number will surpass 7.5 billion by 2025, contributing to substantial increases in mobile data generation [44, 45]. To address spectrum scarcity and enhance utilization, innovative solutions like CR and D2D communication have been proposed alongside traditional Cellular Networks (CNs) [46]. D2D communication allows direct interaction between devices using the same radio resources as Cellular Users (CUs), reducing delay and power consumption while improving spectrum efficiency. D2D networks typically operate in two modes: underlay where CUs and D2D pairs share the spectrum and transmit simultaneously under interference constraints and overlay where part of the spectrum is reserved exclusively for D2D pairs, with the remainder used by CUs [47].

CR facilitates dynamic and efficient spectrum use by enabling devices to sense their environment, adjust transmission parameters, and access underutilized spectrum without interfering with licensed users. It enhances SE by promoting cooperative sharing between PUs and SUs to limit interference [11]. CR can be integrated with D2D communication to form C-D2D systems, which are pivotal for developing 6G networks. This integration enhances energy efficiency, spectrum utilization, and coverage while reducing power consumption and signaling overhead. As a result, C-D2D systems have attracted significant research attention [11], enabling devices to sense interference and make informed channel use decisions [48].

In C-D2D networks, devices can connect directly without relying on a centralized BS. They use spectrum sensing and cognitive decision-making to manage resources and interference dynamically in a distributed manner. However, C-D2D networks face several challenges, including interference management, resource allocation, and optimization. Addressing these challenges requires advanced techniques in machine learning, optimization, and signal processing,

as well as careful design and integration of protocols and algorithms tailored to the specific requirements of C-D2D communication. Moreover, in many practical situations, obtaining complete CSI knowledge or accurately estimating statistical information can be challenging in highly dynamic environments. Because of these difficulties, model-free learning approaches are becoming essential and effective in communication scenarios where prior knowledge is limited or uncertain [49].

AI, particularly DRL, has advanced significantly, offering solutions for interference mitigation, robust channel coding, and efficient resource allocation. DRL combines RL with deep learning, enabling agents to learn through trial and error, adapt to dynamic conditions, and handle complex decision-making tasks [11]. Unlike supervised learning, which typically operates in an input-output fashion, RL is goal-oriented and suited for decision-making across a sequence of actions [44]. DRL is particularly useful in high-dimensional and multi-agent scenarios, making it ideal for optimizing resource allocation in distributed D2D networks [50, 51]. DRL also effectively addresses decision-making under uncertainty by learning hidden patterns through interaction and historical data analysis, optimizing decisions based on these learned patterns [47, 52]. This approach reduces the extensive signaling demands of centralized schemes, which require instantaneous global CSI for optimizing resource allocation in D2D communication networks [47, 50]. The RL policy search operates as a Markov Decision Process (MDP) with agents updating policies independently. However, in multi-agent environments, where multiple agents update their policies simultaneously, the environment becomes non-stationary [53]. MADRL is effective in optimizing policies for D2D networks, especially in complex scenarios involving cooperation, competition, or coexistence among multiple agents [52].

Numerous resource allocation schemes based on traditional optimization methods have been explored for C-D2D communication underlaying CN. While optimization-based centralized control methods are relatively easy to implement, they may not be well-suited for dynamic and uncertain spectrum environments [54]. Further, future wireless networks with high user density and rapidly changing conditions introduce new challenges for resource allocation, including significant signaling overhead due to the need for instantaneous CSI and solving optimization problems using traditional techniques with nonlinear constraints. Therefore, this chapter investigates optimization-based distributed resource allocation in C-D2D systems using MADRL underlaying CN with perfect and imperfect CSI.

3.1 Related Work

Conventional D2D resource allocation has been extensively investigated using optimization-based, heuristic-based, and reinforcement learning approaches. Traditional optimization techniques generally focus on spectrum reuse, interference mitigation, and power allocation under predefined communication settings. An evolutionary optimization algorithm for optimal D2D user selection based on channel conditions was introduced in [45]. Although optimization and heuristic-based methods can achieve efficient resource allocation in relatively static wireless environments, their dependence on accurate CSI and centralized iterative optimization limits their applicability in dynamic D2D networks with distributed decision-making and spectrum uncertainty.

Recently, DRL techniques have attracted significant attention for wireless resource management due to their adaptability in dynamic environments without requiring explicit mathematical models [55]. A comprehensive survey of DRL techniques for networking applications was presented in [56]. Several DRL-based approaches have been proposed for D2D communication systems. A DRL-based overlay D2D resource allocation framework employing a priority sampling dueling Double Deep Q Network (DDQN) was proposed in [18] for power control and resource allocation. Joint uplink–downlink subcarrier assignment and power allocation using a DDQN was investigated in [59]. Distributed multi-agent Deep Q-Network (DQN)-based mode selection and channel allocation in heterogeneous cellular networks were studied in [19] to maximize the system sum rate under Quality of Service (QoS) constraints. In [61], a DRL-based framework for maximizing the sum rate and fairness among D2D pairs using non-orthogonal multiple access was proposed with centralized control at the BS. A distributed DRL-based approach for joint channel selection and power control in overlay D2D networks was introduced in [47], while [57] and [62] proposed a DDQN-based D2D spectrum access algorithm allowing dynamic switching between underlay and overlay modes using a centralized agent. Resource management optimization under partially observed CSI was explored in [63] using a deep recurrent reinforcement learning framework.

To improve scalability and distributed decision-making, multi-agent DRL has recently emerged as an effective framework for wireless resource management. By considering the interactions among multiple agents, multi-agent DRL alleviates non-stationarity issues and improves policy stability compared with conventional single-agent RL methods [64]. In [65], a partially observable trust-region-based multi-agent DRL framework was proposed. A fully

decentralized multi-agent DRL algorithm employing soft actor-critic learning for cooperative caching and fetching in D2D networks was developed in [66]. In [53], multi-agent actor-critic and neighbor-agent actor-critic frameworks were introduced for distributed spectrum allocation in D2D downlink scenarios using centralized training and decentralized execution. Similarly, a joint Resource Block (RB) scheduling and power control scheme based on Distributed Deep Deterministic Policy Gradient (D3PG) multi-agent DRL was proposed in [67] to improve network sum rate and fairness. A distributed device-centric RL-based resource allocation approach was introduced in [68]. Moreover, a multi-agent Proximal Policy Optimization (PPO) algorithm for multiple unmanned aerial vehicles was given in [69] to minimize interaction overhead and optimize deployment efficiency. An innovative framework named Federated multi-agent DRL was developed in [70] to enhance privacy protection and minimize overhead. It integrates federated learning with multi-agent DRL underlaying vehicle-to-vehicle communications. In [71], the joint client selection and dynamic resource allocation problem was investigated for federated edge learning with imperfect CSI-based DRL. However, these studies did not explicitly consider multi-agent DRL in cognitive D2D communication with opportunistic spectrum access and spectrum sensing uncertainty.

In cognitive radio networks, Machine Learning (ML) and DRL techniques have been widely investigated for spectrum sensing, dynamic spectrum access, and interference management. Cooperative spectrum prediction using Hidden Markov Models was investigated in [26] to improve spectrum sensing performance in CRNs. DRL-based spectrum access schemes for secondary users were explored in [72] and [73]. In [49], DRL-based dynamic uplink access for cognitive IoT networks with energy harvesting was studied using Deep Deterministic Policy Gradient (DDPG) learning. A hierarchical multi-agent DRL framework combining spectrum sensing and power allocation was proposed in [54] for CRNs. Dynamic spectrum access optimization using DDQN was investigated in [51], while Centralized Training with Decentralized Execution (CTDE)-based multi-agent DRL was explored in [74] to maximize access success rates in interweave CRNs.

Compared with conventional underlay D2D communication, cognitive D2D communication introduces additional challenges due to opportunistic spectrum access, spectrum sensing uncertainty, and primary-user protection requirements. In conventional D2D communication, spectrum reuse is generally predefined and coordinated within the cellular network, whereas C-

D2D users must dynamically identify and opportunistically access available spectrum while avoiding harmful interference to primary users. Consequently, C-D2D resource allocation becomes a coupled spectrum sensing, access, interference management, and power allocation problem under dynamic and partially observable wireless conditions [11]. In this context, in [112], a centralized GA utilizing global CSI was proposed to optimize the quality of experience in CR-based heterogeneous networks with coexisting C-D2D pairs and cellular users through resource allocation and power control. Cognitive and energy harvesting-based D2D communication performance was investigated in [46] using stochastic geometry. Efficient resource allocation in C-D2D communication using Lagrangian optimization and geometric water-filling techniques was studied in [48]. DRL-based resource management has also been investigated for cognitive D2D-enabled IoT networks. In [113], a coordinated multi-agent DRL framework was proposed for joint RB assignment and transmission power control in social-aware cognitive D2D-based IoT networks, where spectrum sensing and QoS-aware optimization were incorporated to improve energy efficiency and communication reliability. However, the proposed framework mainly focused on coordinated actor–critic learning and did not investigate policy-based learning architectures, DTDE/CTDE learning structures, or imperfect CSI-aware resource allocation. Unlike [113], which mainly focused on energy-efficiency and QoS-aware optimization for cognitive IoT networks, this work focuses on sum-rate maximization and spectrum reuse efficiency in C-D2D underlay cellular networks. Furthermore, this work investigates the impact of perfect and imperfect CSI on both CTDE and DTDE learning architectures, while also analyzing how centralized and fully distributed learning structures affect fairness among C-D2D users. Furthermore, the comprehensive survey presented in [5] identified several open challenges in C-D2D communications, including resource allocation and power control, while highlighting that ML-based C-D2D solutions remain very limited in the existing literature. Similarly, the survey in [114] identified AI /ML-aided D2D communication as a promising future research direction.

Motivated by these research gaps, this work investigates a multi-agent PPO-based framework for joint RB allocation, power control, and spectrum-aware resource management in C-D2D underlay cellular networks under both perfect and imperfect CSI conditions. As summarized in Table 3.1, existing MADRL-based D2D studies mainly focus on conventional underlay or overlay D2D networks without explicit cognitive-radio functionality, spectrum sensing, or opportunistic PU/ SU-aware spectrum access. In contrast, existing C-D2D studies

mainly rely on analytical, optimization-based, or game-theoretic approaches, while limited attention has been given to distributed PPO-based MADRL frameworks with DTDE/CTDE learning architectures under imperfect CSI conditions. Therefore, the proposed framework aims to bridge these research directions by jointly considering cognitive spectrum reuse, distributed multi-agent learning, and imperfect CSI-aware resource allocation in C-D2D underlay cellular networks.

Table 3.1. Comparison of existing D2D and C-D2D resource allocation approaches

Ref	System Model	Method	Cognitive D2D	Optimization Objective	CSI	Learning Architecture	Main limitation
[47]	Overlay D2D network	FP optimization + distributed DRL	✗	Joint channel selection and power control for weighted sum rate maximization	Local CSI + outdated nonlocal CS	Distributed execution	No cognitive radio, no spectrum sensing
[53]	Underlay D2D-assisted cellular networks	Multi-Agent/Neighbor-Agent Actor-Critic / (multi-Agent DRL)	✗	Spectrum allocation	Partial observation	CTDE	No cognitive radio, no spectrum sensing
[54]	Dynamic spectrum access CRN	Hierarchical multi-Agent DRL	✗	Joint spectrum sensing and power allocation	Partial observation	CTDE	No C-D2D communication
[57], [62]	Hybrid underlay/overlay D2D-assisted cellular network	Generalized DRL-DDQN	✗	location-based spectrum access strategy	Partial observation	centralized DRL controller	No MARL cooperation; no joint power allocation
[58]	Underlay D2D-assisted cellular network	priority sampling-based dueling DDQN (multi-agent DRL)	✗	Joint power control and resource allocation	Partial observation	DTDE	No cognitive radio/spectrum sensing.
[60]	D2D-enabled heterogeneous cellular network with mmWave	Distributed multi-agent DRL (DQN)	✗	Joint mode selection and channel allocation for sum-rate maximization under QoS constraints	Partial observation	DTDE	No cognitive radio, no power control optimization,
[61]	NOMA assisted underlay D2D-	DDPG + D3PG + Successive Convex Approximation	✗	RB/resource and power allocation for sum-rate/fairness maximization	Global CSI at BS; limited CSI at agents	Hybrid / CTDE	No cognitive spectrum sensing or opportunistic PU/SU access

	cellular network						
[67]	Underlay D2D cellular / IoT network	Multi-agent DQN + D3PG + Convex Optimization	✗	Joint RB scheduling and power control for proportional fairness and sum-rate improvement	Partial / Dynamic CSI	CTDE	No cognitive radio modeling, no spectrum sensing
[46]	Cognitive D2D with energy harvesting-assisted cellular network	Stochastic Geometry	✓	Outage probability analysis and spectrum access evaluation	Statistical channel knowledge	N/A	No DRL/multi-agent DRL-based optimization
[48]	Cognitive D2D network	Lagrangian optimization and geometric water filling	✓	Joint subcarrier and power allocation	Estimated CSI	N/A	Iterative Centralized optimization, No learning capability
[112]	CR-enabled heterogeneous with C-D2D	Centralized GA and semi-distributed Stackelberg Game	✓	Quality of Experience optimization with resource and power allocation	Global CSI/ local CSI in semi-distributed mode	N/A	No DRL/multi-agent DRL framework or distributed learning under CSI uncertainty
[113]	Social-aware cognitive D2D-based IoT network	Coordinated multi-agent DRL-Prioritized Experience Replay (Actor–Critic)	✓	Joint RB assignment and power control for energy-efficiency and QoS optimization	Instantaneous/local observations	Coordinated distributed MADRL	No PPO-based learning, no CTDE/DTDE comparison, and no imperfect CSI analysis
This work	C-D2D underlay cellular network	Multi-agent PPO	✓	Joint RB/channel and power allocation for C-D2D resource allocation	Perfect and imperfect CSI	DTDE and CTDE	—

3.2 Contributions of This Chapter

Inspired by the research gaps identified in the literature and the open challenges highlighted in the comprehensive survey in [11], this paper investigates joint RB assignment and power allocation for C-D2D communications underlaying cellular networks. In the considered system, CUs act as PUs, while C-D2D pairs operate as SUs that opportunistically reuse cellular spectrum resources while satisfying CU QoS constraints. Unlike conventional underlay D2D

communication, C-D2D communication introduces additional challenges related to spectrum sensing, opportunistic spectrum access, dynamic interference management, and imperfect CSI.
The uplink joint RB and power allocation problem is formulated as a MINLP problem involving binary RB assignment variables and non-convex SINR constraints. Since conventional optimization and heuristic-based approaches generally rely on centralized processing and accurate CSI, their applicability becomes limited in dense and dynamic C-D2D environments with distributed decision-making and spectrum uncertainty. To address these limitations, this work develops a multi-agent PPO-based DRL framework for distributed C-D2D resource allocation.
In the proposed framework, each C-D2D pair acts as an autonomous learning agent that jointly selects RBs and transmission power levels based on local observations without requiring global network information. Spectrum-awareness is incorporated through energy detection, where C-D2D users measure the received energy levels over candidate frequency bands and use the sensing outcomes as part of the DRL observation space. This enables opportunistic spectrum reuse while limiting co-channel interference to primary users. The proposed framework supports adaptive and scalable distributed decision-making in dynamic wireless environments and enables efficient spectrum reuse among spatially separated C-D2D pairs.

Motivated by the challenges of centralized and independent learning discussed in [64] and inspired by distributed learning approaches such as [75], this work investigates two multi-agent learning architectures: DTDE and CTDE. The impact of perfect and imperfect CSI on both learning architectures is analyzed, and the trade-off between centralized-assisted and fully distributed learning is investigated in terms of sum-rate performance, convergence behavior, scalability, and fairness among C-D2D users. Unlike existing optimization-based C-D2D approaches and conventional DRL-based D2D frameworks, the proposed framework jointly considers cognitive spectrum reuse, distributed multi-agent learning, and imperfect CSI-aware resource allocation.

The main contributions of this work are summarized as follows:

• A realistic C-D2D underlay cellular network model is developed in which CUs act as primary users and C-D2D pairs operate as secondary users under both perfect and imperfect CSI conditions. The proposed system incorporates spectrum sensing, opportunistic spectrum reuse, and co-channel interference management. The network environment is implemented using the OpenAI Gym framework.

• The uplink joint RB assignment and power allocation problem for C-D2D communication is formulated as a MINLP optimization problem involving binary RB allocation variables and coupled non-convex SINR expressions. A one-to-many RB reuse strategy is adopted to improve spectrum efficiency while maintaining CU QoS requirements.

• A multi-agent PPO-based DRL framework is proposed for distributed C-D2D resource allocation, where each C-D2D pair acts as an independent learning agent. The proposed framework jointly optimizes RB selection and transmission power control using local observations and spectrum sensing outcomes without requiring prior global system knowledge.

• Two learning architectures, DTDE and CTDE, are investigated and compared for C-D2D communication networks. The impact of perfect and imperfect CSI on both architectures is evaluated, while the effect of centralized-assisted and fully distributed learning on fairness and resource allocation performance among C-D2D users is analyzed.

• Unlike conventional optimization-based, heuristic-based, and centralized DRL approaches, the proposed multi-agent PPO framework addresses the joint discrete-continuous optimization problem in a scalable, model-free, and distributed manner. The proposed framework adapts efficiently to dynamic wireless environments and dense spectrum reuse scenarios.

• The proposed framework is extensively evaluated against DRL-based benchmark algorithms, including SAC and DQN, as well as heuristic approaches such as GA and random search. Simulation results demonstrate improved sum-rate performance, robustness under imperfect CSI, improved fairness performance observed among C-D2D users, and better scalability with increasing network density.

3.3 System Model

The C-D2D communication underlaying the CN system includes one BS, M CUs, and N C-D2D pairs (sender and receiver) as shown in Figure 3.1. In this system, C-D2D pairs can share the uplink spectrum resources of CUs. FDMA is used to facilitate multiple access for both CU and C-D2D communication in the uplink during a time slot t. Within this framework, a set of K RBs is available for spectrum allocation. An RB represents the smallest unit of spectrum resources that can be assigned to a user. We assume that each CU has been assigned an RB and that RBs can be allocated to multiple C-D2D pairs. Both CUs and C-D2D pairs can only use a single RB at a time. In C-D2D communication, C-D2D pairs establish connections before spectrum allocation.

Spectrum sensing is typically performed by the C-D2D receiver over the sensing interval τ, using ED on their selected RB. The energy received during this interval is denoted by E_t^n.

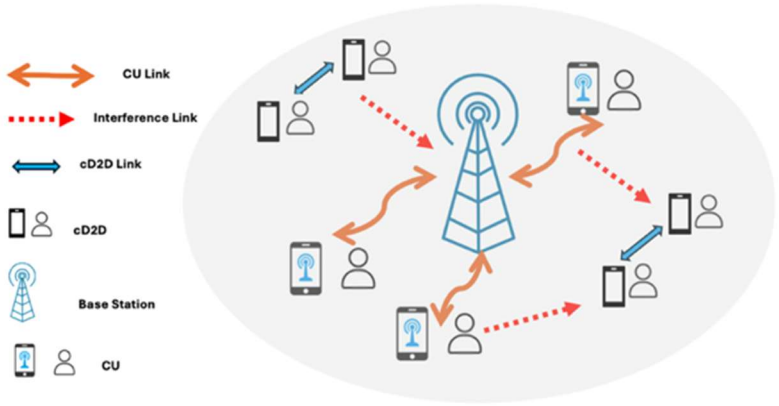


Figure 3.1: System model of the cognitive D2D network underlaying the cellular network.

3.3.1 Channel Model

Two CSI cases are considered: imperfect CSI and perfect CSI. In both cases, large-scale path-loss and small-scale block Rayleigh fading are considered. Let $\overline{g_{x,y}} = PL(d_{x,y})$ represent the large-scale path-loss component between x and y where $d_{x,y}$ is the distance between them. The small-scale Rayleigh fading channel coefficient $h_{x,y}$ is modeled as a complex Gaussian random variable $h_{x,y} \sim \mathcal{CN}(0,1)$. Thus, the channel gain in time slot t is $g_{x,y} = |h_{x,y}|^2 \overline{g_{x,y}}$ [56].

In the case of imperfect CSI, the receiver does not know $h_{x,y}$. Instead, it has an estimated channel gain $\widehat{h_{x,y}}$which can be considered a noisy version of the actual gain. Thus, the imperfect channel estimate can be expressed as $\widehat{h_{x,y}} = h_{x,y} + e_{x,y}$, where $e_{x,y} \sim \mathcal{CN}(0, \sigma_e^2)$ is the channel estimation error which can be modeled as a complex Gaussian random variable with variance σ_e^2. This variance represents the degree of uncertainty in the channel estimation with a larger σ_e^2 indicating worse CSI quality. With imperfect CSI, the channel gain $\widehat{g_{x,y}}$ (based on the estimated gain) is $\widehat{g_{x,y}} = |\widehat{h_{x,y}}|^2 \overline{g_{x,y}}$. Substituting $\widehat{h_{x,y}} = h_{x,y} + e_{x,y}$ gives $\widehat{g_{x,y}} = |h_{x,y} +$

$e_{x,y}\big|^2\ \overline{g_{x,y}}$. When the estimation error $e_{x.y}$ is zero (perfect CSI), the estimated channel gain equals the actual channel gain [76, 77].

3.3.2 SINR Model for CUs and C-D2D Users

In the uplink, C-D2D communication can cause significant interference to cellular links when sharing the same RB. The instantaneous SINR of the received signal at the BS from the *m*th CU over a given RB can be written as

$$\xi_b^m = \frac{P_c\ g_{c,b}^m}{\sum_{n=1}^{N} \alpha^{m,n}\ P_d^n\ g_{s,b}^n + \sigma^2} \tag{3.1}$$

where P_c and P_d^n are the transmit powers of the CU and the *n*th C-D2D sender over the RB, respectively, and $g_{c,b}^m$ and $g_{s,b}^n$ are the channel gains from the *m*th CU to the BS and from the *n*th C-D2D sender to the BS, respectively. The AWGN power at a receiver is denoted by σ^2. Finally, $\alpha^{m,n}$ is the RB allocation indicator, which is 1 if the *n*th C-D2D pair occupies the *k*th RB of the *m*th CU, and 0 otherwise.

The instantaneous SINR of the received signal at the *n*th C-D2D receiver from its sender over a given RB can be expressed as

$$\xi_d^n = \frac{P_d^n\ g_{s,r}^n}{P_c\ g_{c,r}^m + \sum_{i \neq n}^{N} \alpha^{i,n}\ P_d^i\ g_{s,r}^i + \sigma^2} \tag{3.2}$$

where P_d^i is the transmit power of the *i*th C-D2D sender, $i \neq n$, and $g_{s,r}^n$, $g_{c,r}^m$, and $g_{s,r}^i$ are the channel gains from the *n*th C-D2D sender to its receiver, from the *m*th CU to the *n*th C-D2D receiver, and from the *i*th C-D2D sender to the *n*th C-D2D receiver. Finally, $\alpha^{i,n}$ represents the co-channel interference among the C-D2D pairs, which is 1 if the *i*th C-D2D pair is using the same RB as the *n*th C-D2D pair, and 0 otherwise.

The instantaneous data rates of the *m*th CU and *n*th C-D2D pair can be expressed as

$$\mathcal{R}_c^m = W(1 + \xi_b^m) \tag{3.3}$$

and

$$\mathcal{R}_d^n = \frac{T_{total} - \tau}{T_{total}} W(1 + \xi_d^n) \tag{3.4}$$

respectively, where W is the RB bandwidth, T_{total} is the total time available for communication, and τ is the sensing time.

3.4 Problem Formulation and Proposed Solution

This section considers the joint RB-power allocation problem with resource sharing to maximize the total sum rate of the C-D2D pairs while considering minimum data rate requirements for the CUs. The optimization problem can be formulated as

$$\max_{P_n^d,k} \quad \sum_{n=1}^{N} \mathcal{R}_{\mathrm{d}}^{\mathrm{n}} \tag{3.5}$$

$$s.t. \quad C_1\colon \xi_b^m \geq \xi_{th}^m \quad \forall m \in M \tag{3.5.a}$$

$$C_2\colon P_d^n \leq P_d^{max} \ \forall n \in N \tag{3.5.b}$$

$$C_3\colon \alpha^{m,n} \in \{0,1\} \quad \forall m \in M, \forall n \in N \tag{3.5.c}$$

$$C_4\colon \sum_{k=1}^{K} \alpha^{m,n} \leq 1 \quad \forall k \in K \tag{3.5.d}$$

where constraint C1 is the minimum SINR requirements of the M CUs, constraint C2 is the transmit power limits of the N C-D2D pairs, constraint C3 is the RB allocation indicators for the CUs and C-D2D pairs, and constraint C4 ensures that at most one RB is allocated to each C-D2D pair. The reuse of an RB by a C-D2D pair is constrained by a predefined threshold. If the interference exceeds this threshold, no C-D2D pairs are allowed to transmit using that RB.

The optimization problem (3.5) cannot be solved directly because it is a MINLP problem. This arises from the combinatorial nature of the integer variable k and the nonconvexity due to the interference terms in ξ_b^m and $\mathcal{R}_d^n$. Nonconvex optimization problems are particularly challenging to solve. Moreover, in practical scenarios, the C-D2D and CU channels are dynamic, making it impractical to obtain instantaneous channel information, especially in high-mobility and/or large-scale networks. As a result, traditional optimization methods are typically unsuitable for the resource allocation problem considered here.

To address these limitations, the optimization problem in (3.5) is reformulated as a multi-agent MDP framework in which each C-D2D pair acts as an autonomous DRL agent. In this framework, the discrete RB allocation variables and continuous power-control variables are mapped into the action space of the DRL agents, while channel conditions, interference levels, and

spectrum sensing outcomes are incorporated into the state space. The reward function is designed according to the original optimization objective in (3.5), namely maximizing the C-D2D sum rate while satisfying the QoS constraints of the CUs.

Unlike conventional optimization methods, multi-agent DRL enables distributed C-D2D agents to learn adaptive resource allocation policies directly through interaction with the wireless environment without requiring explicit mathematical models or repeated optimization at each time slot. Furthermore, multi-agent DRL supports decentralized decision-making aligned with the distributed nature of C-D2D communication and adapts efficiently to time-varying wireless conditions. Therefore, multi-agent DRL provides a scalable and adaptive framework for dynamic C-D2D resource allocation under imperfect CSI and distributed spectrum access conditions.

The detailed state representation, action definition, reward design, and PPO learning framework are presented in Section 3.5.

3.5 Proposed MADRL Algorithm

In this section, we first introduce a general RL framework. In the proposed system, multiple C-D2D pairs reuse RBs already assigned to CUs, making the use of MADRL techniques essential. Thus, we reformulate the optimization problem in (3.5) into a model-free, MADRL approach using the MDP framework. Finally, we propose a model-free, PPO-based MADRL algorithm to optimize resource allocation for C-D2D communication underlaying CN using the CTDE and DTDE frameworks.

3.5.1 MDP Formulation

RL is a model-free algorithm that develops policies to maximize the expected discounted reward through agent-environment interactions. In a typical DRL setup, the environment is modeled as an MDP, defined by a tuple (S, A, P, R, γ) where S is the set of possible states, A is the set of possible actions, P is the state transition probabilities, R is the reward function, and $\gamma \in (0,1)$ is the discount factor that balances immediate and future rewards. At each time step, an agent observes their current state, selects an action based on its policy, receives a reward, and transitions to a new state. The goal of RL is to discover the optimal policy that maximizes the expected discounted reward by approximating the state-action value function using one of the RL techniques

[51]. In DRL, agents interact with the environment in discrete time steps $t = 0,1, \dots, T$ where T is the final time step.

At time step t, an agent in state s_t takes an action $a_t \in A(s_t)$, receives a reward r_{t+1}, and transitions to state s_{t+1}. The policy $\pi: A \times S \rightarrow [0,1]$, $\pi(s,a) = \Pr(s_t = s | a_t = a)$, defines the probability of taking a particular action in a given state. It is updated based on agent experience and reward [44]. The objective of RL is to maximize cumulative discounted reward functions by finding an optimal policy π^*. The long-term reward R_t is defined as the accumulated and discounted reward by $R_t \triangleq \sum_{v=0}^{T-t-1} \gamma^v r_{t+v+1}$, where $\gamma \in [0,1]$ is the discount factor [44]. The policy π^* specifies the action a_t to take in response to an arbitrary state s_t. In other words, π^* maps the optimal action (such as allocated RB and power) for every state [49].

DRL algorithms are generally classified into two categories: 1) value-based methods, which alternate between evaluating the policy and improving it, and 2) policy-based methods, which directly optimize the policy using the estimated gradient of the expected return [65]. In the proposed model, each C-D2D pair acts as a DRL agent with respect to the cellular environment given that an agent can only observe its local state and cannot access or infer the states or actions of other agents or the CUs at the current time. The MDP of the proposed model is given in the following subsections.

- States: At time t, the local state $s_t^n \in S$ observed by agent n can be defined as $s_t^n = \{E_t^n, g_{s,r}^n, I_{t-1}^n, P_{d,t-1}^n, K_{t-1}^n\}$ where E_t^n is the energy of the received signal over a sensing interval from the spectrum sensing process, $g_{s,r}^n$ is the channel gain from the nth C-D2D sender to its receiver, I_{t-1}^n is the interference at the C-D2D receiver in the previous time slot; $P_{d,t-1}^n$ is the transmit power for the C-D2D sender in the previous time slot, and K_{t-1}^n is the RB selected in the previous time slot.
- Transition Probability: $P(s_{t+1}^n | s_t^n, a_t^n)$ is the probability of moving from the current state $s_t^n \in S$ to a new state $s_{t+1}^n \in S$ after taking an action $a_t^n \in A$.
- Actions: At each time slot, the nth C-D2D sender takes an action a_t, consisting of two decisions: selecting the RB and the transmit power, $a_t^n = [K_t^n, P_{d,t}^n]$.
- Reward: The total reward r_t^{total} is based on maximizing the overall sum rate for the C-D2D links given in (3.5) and can be expressed as

$$r_t^{total} = \sum_{n=1}^{N} r_t^n = \sum_{n=1}^{N} \mathrm{F}_d \, \mathcal{R}_d^n \tag{3.6}$$

where $F_d = 1$ if all constraints are satisfied, and 0 otherwise, to act as a penalty for the C-D2D pairs.

3.5.2 Multi-agent PPO for Resource Allocation

In MADRL algorithms, environment dynamics are nonstationary for each agent due to the influence of other agent actions. This chapter addresses the sum rate optimization problem for RB and power allocation using PPO-based MADRL. Two frameworks are considered for MADRL, namely DTDE and CTDE, as shown in Figure 3.2.

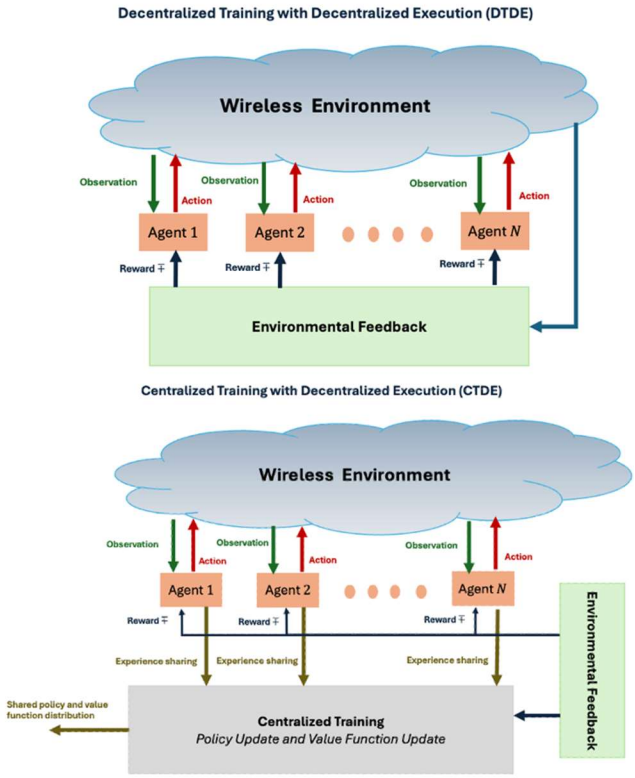


Figure 3.2: The CTDE and DTDE based MADRL frameworks.

In the DTDE framework, both the training and execution phases rely solely on local observations by the agents. They do not have access to centralized or global information at any stage. This makes problem solving more challenging as agents must cooperate or compete with limited information. In the CTDE framework, agents are trained centrally which helps to coordinate strategies. Thus, agents gather data independently, and this data is used to update a

single shared policy in a centralized manner. However, each agent acts independently using the shared policy to interact with its environment.

PPO, an on-policy algorithm introduced by OpenAI [78], has an actor-critic structure. The actor selects actions, while the critic evaluates expected rewards. PPO stabilizes training with a clipping mechanism that penalizes large deviations from the old policy, preventing excessive policy updates. This makes PPO robust for practical reinforcement learning. The objective function uses the probability ratio between the old and new policies and a clipping parameter ϵ to control the ratio range to ensure stable updates within $[1-\epsilon, 1+\epsilon]$ [64]. In the proposed PPO-MADRL model, two frameworks are proposed, and they are described below.

A. CTDE-based PPO-MADRL

In the CTDE-based PPO-MADRL framework, the central server uses the local experience from the agents (which is not shared directly between the agents), to update the shared policy. Then each agent uses the shared policy to act based on its local observation of the environment. This is decentralized because the agents do not rely on the decisions of others during this phase, even though they share the same policy. In this framework, learning begins with parameter initialization. During training, each agent interacts with the environment individually, while contributing to the overall policy and value function updates. At time step t, agent n selects an action a_t^n according to the shared policy π_θ conditioned on its local observation s_t^n. After executing this action, the agent receives a reward r_t^n based on the current environment state and observes the next state s_{t+1}^n.

The agent then stores their experience tuple $(s_t^n, a_t^n, r_t^n, s_{t+1}^n)$ in its local buffer. Every iteration, the experience from the agents is aggregated to update the shared policy and value function. To do so, the advantage function $\hat{A}_t^n$ for each agent is computed as the difference between the expected return from taking an action and the actual return, which is given by

$$\hat{A}_t^n = r_t^n + \gamma V(\phi)(s_{t+1}^n) - V(\phi)(s_t^n) \tag{3.7}$$

The policy update is then performed by maximizing the PPO objective. The objective function uses a clipping mechanism to limit the change in the policy between updates, ensuring that the new policy does not deviate too much from the old one. Specifically, the PPO objective is given by

$$\mathcal{L}_{PPO}(\theta) = \widehat{\mathbb{E}_t}\left[\min\left(c_t(\theta)\hat{A}_t, \text{clip}(c_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{3.8}$$

where $c_t(\theta)$ is the probability ratio

$$c_t(\theta) = \frac{\pi_\theta\left(a_t^n | s_t^n\right)}{\pi_{\theta_{old}}\left(a_t^n | s_t^n\right)} \tag{3.9}$$

The clipping ensures the updates are stable by preventing large changes in the policy between iterations. After computing the PPO objective, the shared policy parameters θ and the value function parameters ϕ are updated using the aggregated experiences of all agents. The shared policy is updated to reflect the collective learning of the multi-agent system. Finally, the current state of each agent is updated to the newly observed state s_{t+1}^n, and this process is repeated until the agents converge to a near-optimal policy that maximizes the total sum rate for the C-D2D network. This algorithm strikes a balance between centralized learning, where agents share their experience to optimize a global policy, and decentralized execution, where each agent independently selects actions based on their local observations during execution. This design allows for efficient coordination and learning during training while enabling agents to operate autonomously during execution. The algorithm steps are summarized in Algorithm 3.1.

Algorithm 3.1: CTDE-based PPO Multi-agent DRL
1**:** Initialization: Set the learning rate α, discount factor γ, clipping parameter ϵ, mini-batch size B, and number of agents N
2**: for** each C-D2D agent $n \in N$ **do**
3**:** Initialize the shared policy π_θ and shared value function $V(\phi)$
4**:** Initialize the initial state s_0^n for all n
5**: end for**
6**: until** convergence **do**
7**: for** each C-D2D agent $n \in N$ **do**
8**:** Select action a_t^n according to the shared policy π_θ, $a_t^n \sim \pi_\theta(a_t^n \mid s_t^n)$

9: Compute reward r_n^t based on the current environment state according to (6)
10: Observe the new state s_{t+1}^n
11: Store experience $(s_t^n, a_t^n, r_t^n, s_{t+1}^n)$ in the agent's local buffer
12: Every iteration, aggregate the experience from all agents and update the shared policy $\pi_\theta(\theta)$ and shared value function $V(\phi)$ using the PPO objective according to (7) and (8) in the central server
13: Update the current state $s_t^n = s_{t+1}^n$
14: **end for**
15: **end until**

B. DTDE-based PPO-MADRL

In this framework, each agent learns independently based on its observations and rewards without any explicit central coordination. Agents observe how their actions influence the shared environment and adjust their policies based on this feedback. The interactions between agents are learned indirectly through experience in the environment. However, learning can be difficult because agents must adapt to a non-stationary environment where the policies of other agents are constantly changing. In DTDE, learning begins with parameter initialization. During training, which continues until the agents converge to a near-optimal policy that maximizes the total sum rate for the C-D2D network, each agent operates independently, interacting with the environment and learning from its own experience. At time step t, agent n selects an action a_t^n according to its local policy π_{θ^n} based on its current observation s_t^n. After executing this action, the agent receives a reward r_t^n based on the current state of the environment and observes the next state s_{t+1}^n. The agent then stores their experience tuple $(s_t^n, a_t^n, r_t^n, s_{t+1}^n)$ in its local buffer, which will be used later to update their policy and value function. Every iteration, each agent updates its local policy $\pi^n(\theta^n)$ and local value function $V^n(\emptyset^n)$ independently, using only its own experience. The

update begins by calculating the advantage function $\hat{A}_t^n$ for each agent, which is used to estimate how much better or worse a particular action was compared to the expected value

$$\hat{A}_t^n = r_t^n + \gamma V^n(\emptyset^n)(s_{t+1}^n) - V^n(\emptyset^n)(s_t^n) \tag{3.10}$$

Algorithm 3.2: DTDE-based PPO Multi-agent DRL
1: Initialisation: Set the learning rate α, discount factor γ, clipping parameter ϵ, mini-batch size B, and number of agents N
2: **for** each C-D2D agent $n \in N$ **do**
3: Initialize the local policy π_{θ^n} and local value function $V^n(\emptyset^n)$
4: Initialize the initial state s_0^n for each n
5: **end for**
6: **until** convergence **do**
7: **for** each C-D2D agent $n \in N$ **do**
8: Select action a_t^n according to the local policy π_{θ^n}, $a_t^n \sim \pi_{\theta^n}(a_t^n \mid s_t^n)$
9: Compute the reward r_t^n based on the current environment state according to (3.6)
10: Observe the new state s_{t+1}^n
11: Store experience $(s_t^n, a_t^n, r_t^n, s_{t+1}^n)$in the agent's local buffer.
12: Every iteration, update the local policy $\pi_{\theta^n}$and local value function $V^n(\emptyset^n)$ using the PPO objective according to (10) and (11)
13: Update the current state $s_t^n = s_{t+1}^n$
14: **end for**
15: **end until**

Each agent then maximizes its own PPO objective to update its local policy. The PPO objective for agent n is designed to stabilize learning by preventing large updates in the policy using the objective function

$$\mathcal{L}_{PPO}^n(\theta^n) = \widehat{\mathbb{E}_t}\left[\min\left(c_t(\theta^n)\,\hat{A}_t^n, \text{clip}(c_t(\theta^n), 1-\epsilon, 1+\epsilon)\,\hat{A}_t^n\right)\right] \tag{3.11}$$

where $c_t(\theta^n)$ is the probability ratio for agent n, which compares the new policy to the old policy

$$c_t(\theta^n) = \frac{\pi_{\theta^n}(a_t^n|s_t^n)}{\pi_{\theta^n_{old}}(a_t^n|s_t^n)} \tag{3.12}$$

The clipping ensures that the new policy does not change significantly from the old one, thereby maintaining training stability. After the PPO objective is maximized, each agent updates its policy parameters θ^n and value function parameters $\emptyset^n$ based solely on its local experience. Finally, after each time step, the state of the agent is updated to reflect the next observed state s_{t+1}^n. This process continues independently for each agent until training converges, at which point the agents will have learned their respective policies. The decentralized nature of this framework allows each agent to act and learn autonomously, making it efficient in environments where coordination is not essential or practical. However, the absence of experience sharing between agents may lead to slow learning in tasks requiring cooperation. The steps of the algorithm are summarized in Algorithm 3.2.

3.6 Performance Evaluation

In this section, the performance of the proposed algorithm is evaluated, and the impact of different parameters on the system performance is studied.

3.6.1 Computational Complexity Analysis

To evaluate the computational efficiency of the proposed multi-agent PPO framework, its complexity is analyzed and compared with conventional heuristic approaches, including GA and Random Search. GA is adopted as a benchmark because it is widely used in wireless resource allocation literature for solving MINLP-based joint RB and power allocation problems and is commonly employed as a baseline in DRL-based wireless optimization studies. In the proposed PPO framework, each C-D2D pair is modeled as an independent agent that interacts with the wireless environment and updates its policy through neural network training. The computational complexity of the proposed framework consists of two stages: offline training and online execution. During training, the computational complexity per iteration can be approximated as $O(N \cdot B \cdot T \cdot C)$, where N is the number of agents (C-D2D pairs), B is the batch size, T is the number of time steps per episode, and C is the computational cost of a forward/backward neural

network update. Although the training phase introduces additional computational overhead, it is performed offline before deployment. During execution, each trained agent only performs a forward neural network evaluation to determine the RB allocation and transmission power actions. Therefore, the online inference complexity is approximately $O(N \cdot C)$, enabling fast distributed decision-making suitable for real-time C-D2D resource allocation in dynamic wireless environments. For the GA implementation, each chromosome jointly encodes the RB allocation indices and transmission power levels of all C-D2D pairs. The fitness function is designed according to the optimization objective in (3.5), namely maximizing the C-D2D sum rate while satisfying the CU SINR constraints. The computational complexity of GA can be approximated as $O(G \cdot PS \cdot E)$, where G is the number of generations, PS is the population size, and E is the cost of evaluating each candidate solution Unlike DRL-based methods, GA does not require offline training; however, it must repeatedly execute iterative optimization procedures for each resource allocation instance, resulting in high online computational overhead in dynamic wireless environments. RS performs an exhaustive search over the solution space with a complexity of $O(M \cdot P)^N$, where M is the number of CUs and P is the number of discrete power levels. This exponential growth makes RS computationally impractical for dense C-D2D deployments and large-scale wireless networks.

Overall, although the proposed PPO framework introduces additional offline training complexity, it significantly reduces the online computational burden during execution compared with iterative heuristic optimization approaches. Therefore, the proposed framework provides a scalable and computationally efficient solution for dynamic C-D2D resource allocation under distributed wireless environments.

3.6.2 Simulation Setup

A single cell scenario with a radius of 500 m is considered. The number of CUs *M* is set to 20, and the number of C-D2D pairs N varies between 5 and 60. The distance between these pairs is limited to 50 m. The total bandwidth of *K* RBs is 180 kHz. A sensing interval $\tau = 5$ ms is used within a transmission time of 100 ms. Both the large-scale and small-scale fading are assumed to be static within one transmission time. The large-scale path loss model $PL(d_{x,y}) = 37.6 \log_{10}(d_{x,y}) + 128$ is used. The estimation error variance for imperfect CSI is chosen to be 0.1 [75]. The CU transmission power is set to 23 dBm and the maximum C-D2D transmission

power is set to 20 dBm. The noise power spectrum density is -174 dBm/Hz and the CU SINR threshold is 0 dB.

Multi-agent PPO was implemented using Python 3.10, OpenAI Gym, and Ray RLlib. OpenAI Gym is an open-source toolkit for RL developed by OpenAI, while Ray RLlib is a scalable and flexible library for RL. To train a DRL agent with RLlib, an environment for agent interaction was created using OpenAI Gym. Training was conducted using a virtual Intel Xeon CPU with 13GB of RAM on Google Colab. The hyperparameters of the PPO algorithm are learning rate 0.0001, discount factor 0.99, PPO clip parameter 0.3, batch size 512 time-steps, and maximum number of training iterations 200. The actor and critic neural networks were modeled as fully connected feedforward neural networks with two hidden layers, each containing 256 neurons. The tanh activation function was used in the hidden layers. The policy and value networks were optimized using the Adam optimizer. However, the remaining PPO parameters are the default for PPO in Ray RLlib. Different PPO hyperparameter configurations, including learning rate, clipping parameter, batch size, and training iterations, were evaluated during preliminary experiments to achieve stable convergence and reliable learning performance in the considered C-D2D environment. The reported hyperparameters correspond to the configuration that provided the most stable training behavior and consistent performance. To evaluate the robustness of the PPO learning behavior, the convergence-stability analysis was repeated over five independent random seeds. The empirical stabilization iteration is defined as the first training iteration at which the mean reward remains within $\pm 2\%$ of the final-window average for 20 consecutive iterations. The GA uses a population size of 50, 200 generations, and tournament selection size 6. The crossover and mutation probabilities are set to 0.6 and 0.01, respectively. The random search algorithm explores possible power and resource allocation configurations for 20000 iterations, where the solution space size for power and channel allocation is $(M \times P)^N$ and P is the number of discrete power levels derived from P_d^{max}.

The simulation study is designed to assess both the training performance and the execution performance of the proposed multi-agent DRL framework. During the training phase, we analyze the learning behavior of the agents by tracking the average reward versus the number of training iterations. Since the reward function is defined in terms of the achieved sum rate of the agents. This metric serves as an indirect measure of the policy quality during learning. These results demonstrate the convergence and stability of different training strategies. In the execution phase,

the learned policies are evaluated under various network configurations, including different numbers of C-D2D pairs and CU SINR thresholds. The sum rate and fairness index are measured to quantify the performance of each method after training is completed. This distinction ensures a clear separation between learning dynamics and final policy performance in practical applications.

Moreover, the performance of the proposed framework is evaluated and compared with several benchmark approaches in terms of C-D2D sum-rate performance and fairness behavior. The first benchmark is Random Resource Allocation (RRA), where C-D2D pairs randomly select RBs and transmission powers without applying any optimization strategy. The second benchmark is a conventional C-D2D spectrum sensing scheme, where C-D2D users adjust their transmission power according to the sensed PU activity using predefined transmission power levels. The third benchmark is RS, which randomly samples resource allocation actions and selects the configuration achieving the highest immediate reward. The fourth approach involves GA optimization, which is a popular nature-inspired optimization technique that mimics biological evolutionary processes, such as natural selection, crossover, and mutation. GA is selected because it is widely used in wireless resource allocation literature for solving MINLP-based joint RB and power allocation problems and is commonly adopted as a heuristic baseline in DRL-based wireless optimization studies. In the implemented GA framework, each chromosome jointly encodes the RB allocation indices and transmission power levels of all C-D2D pairs. The fitness function is defined according to the optimization objective in (3.5), namely maximizing the C-D2D sum rate while satisfying the CU SINR constraints. Tournament selection, crossover, and mutation operations are iteratively applied to evolve the candidate solutions. Both centralized-GA and Distributed-GA implementations are considered for comparison. The centralized version (Centralized-GA) uses global information to optimize resource allocation, while the distributed version (Distributed-GA) relies on local decisions by each C-D2D user. The proposed framework is further compared with DRL-based benchmark approaches, including SAC-based DRL [26] and DQN-based DRL [58].

The selected baseline algorithms represent different categories of resource allocation approaches commonly adopted in wireless communication and DRL literature. RRA is included as a simple non-optimized benchmark to evaluate the gain achieved by learning-based resource allocation. GA is selected as a representative heuristic optimization method because it is widely used for solving non-convex MINLP-based wireless resource allocation problems. DQN is

adopted as a representative value-based DRL approach, whereas SAC is selected as an advanced actor-critic DRL algorithm with continuous action optimization capability. These baselines enable a comprehensive comparison between conventional heuristics, value-based DRL methods, and policy-based DRL approaches for wireless network resource allocation.

The simulation study is designed to evaluate both the training performance and execution performance of the proposed multi-agent DRL framework. During the training phase, the learning behavior of the agents is analyzed by tracking the average reward with respect to the number of training iterations. Since the reward function is defined based on the achieved C-D2D sum rate, this metric provides an indirect measure of policy quality during learning and demonstrates the convergence and stability of different training architectures.

During the execution phase, the learned policies are evaluated under different network configurations, including varying numbers of C-D2D pairs and different CU SINR thresholds. The total sum rate and Jain's fairness index are measured to evaluate the effectiveness of the trained policies after convergence. This distinction separates learning performance from final deployment performance in practical wireless environments.

3.6.3 Simulation Results

In this section, the sum rate of the system is evaluated considering the number of D2D pairs, number of CUs, and minimum SINR requirements. Figure 3.3 presents the D2D sum rate for the RRA algorithm with respect to the number of D2D pairs. This shows that as the number of CUs in a cell increases, more RBs are available for spectrum sharing. As a result, D2D pairs get an opportunity to transmit with less interference to CUs. Therefore, the sum rate of the D2D network increases. In addition, as the number of D2D pairs in a network increases, the sum rate of the D2D pairs increases until the number of available RBs is less than the number of D2D pairs. Then, the D2D sum rate decreases as the number of D2D pairs increases. This is because the co-channel interference among the D2D pairs and CUs increases, and to satisfy the minimum CU SINR condition, D2D pairs are penalized. Thus, an efficient resource allocation algorithm is important to improve the performance of the D2D network underlaying CN.

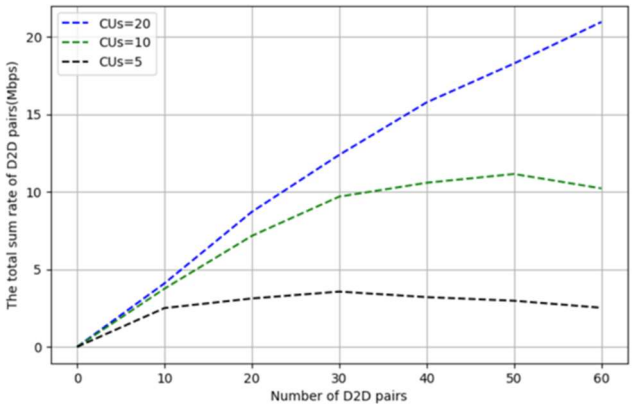


Figure 3.3: The total D2D sum rate versus the number of D2D pairs.

Figure 3.4 gives the average reward over training iterations for three DRL algorithms: PPO, SAC, and DQN, simulated with 20 C-D2D pairs using the DTDE framework. PPO outperforms the others, achieving the highest average reward (~15) with better stability. DQN also performs well, though with more variance, converging around an average reward of 14. SAC, despite being a powerful algorithm, struggles with stability in this setup and has a lower average reward, fluctuating between 8 and 11. PPO will be selected for the remaining network performance analysis due to its superior empirical performance and stability, which align with the objectives of optimizing sum rate in DRL strategies. Although the algorithm was tested for a maximum of 200 iterations, the convergence of PPO is observed around 120 iterations for 20 C-D2D. This limit of 200 iterations was selected based on empirical tests where the average reward performance of the proposed system generally stabilized within this range.

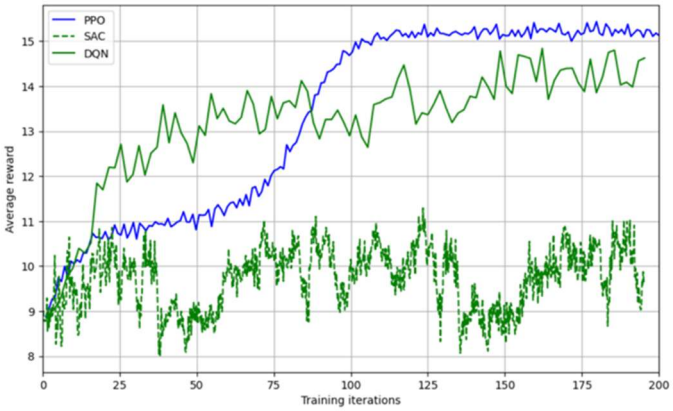


Figure 3.4: The average reward over the training iterations for different DRL algorithms.

Figure 3.5 presents the variation in sum rate with respect to the minimum SINR requirement of the CUs. These results indicate that as the SINR requirement of CUs decreases, more D2D pairs can transmit without degrading CU network performance. However, as the SINR requirement increases, the D2D network sum rate decreases more rapidly due to fewer transmission opportunities for D2D pairs so as to limit co-channel interference to CUs. Additionally, for 20 CUs, the DRL scheme outperforms the RRA scheme. As the number of D2D pairs exceeds the available RBs, the DRL scheme further improves D2D network performance by allowing D2D pairs to acquire the best RBs to minimize interference.

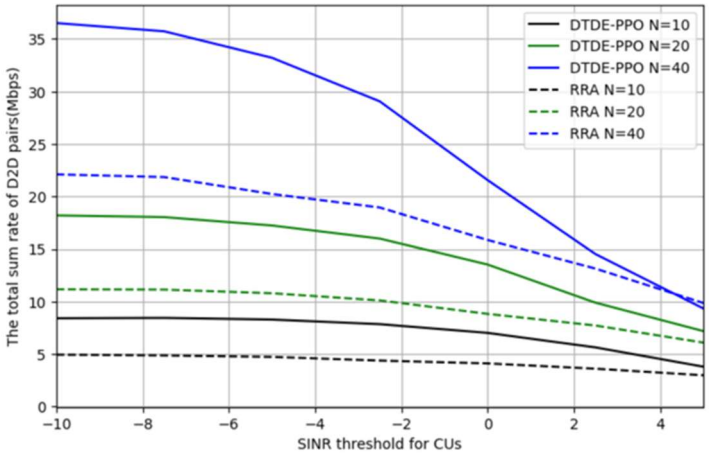


Figure 3.5: Performance of the RRA and DTDE-PPO algorithms versus the CU SINR threshold.

Figure 3.6 presents the total sum rate of different transmission algorithms with respect to the number of D2D pairs. The MADRL PPO-based algorithms achieve a higher sum rate compared to C-D2D and RRA due to their ability to employ decision policies with rewards to optimize performance. In DTDE-PPO C-D2D, the received signal threshold from spectrum sensing in the observation state improves the learning ability of the agent to autonomously choose the best RB and power (lower interference) for transmission that might be assigned to physically distant CUs. These results also show that when the number of D2D pairs in a cell is higher than the number of available RBs, indicating the proposed scheme is effective. However, with 40 D2D pairs, the proposed DTDE-PPO C-D2D scheme achieves a 13%, 20%, and 26% higher sum rate than the distributed DTDE-PPO D2D scheme, C-D2D scheme, and RRA scheme, respectively. However, the DTDE PPO framework performance declines as the number of D2D pairs exceeds 50. This will be discussed later.

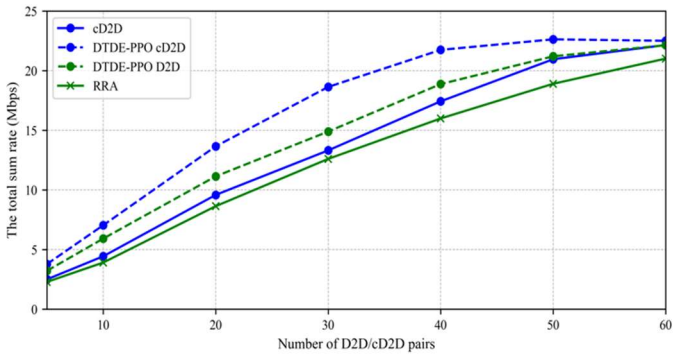


Figure 3.6: The sum rate with four algorithms versus the number of D2D pairs.

Figure 3.7 gives the average reward over the training iterations for both DTDE and CTDE based PPO with perfect and imperfect CSI, estimation error variance ≈ 0.1, and 40 C-D2D pairs. In DTDE, the learning process is more challenging because agents must adapt to a non-stationary environment, where the policies of other agents are constantly changing. This results in slower convergence compared to CTDE, where all agents share the same policy and can pool their experiences, leading to faster convergence. However, CTDE faces limited exploration diversity since all agents share the same policy, causing them to converge to similar behavior patterns, which can be suboptimal in environments requiring diverse strategies. Consequently, DTDE achieves a higher reward than CTDE after convergence.

Furthermore, the results demonstrate the impact of perfect and imperfect CSI on the performance of MADRL. With imperfect CSI, the performance of both frameworks is affected, showing lower reward performance, though both exhibit similar outcomes. This confirms that coordination of all agents in CTDE leads to better performance compared to DTDE. Nevertheless, with perfect CSI, DTDE converged to a higher reward than CTDE.

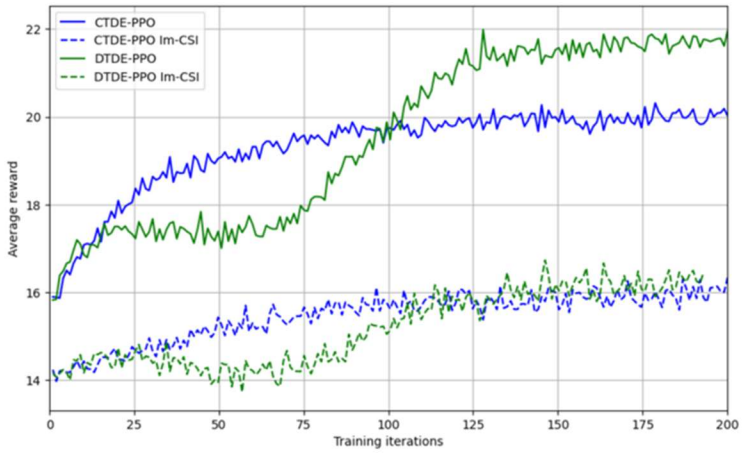


Figure 3.7: The average reward over the training iterations for CTDE and DTDE frameworks.

To further assess the convergence stability and reproducibility of the proposed PPO-based C-D2D resource allocation framework, Figure 3.8 gives the learning curves for $N = 40$ C-D2D pairs scenario with five independent random seeds. The solid curves represent the average reward across random seeds, while the shaded regions indicate ± 1 standard deviation. CTDE-PPO achieves a final average C-D2D rate of 19.97 ± 0.06 Mbps and empirically stabilizes at iteration 125. DTDE-PPO achieves 21.78 ± 0.17 Mbps and stabilizes at iteration 134. These results show that CTDE-PPO provides faster and more stable convergence, whereas DTDE-PPO attains a higher final average reward in this dense setting but requires more training iterations to stabilize. This behavior is expected because centralized training improves coordination during learning, while independent DTDE policies may provide greater policy diversity after sufficient training.

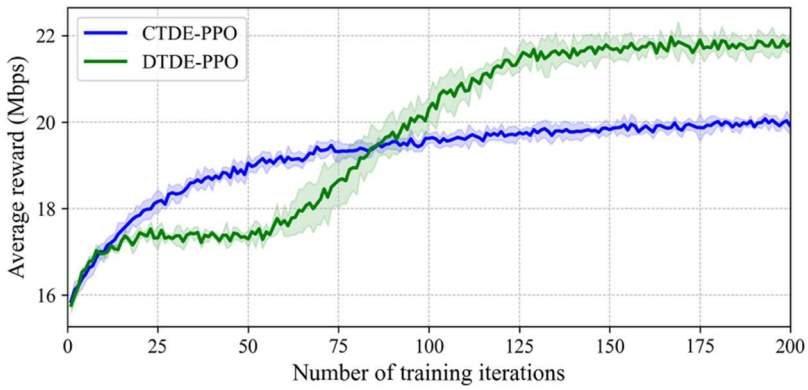


Figure 3.8: The average reward across random seeds versus the number of training iterations.

Figure 3.9 presents the total sum rate of the trained CTDE-based multi-agent PPO and DTDE-based multi-agent PPO frameworks with respect to the number of C-D2D pairs, compared to GA and random search algorithms. When $N < 40$, DTDE-based multi-agent PPO C-D2D outperforms CTDE-based multi-agent PPO C-D2D, the GA-based approaches, and random search. This demonstrates its efficiency in optimizing resource allocation in smaller networks. Distributed-GA is worse because it relies on local information, limiting its ability to achieve globally optimal solutions. For $N > 40$, random search can achieve competitive performance in smaller search spaces due to extensive exploration; however, its computational complexity increases exponentially with network size, making it impractical for real-time deployment. Random search, CTDE-based multi-agent PPO C-D2D, and Centralized-GA, with their reliance on global network information, outperform the distributed methods. The fully distributed DTDE-based multi-agent PPO C-D2D has a better sum rate and interference management than Distributed-GA because it continuously learns and adapts to dynamic conditions. This highlights the superiority of the proposed algorithm and the effectiveness of DRL in distributed scenarios.

GA requires re-execution for every spectrum access, making it inefficient in dynamic environments. In contrast, DRL is trained once, which enables rapid, real-time decisions. Thus, it is superior in rapidly changing network conditions. In terms of scalability, Figure 3.9 shows that DTDE plateaus as the number of C-D2D pairs increases. This is because DTDE agents learn independently and so fail to leverage shared experiences, resulting in slow convergence and inefficiency in large-scale networks. In contrast, CTDE utilizes shared experience which enhances coordination, making it more effective for large networks. Although CTDE improves convergence behavior and learning stability, its centralized training process may introduce additional communication overhead in dense wireless environments, motivating future research on communication-efficient and fully decentralized MARL architectures.

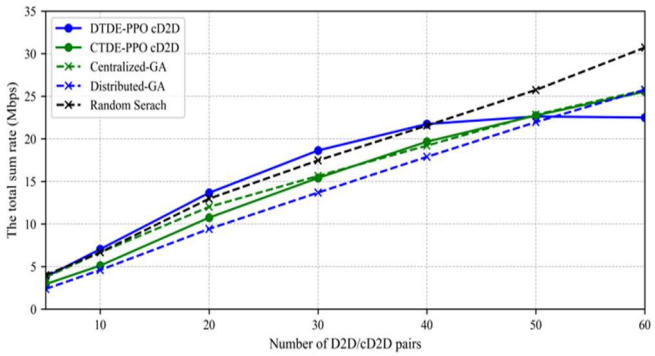


Figure 3.9: The sum rate with four algorithms versus the number of D2D pairs.

To evaluate the efficiency of resource allocation, Jain's Fairness Index (J) is considered, which is defined as $J(\mathcal{R}_d^1, \mathcal{R}_d^2, \ldots, \mathcal{R}_d^n) = \frac{(\sum_{n=1}^{N} \mathcal{R}_d^n)^2}{N.\sum_{n=1}^{N} (\mathcal{R}_d^n)^2}$. This index measures fairness across users based on their average data rates to quantify fairness in resource allocation for D2D pairs [46, 63]. The fairness index ranges from 0 to 1, where larger values indicate a more balanced resource distribution among users.

It should be noted that fairness is not explicitly incorporated into the reward function or optimization objective of the proposed MARL framework. Instead, fairness is evaluated as an additional performance metric to analyze the impact of different learning architectures and resource allocation policies on user-level resource distribution.

The fairness performance is analyzed by varying the number of C-D2D pairs, as illustrated in Figure 3.10. The results show that the fairness index decreases as the number of C-D2D pairs increases due to the higher co-channel interference caused by more aggressive spectrum reuse among users. Nevertheless, the proposed multi-agent PPO-based C-D2D framework achieves better fairness performance compared with the RRA scheme, indicating that DRL-based resource allocation can provide more balanced resource distribution among C-D2D users even without explicitly optimizing fairness in the reward design.

Furthermore, the results demonstrate that the CTDE framework achieves higher fairness performance than DTDE because the centralized training process allows agents to learn more coordinated resource allocation policies through shared training experiences. In contrast, DTDE agents rely entirely on independent local learning, which may lead to less balanced resource allocation decisions in dense network scenarios. However, the fairness index still decreases in

highly dense deployments, reaching approximately 0.4 when the number of C-D2D pairs approaches 60. Future work may investigate fairness-aware reward shaping strategies by incorporating fairness-related terms directly into the reward function. Such an approach could guide the agents toward policies that jointly optimize network throughput and equitable resource allocation among C-D2D users.

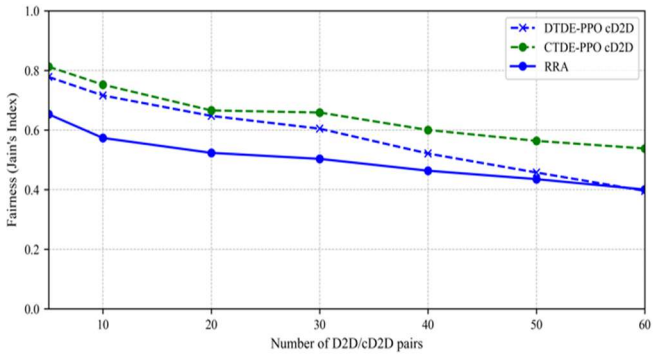


Figure 3.10: Fairness versus the number of D2D pairs.

3.7 Conclusion

This chapter presented a MADRL framework for joint power control and channel allocation in C-D2D networks. By formulating the problem as an MDP and employing a PPO-based approach, the proposed framework enables distributed and adaptive resource allocation under dynamic and uncertain network conditions. Both centralized and decentralized training strategies were investigated, allowing agents to learn efficient policies with limited channel state information. Simulation results demonstrated that the proposed MADRL framework significantly outperforms conventional optimization methods and existing DRL approaches in terms of sum-rate performance, scalability, and robustness. The results also highlight the effectiveness of multi-agent learning in enabling decentralized decision-making without requiring explicit system models, making it well-suited for complex and rapidly changing wireless environments. Several important insights can be drawn from this work. First, learning-based approaches eliminate the need for accurate mathematical models, enabling efficient operation under uncertainty. Second, multi-agent frameworks provide scalable solutions for distributed networks. However, these

advantages come with challenges, including high computational complexity during training, sensitivity to reward design, and coordination difficulties under partial observability.

Despite these improvements, the proposed framework does not fully exploit advanced spectrum sharing techniques, particularly NOMA, which can further enhance spectrum reuse and network capacity. This limitation motivates the work presented in the next chapter, where NOMA is integrated with C-D2D communication within a multi-agent learning framework to further improve system performance.

Chapter 4

NOMA-Enabled Cognitive D2D Networks with Multi-Agent Deep Reinforcement Learning

The rapid growth of emerging wireless services and the proliferation of IoT devices—projected to exceed 75 billion connections by 2025—continue to intensify the demand for high-capacity and low-latency communication. D2D communication has been recognized as a key enabler for enhancing spectrum reuse and reducing network load by allowing proximate users to communicate directly without routing traffic through the BS [79]. Due to short-range operation, D2D transmission improves EE, supports local data exchange, and enhances service quality in areas with weak cellular coverage. However, conventional D2D systems rely on OMA, which limits spectral efficiency and leads to underutilization of radio resources [80], motivating the need for more advanced resource-sharing strategies.

NOMA has emerged as a promising solution for supporting massive connectivity by allowing multiple users to share the same time–frequency resource through power-domain multiplexing. NOMA enhances spectral efficiency but introduces challenges related to SIC, which requires accurate CSI and robust decoding in the presence of user mobility, fading, and interference [81]. When NOMA is combined with D2D communications, it can significantly improve throughput and reduce congestion; however, these benefits also increase co-channel interference and impose stricter SIC and power-control requirements [82].

To further increase spectrum utilization, researchers have explored integrating D2D communication with CR, enabling SUs to opportunistically reuse licensed spectrum while protecting the QoS of primary CUs [83]. In underlay CR operation, D2D transmitters must regulate their power to satisfy interference temperature constraints at the BS. When NOMA is introduced into this setting, the joint design of RB reuse, power allocation, and SIC feasibility becomes a highly complex MINLP problem. Traditional optimization-based resource allocation methods struggle to cope with the high dimensionality and uncertainty introduced by NOMA-enabled D2D communication under CR constraints. These challenges are further intensified in large-scale IoT

deployments, where distributed devices must adapt to time-varying channels, interference conditions, and energy availability. In such environments, model-free learning approaches provide a viable alternative by enabling adaptive decision-making without requiring accurate channel models or repeated real-time optimization. In particular, DRL has recently been introduced as a powerful tool for wireless resource management because it enables agents to learn adaptive policies without relying on explicit channel models. DRL techniques can handle high-dimensional observations, stochastic environments, and multi-objective trade-offs, making them suitable for dynamic multi-user networks [84], [85]. In decentralized D2D settings, MADRL allows each device to autonomously learn optimal spectrum and power strategies while implicitly coordinating with others through environmental feedback.

In this context, this chapter proposes a novel approach that integrates NOMA technology with C-D2D communication to improve radio resource utilization and enhance the overall performance of cellular networks. The joint management of RB reuse, NOMA power allocation, CR interference constraints, SIC feasibility, and fairness among D2D groups calls for a scalable, model-free, and distributed decision-making framework. These challenges motivate the investigation of advanced multi-agent learning frameworks capable of jointly addressing spectrum reuse, power control, and fairness in next-generation wireless networks.

4.1 Related Work and Motivation

Recent studies such as [61], [86]-[88] examine D2D communication under NOMA-based cellular networks, but a common limitation is that NOMA is applied only to CUs, while D2D links remain OMA-based to simplify interference management and reduce computational complexity. As a result, these works improve spectral efficiency at the cellular level but do not fully exploit the potential gains of applying NOMA within D2D pairs themselves.

In [80] and [89]-[97], NOMA was directly applied within D2D groups to enhance spectral efficiency, support multi-receiver D2D links, and improve rate performance through optimized power and subchannel allocation. These studies typically rely on mathematical optimization tools—such as matching theory, dual decomposition, convex approximation, or heuristic search—to manage interference and satisfy QoS or SIC constraints. While they demonstrate notable gains over OMA-based D2D, most of these solutions rely on centralized optimization, assume perfect CSI, or require repeated problem solving at each time slot. Furthermore, they do not incorporate

learning-based distributed decision making nor consider multi-agent interactions among decentralized D2D transmitters. They are largely limited to static or single-objective formulations.

Recent studies such as [83], [98], and [99] investigate CR techniques combined with D2D communication to improve spectrum utilization, throughput, and EE. These works typically optimize power allocation, channel assignment, user pairing, or computation offloading under CR constraints, often incorporating RF EH or NOMA at the cellular tier. However, a common limitation is that D2D links themselves do not employ NOMA; instead, they rely on orthogonal access or serve only as relays/offloading partners to protect PUs and simplify interference management. As a result, existing CR–D2D frameworks do not fully exploit the spectral efficiency gains achievable with NOMA-enabled C-D2D transmissions. Moreover, [99] assumes perfect instantaneous global CSI. This enables centralized convex optimization but may impose significant signaling overhead and practical implementation challenges in large-scale networks.

DRL has recently been explored to address the non-convexity and dynamic adaptation challenges of D2D underlay networks. [100] demonstrated the benefits of AI-driven architectures for 6G-oriented D2D communication in reducing latency and improving throughput. In Single-Agent DRL (SA-DRL), a single centralized agent—typically located at the base station—observes the global network state and optimizes resource allocation decisions for all users. While this approach benefits from global observability with full CSI and stable convergence, it suffers from scalability limitations and high signaling overhead as network size grows. In this category, [101] proposed a centralized SA-DRL framework combining fractional programming with DRL, where a Dinkelbach-twin delayed DDPG algorithm jointly optimized D2D-NOMA clustering and power allocation to maximize long-term EE. Simulation results demonstrated substantial EE gains and improved convergence behavior compared with DQN and DDPG baselines. To overcome scalability issues, MADRL frameworks have been investigated, where multiple agents learn control policies while interacting within a shared wireless environment. In centralized-training MADRL, agents are trained using global/full CSI but execute policies locally. In [102], a centralized MADRL framework was developed for NOMA-enabled D2D underlay networks. The proposed DRL-based resource allocation scheme achieved significant improvements in sum rate, EE, fairness, and outage performance compared with the greedy and RRA schemes.

Similarly, the MADRL framework combining DDPG and PPO was proposed in [103] to maximize the D2D sum throughput under SINR constraints for both cellular and D2D users. Although

performance gains were reported over DQN-based baselines, the evaluation was limited to DRL comparisons only, and no optimization-based benchmarks were considered. The work in [104] further extended centralized-training MADRL to Reconfigurable Intelligent Surfaces (RIS)-assisted NOMA-D2D networks, proposing a multi-agent multi-pass DQN algorithm for joint sub-channel assignment, power control, and RIS phase optimization, yielding a 27% improvement over OMA.

To reduce signaling overhead and enhance scalability, decentralized and federated MADRL architectures have also been explored. In such frameworks, model training is performed locally at user devices, while global coordination is achieved through parameter aggregation rather than raw data exchange. In [85], a federated multi-agent Double DQN (DDQN) framework was proposed for D2D-assisted networks, enabling distributed learning while preserving user privacy. The federated approach achieved a 41.52% improvement in system EE and significantly reduced cellular outage probability compared with multi-agent actor-critic, Dirichlet DDPG, and conventional DQN schemes. However, despite decentralized training, the framework assumes full instantaneous CSI, as the agents' state representation includes global SINR information of all users, implying globally observable network conditions.

Table 4.1 summarizes existing works based on learning paradigm, training architecture, CSI assumptions, and optimization objectives. Although prior studies have applied MADRL to NOMA-enabled D2D systems, most focus primarily on centralized algorithmic choices (e.g., DQN, DDPG, PPO) under full or near-global CSI assumptions. As a result, the role of information structure in enabling decentralized coordination remains insufficiently understood. In cognitive underlay NOMA-D2D systems, where interference coupling, CR protection, and SIC decoding constraints must be simultaneously satisfied, limited observability introduces additional challenges in maintaining both performance and feasibility. In such settings, the amount and structure of available CSI directly influence coordination capability, learning stability, and constraint satisfaction. This motivates a systematic investigation of information-aware decentralized MADRL design. In this work, we explicitly analyze how different CSI levels (local, partial, and full) affect convergence behavior, feasibility enforcement, and achievable performance under both centralized learning and fully decentralized learning architectures. This perspective shifts the focus from algorithm-centric design to understanding the fundamental role of information availability in decentralized wireless resource allocation.

Table 4.1: Comparison with related work

Ref.	DRL	SA/MA	Algorithm	CSI	D2D	NOMA included	CR included	Objective	Other Metrics
[85]	✓	MA-Semi-Distributed (Federated)	DDQN	Full CSI	✓	×	×	Maximize EE	Sum rate, QoS, Outage Probability
[99]	×	×	Karush–Kuhn–Tucker Conditions	Full CSI	✓	✓	✓	Maximize Sum rate	EH
[101]	✓	SA-Centralized Training	Dinkelbach-Twin Delayed DDPG	Partial/ Delayed CSI	✓	✓	×	Maximize EE	×
[102]	✓	MA-Centralized Training	DQN	Full CSI	✓	✓	×	Maximize Sum Rate	Outage Probability, EE, Fairness
[103]	✓	MA-Centralized Training	DDPG/PPO	Full CSI	✓	✓	×	Maximize Sum Rate	×
[104]	✓	MA-Centralized Training	DQN	Full CSI	✓	✓		Maximize Sum Rate	×
This work	✓	MA-Centralized and Distributed Training	PPO	Local/Partial/Full CSI	✓	✓	✓	Maximize Sum Rate	EE, Fairness, SIC Feasibility, Cellular Uutage Probability

SA: Single Agent; MA: Multi Agent.

4.2 Contributions of This Chapter

This chapter investigates joint RB reuse and power allocation for cognitive NOMA-enabled D2D groups under CR protection and SIC decoding constraints. Unlike prior works that focus on algorithmic design under full CSI, this work systematically studies how information structure, constraint enforcement, and fairness regularization affect decentralized coordination in interference-coupled networks. This work shifts the focus from algorithm-centric design to understanding how information structure and constraint-aware reward design enable decentralized coordination in interference-coupled cognitive NOMA-D2D networks. The main contributions of this chapter are

- A unified cognitive underlay NOMA-D2D framework is considered, where spectrum sharing with CUs and intra-group NOMA interference are jointly modeled. This results in a tightly coupled interference structure involving both cross-tier (C-D2D–cellular) and intra-group (NOMA) interactions, providing a challenging setting to study coordinated resource allocation under simultaneous CR protection and SIC decoding constraints.
- A structured observation-state design is developed to enable scalable multi-agent learning under local and partial CSI. The impact of CSI availability on convergence, coordination, and constraint satisfaction is systematically quantified.
- CR protection and SIC decoding constraints are consistently integrated with adaptive-rate modeling through feasibility-gated throughput and penalty mechanisms, ensuring alignment between physical-layer feasibility and learning objectives.
- Dual Optimization Objectives (P1 and P2) are proposed:
    - P1 (Constrained Sum-Rate Maximization): Maximization of long-term C-D2D sum rate subject to CR interference and SIC decoding constraints.
    - P2 (Fairness-regularized Sum-Rate Maximization): An extension of P1 that incorporates the long-term Jain's fairness index via reward shaping, enabling fairness-aware resource allocation while preserving sum-rate efficiency.
- A structured evaluation of two complementary multi-agent learning paradigms is conducted: CTDE and DTDE architectures are compared to reveal the tradeoffs between coordination capability, scalability, and signaling overhead under different CSI conditions.

- The impact of fairness regularization is analyzed, revealing a non-monotonic tradeoff between throughput, fairness, and energy efficiency. Moderate fairness shaping is shown to improve coordination stability and interference management.
- A comprehensive performance evaluation is conducted. The proposed multi-agent PPO scheme is benchmarked against MADRL baselines (DQN, Importance Weighted Actor-Learner Architecture (IMPALA)) and optimization-based heuristics (Genetic Algorithm (GA) and RS). Simulation results demonstrate significant improvements in sum rate, fairness, energy efficiency, cellular outage probability, and SIC decoding success.

4.3 System Model

4.3.1 Network Architecture and Spectrum Sharing Model

We consider a single-cell uplink cellular network underlaid with C-D2D communication employing PD-NOMA**,** as illustrated in Figure 4.1. The network consists of one BS, a set of M CUs, and a set of G C-D2D-NOMA groups denoted by C-DGs. Each CU is allocated at most one uplink RB using OMA, while multiple D2D-NOMA groups are allowed to reuse the same CU uplink RB in an underlay manner. The cellular uplink transmissions are regarded as the primary system, whereas the D2D-NOMA groups operate as SUs following the CR underlay paradigm. Accordingly, D2D transmissions must satisfy predefined interference protection requirements to ensure the QoS of the primary cellular links. Each C-DG consists of a C-D2D transmitter (C-DT) and two C-D2D receivers (C-DRs), strong user d_s with better channel gain and weak user d_w with poorer channel gain. The BS is located at the cell center, while the CUs and C-DGs are assumed to be uniformly and independently distributed within the cell coverage area.

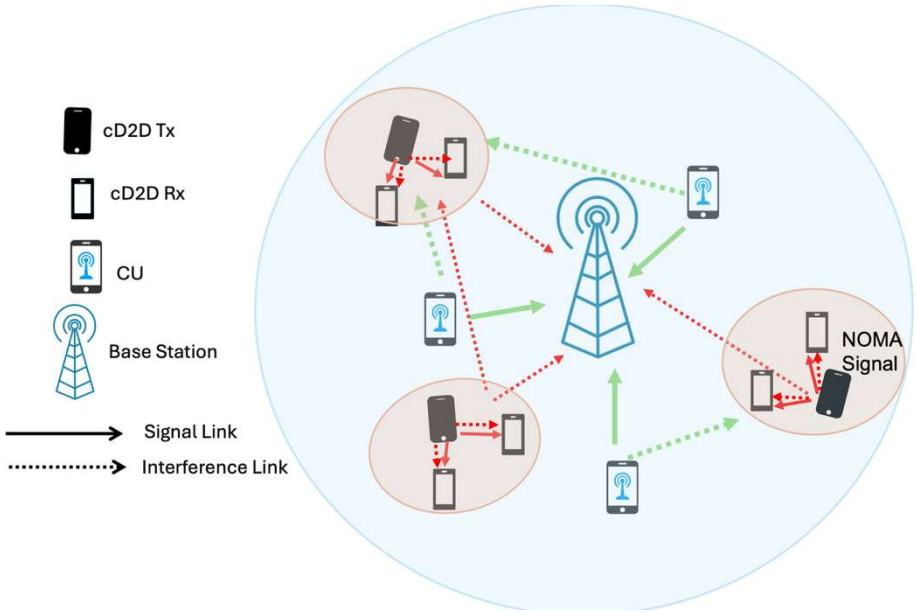


Figure 4.1: The system model.

4.3.2 Resource Reuse and Association Model

The total available bandwidth B is divided into K orthogonal subchannels, denoted by $K \in \{1,2,\dots,k,\dots,K\}$, each of bandwidth $B_k = \frac{B}{K}$. Let $a_{m,k} \in \{0,1\}$ be the binary assignment variable, where $a_{m,k} = 1$ indicates that CU m occupies RB k. Where $\sum_{k=1}^{K} a_{m,k} \leq 1, \forall m$. Each CU m transmits at a fixed maximum uplink power $P_m = P_m^{max}$.

Each C-DG opportunistically reuses the subchannels of the CUs under an underlay spectrum-sharing paradigm. Let $z_{g,k} \in \{0,1\}$ denote the binary reuse variable, where $z_{g,k} = 1$ indicates that C-DG g reuses RB k. with the constraint $\sum_{k=1}^{K} z_{g,k} \leq 1, \quad \forall g$, which ensures each group occupies no more than one RB.

The C-DT transmits a superimposed signal containing messages for both users using the PD-NOMA principle. Let p_g denote the transmit power of C-DT g, and let $\alpha_g \in (0,1)$ denote the NOMA power allocation factor assigned to the weak receiver, such that the strong receiver is allocated power $p_{g,s} = (1-\alpha_g)p_g$ and the weak receiver is allocated power $\mathrm{p_{g,w}} = \alpha_g \, \mathrm{p_g}$. Without loss of generality, the receivers are ordered such that the strong receiver has a larger channel gain than the weak receiver. Where $x_{g,s}$ and $x_{g,w}$ denote the data symbols intended for the

strong and weak users, respectively. Then the transmitted NOMA signal is $x_g = \sqrt{p_{g,s}}\, x_{g,s} + \sqrt{p_{g,w}}\, x_{g,w}$, where the total transmit power satisfies $p_{g,s} + p_{g,w} \leq P_g^{max}$.

During the uplink phase, the CU transmits its signal to the BS on the assigned RB, while the C-DT simultaneously transmits superposed NOMA signals to its two receivers on the same RB. As a result, the BS experiences aggregate interference from all D2D-NOMA groups reusing that RB, and each C-DR experiences interference from the CU transmission and from other D2D-NOMA groups reusing the same RB.

4.3.3 Channel Model and Interference Analysis

All wireless links experience large-scale path loss and small-scale Rayleigh fading. The channel coefficient between nodes x and y on RB k is $\left|h_{x,y}^k\right|^2 = \left|\widehat{h_{x,y}^k}\right|^2 d_{x,y}^{-\beta}$, where $d_{x,y}$ is the distance, β is the path-loss exponent and $\widehat{h_{x,y}^k} \sim \mathcal{CN}(0,1)$ is Rayleigh fading. We assume a block-fading channel model, where channel coefficients remain constant within one transmission interval and vary independently across different intervals. Perfect instantaneous CSI is assumed at the receivers for local SINR calculation and decoding.

When CU m transmits over RB k, the BS receives: the intended CU signal, interference from all C-D2D transmitters reusing the same RB, and the additive white Gaussian noise. The received signal at the BS is given by $y_{m,b}^k = a_{m,k} h_{m,b}^k \sqrt{P_m}\, x_m + \sum_{g=1}^{G} z_{g,k}\; h_{g,b}^k \sqrt{p_{g,s} + p_{g,w}}\, x_g + n_b$. Where $h_{m,b}^k$ represents the CU–BS channel, $h_{g,b}^k$ represents the C-DT–BS channel, and $n_b \sim \mathcal{CN}(0, \sigma_b^2)$ is the noise at BS. Thus, the uplink SINR of CU m is given by

$$\gamma_{m,b}^k = \frac{a_{m,k}\, P_m \left|h_{m,b}^k\right|^2}{\sum_{g=1}^{G} z_{g,k}(p_{g,s}+p_{g,w})\left|h_{g,b}^k\right|^2 + \sigma_b^2.} \tag{4.1}$$

The SINR $\gamma_{m,b}^k$is defined only for the RB assigned to CU m and each CU must satisfy $\gamma_{m,b}^k \geq \gamma_{\mathrm{CU,th}}$.

The NOMA principle requires that the user with the poorer channel gain is allocated more power, and the user with the better channel gain is allocated less power for successful SIC. Therefore, for C-DG g, the power allocation must satisfy the condition $p_{g,w} \geq p_{g,s}, \quad \text{if } \left|h_{g,w}^k\right|^2 < \left|h_{g,s}^k\right|^2$. Each C-DR experiences its intended NOMA signal from the C-DT, interference from other

cDTs reusing the same RB, interference from the CU assigned to the RB, and the receiver noise. Thus, the received signal at the weak user reception is given by $y_{g,w}^{k} = h_{g,w}^{k}(\sqrt{p_{g,w}}\, x_{g,w} + \sqrt{p_{g,s}}\, x_{g,s}) + \sum_{ǵ \neq g} z_{ǵ,k} h_{ǵ,w}^{k} \sqrt{p_{ǵ,w} + p_{ǵ,s}}\, x_{ǵ} + h_{m,w}^{k}(\sqrt{P_m}\, x_m) + n_w$. The weak-user SINR is given by

$$\gamma_{g,w}^{k} = \frac{p_{g,w}\left|h_{g,w}^{k}\right|^2}{p_{g,s}\left|h_{g,w}^{k}\right|^2 + I_{g,w}^{\mathrm{D2D}} + I_{g,w}^{\mathrm{CU}} + \sigma^2} \tag{4.2}$$

where the interference from the CU assigned to the RB is given by $I_{g,w}^{\mathrm{CU}} = a_{m,k}\, P_m \left|h_{m,w}^{k}\right|^2$, the interference from other C-DTs reusing the same RB is given by $I_{g,w}^{\mathrm{D2D}} = \sum_{ǵ \neq g} z_{ǵ,k}(p_{ǵ,w} + p_{ǵ,s})\left|h_{ǵ,w}^{k}\right|^2$. The strong user ($d_s$) first decodes the weak user's signal. The corresponding SINR is

$$\gamma_{g,s\to w}^{k} = \frac{p_{g,w}\left|h_{g,s}^{k}\right|^2}{p_{g,s}\left|h_{g,s}^{k}\right|^2 + I_{g,s}^{\mathrm{D2D}} + I_{g,s}^{\mathrm{CU}} + \sigma^2} \tag{4.3}$$

To ensure reliable decoding and cancellation of the weak-user signal at the strong receiver, the following SIC decoding constraint is imposed $\gamma_{g,s\to w}^{k} \geq \gamma_{SIC,\, min}$. This constraint defines the feasibility of a transmission configuration with respect to SIC decoding at the strong receiver. After canceling the weak user's interference, the strong user SINR becomes

$$\gamma_{g,s}^{k} = \frac{p_{g,s}\left|h_{g,s}^{k}\right|^2}{I_{g,s}^{\mathrm{D2D}} + I_{g,s}^{\mathrm{CU}} + \sigma^2} \tag{4.4}$$

where the interference from CUs is expressed as $I_{g,s}^{\mathrm{CU}} = a_{m,k} P_m \left|h_{m,s}^{k}\right|^2$ and $I_{g,s}^{\mathrm{D2D}} = \sum_{ǵ \neq g} z_{ǵ,k}(p_{ǵ,s} + p_{ǵ,w})\left|h_{ǵ,s}^{k}\right|^2$. Here, $\gamma_{SIC,\, min}$ denotes the minimum SINR margin required at the strong receiver for reliable decoding of the weak-user signal prior to interference cancellation.

4.3.4 Interference and Cognitive Radio Constraints

The aggregate interference at the BS on RB k caused by C-DGs reusing that RB is given by the sum of received interference powers from all active C-DTs. To protect the primary cellular uplink, the SINR of each CU must remain above a minimum threshold to ensure reliable uplink communication $\gamma_{m,b}^{k} \geq \gamma_{CU,th}$. These CR protection constraints regulate the admissible RB reuse patterns and D2D transmit powers.

4.3.5 Achievable Rates

The achievable data rates for the cellular and C-D2D links follow Shannon's capacity, the uplink rate of CU m occupying RB k is expressed as

$$R_{m,b}^{k} = B_k \log_2\left(1 + \gamma_{m,b}^{k}\right) \tag{4.5}$$

For the C-D2D links, when C-DG g reuses RB k ($z_{g,k} = 1$), the achievable rates of the weak and strong C-DRs are given by

$$R_{g,w}^{k} = B_k \log_2\left(1 + \gamma_{g,w}^{k}\right), R_{g,s}^{k} = B_k \log_2\left(1 + \gamma_{g,s}^{k}\right) \tag{4.6}$$

Since each C-DG reuses at most one RB, the total NOMA throughput of group g is

$$R_g = \sum_{k=1}^{K} z_{g,k} \left(R_{g,s}^{k} + R_{g,w}^{k}\right) \tag{4.7}$$

Finally, the total C-D2D network throughput is

$$R_{\text{sum}} = \sum_{g=1}^{G} R_g \tag{4.8}$$

The achievable-rate expressions in (4.5) – (4.8) represent the physical-layer link rates under an adaptive-rate abstraction. The SIC decoding and cellular protection constraints define the feasibility of transmission configurations, and only feasible configurations are considered in performance evaluation.

4.4 Problem Formulation

In this section, we formulate the joint RB assignment, power allocation, and NOMA feasibility problems for the C-D2D-assisted uplink network. The following constraints apply to all three optimization problems

$$\text{(C1)}\ \sum_{m=1}^{M} a_{m,k} \leq 1, \quad \forall k \tag{4.9}$$

$$\text{(C2)}\ \sum_{k=1}^{K} z_{g,k} \leq 1, \quad \forall g \tag{4.10}$$

$$\text{(C3)}\ \gamma_{m,b}^{k}(t) \geq \Gamma_{\text{CU}}, \quad \forall m, k: a_{m,k} = 1 \tag{4.11}$$

$$\text{(C4)}\ \gamma_{g,s\to w}^{k}(t)) \geq \gamma_{\text{SIC, min}}, \quad \forall g, k: z_{g,k} = 1 \tag{4.12}$$

$$\text{(C5)}\ \gamma_{g,w}^{k}(t) \geq \Gamma_w, \qquad \gamma_{g,s}^{k}(t) \geq \Gamma_s, \quad \forall g, k: z_{g,k} = 1 \tag{4.13}$$

$$\text{(C6)}\ \ p_{g,s}(t) + p_{g,w}(t) \le P_g^{\max}, \qquad p_{g,s}(t) \ge 0,\ p_{g,w}(t) \ge 0 \tag{4.14}$$

where (C1) enforces the OMA constraint for CUs, (C2) ensures that each C-DG reuses at most one RB, (C3) guarantees CU SINR protection, (C4) enforces the minimum SINR required for decoding the weak-user signal at the strong receiver, (C5) ensures the minimum SINR requirements for both C-DRs and (C6) constrains the C-DT transmit power and enforces non-negativity.

4.4.1 Problem P1 – Baseline Cognitive NOMA–D2D Sum-Rate Maximization

The goal in P1 is to maximize the instantaneous C-D2D network throughput subject to all feasibility constraints. The optimization variables include the RB reuse indicators $z_{g,k}$, and the NOMA transmit powers ($p_{g,s},\ p_{\mathrm{g},w}$). This leads to the following problem:

$$(P1): \quad \max_{\{z_{g,k},\, p_{g,w},\ p_{\mathrm{g},s}\}} R_{\mathrm{sum}} \tag{4.15}$$

$$s.t. \quad (C1) \text{ to } (C6) \tag{4.16}$$

Problem P1 is a mixed discrete–continuous, highly non-convex program due to the binary RB reuse variables $z_{g,k}$, the coupled interference terms in the SINR expressions, and the NOMA power-splitting structure.

4.4.2 Problem P2 – Fairness-Aware Optimization

Maximizing the instantaneous sum throughput may yield highly unequal rate allocation, where groups with favorable channels dominate the resources. To capture inter-group fairness, we introduce a long-term average rate $\bar{R}_g(t)$ for each C-DG g, computed over past time slots, and define Jain's fairness index $J(t) = \dfrac{\left(\sum_{g=1}^{G} \bar{R}_g(t)\right)^2}{G\sum_{g=1}^{G} \bar{R}_g^2(t)}, 0 \le J(t) \le 1$. Using a scalarized formulation, the fairness-aware problem is written as

$$(P2): \quad \max_{\{z_{g,k}, p_{g,s}, p_{g,w}\}} R_{\mathrm{sum}} - \lambda_{\mathrm{fair}}(1 - J(t)) \tag{4.17}$$

$$s.t. (C1) - (C6) \tag{4.18}$$

where $\lambda_{\mathrm{fair}} \ge 0$ controls the throughput–fairness tradeoff.

4.5 Proposed Solution

4.5.1 Motivation for Learning-Based Solution

Problems P1 and P2 are non-convex MINLPs with coupled interference, binary RB reuse decisions, and SIC decoding constraints. Classical approaches (e.g., exhaustive search, MINLP solvers, or successive convex approximation) become prohibitively complex as G increases and must be repeated per slot under time-varying channels. Therefore, we adopt a model-free MADRL approach based on PPO, which learns distributed policies through interaction with the environment and naturally handles stochastic fading, multi-agent coupling, and long-term fairness objectives.

4.5.2 Multi-Agent MDP Formulation

Due to decentralized operation and limited information exchange, the resource allocation problem is modeled as a Partially Observable Markov Decision Process (POMDP), where each C-D2D group acts as an autonomous learning agent. The environment corresponds to the underlying cognitive NOMA–D2D wireless network, including channel dynamics, interference coupling, CR protection constraints, and SIC decoding constraints. Although each agent observes only partial CSI, the environment maintains the full global network state.

Agents: Each C-D2D-NOMA group $g \in \mathcal{G} = \{1,2,\dots,G\}$ is modeled as an autonomous RL agent located at and representing the C-D2D transmitter of the group. The agent independently controls the RB reuse decision and the NOMA transmission parameters of its associated C-D2D transmitter.

State (observation) space: At the beginning of slot *t*, agent *g* observes a local CSI and partial interference state vector $\mathrm{s}_g(t)$ constructed from the system model. In compact form, $s_g(t) = \Big[\left|h_{g,s}^k(t)\right|^2,\ \left|h_{g,w}^k(t)\right|^2,\ \left|h_{g,b}^k(t)\right|^2,\ I_{g,w}(t-1),\ I_{g,s}(t-1), R_g(\mathrm{t}-1),\ I_{g,b}^k(t-1),\ \overline{I_{BS}}(t-1),\ \overline{R_g}(t-1)\Big]$.

The first three components represent local instantaneous channel power gains from the C-D2D transmitter of group g to its strong receiver, weak receiver, and the BS, respectively, and constitute local CSI available at the transmitter via channel reciprocity or receiver feedback. The terms $I_{g,w}(t\text{-}1)$ and $I_{g,s}(t\text{-}1)$denote the previous-slot locally experienced interference at the weak

and strong receivers, respectively, given by $I_{g,w}(t-1) = I_{g,w}^{\mathrm{CU}}(t-1) + I_{g,w}^{\mathrm{D2D}}(t-1)$, $I_{g,s}(t-1) = I_{g,s}^{\mathrm{CU}}(t-1) + I_{g,s}^{\mathrm{D2D}}(t-1)$, $R_g(\mathrm{t}-1)$ which are measured at the receivers and fed back to the associated transmitter. The term $R_g(t$ -1) denotes the effective C-D2D throughput achieved by group g in the previous slot. The quantity $\overline{I_{BS}}(t-1) = \frac{1}{K}\sum_{k=1}^{K} I_{g,b}^{k}(t-1)$ represents the mean aggregate interference at the BS across all RBs**,** computed from previous-slot measurements and broadcast to all C-D2D transmitters as a low-rate global congestion indicator. Finally, $\overline{R_g}(t-1)$ is a smoothed throughput term used in P2. For P1, the fairness component is omitted. All observation features are normalized before being fed to the policy network. The resulting observation corresponds to local CSI with limited broadcast interference statistics, reflecting practical decentralized operation under partial observability. The Partial CSI and Full CSI observation models are formally defined and discussed in Section V.

Action space: At each slot, agent *g* selects an action $a_g(t) \leftrightarrow \left(k_g(t),\, P_g(t)\right)$, where $k_g(t)$ denotes the reused RB, and $P_g(t)$ is discretized into N_P power levels.

State transition: Given the current global network state and the joint action profile $\{a_g(t)\}_{g=1}^{G}$, the environment computes the resulting interference levels, evaluates SINRs and feasibility constraints, and determines the instantaneous rewards. The channel coefficients then evolve according to the block-fading model by sampling new small-scale fading realizations for slot $t+1$. Based on the updated network state, the environment generates the next local observations $\mathrm{s}_g(t+1)$ for all agents.

4.5.3 PPO-Based MADRL Implementation

PPO is adopted due to stable on-policy updates in stochastic environments. The actor–critic framework learns a stochastic policy $\pi_\theta(a_g \mid s_g)$ and a value function $V_\phi(s_g)$ using generalized advantage estimation. PPO optimizes the clipped surrogate objective $\mathbb{E}[min(\rho_t(\theta)A_t,\, clip(\rho_t(\theta), 1-\epsilon, 1+\epsilon)A_t)]$, where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ and A_t is the advantage estimate [32],[34]. An entropy term is included to encourage exploration. We evaluate two architectures:

- CTDE**:** During training, the critic is augmented with global state information $s(t)$, while the actor relies solely on local observation $o_g(t)$ to preserve decentralized execution. training, while execution remains fully decentralized using local observations.
- DTDE: each agent g maintains its own PPO policy π_{θ_g} and value V_{ϕ_g} networks and learns independently from its local trajectories.

The overall multi-agent DRL framework, including agent–environment interaction and the CTDE/DTDE training architectures, is illustrated in Fig. 4.2.

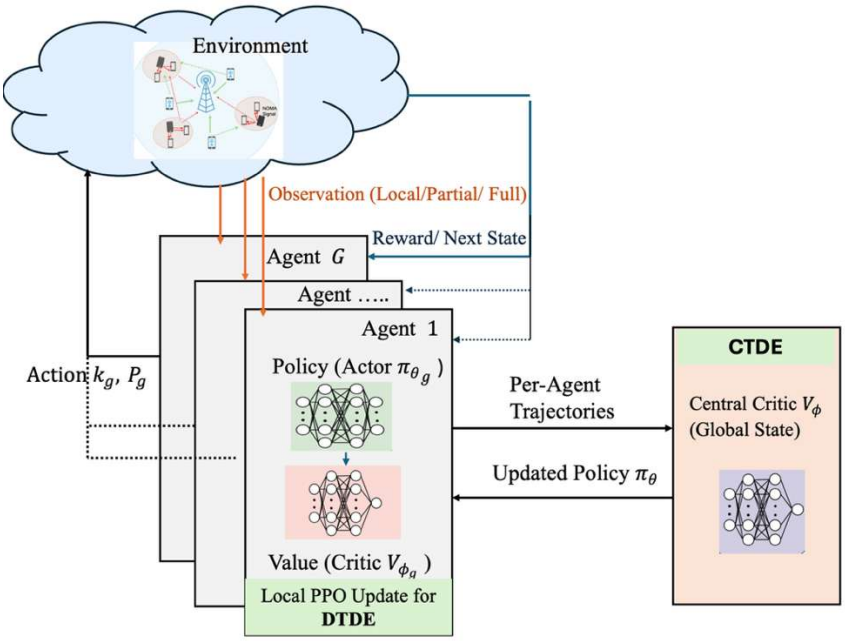


Figure 4.2: Multi-agent DRL framework for cognitive NOMA-D2D resource allocation under CTDE and DTDE architectures.

4.5.4 Reward Design

Constraint handling is implemented through a feasibility-gated reward mechanism. Specifically, if the selected transmission configuration violates the cellular protection or SIC decoding constraints, the effective throughput is set to zero. In addition, penalty terms are imposed to further discourage infeasible actions during learning.

P1: Constrained Sum-Rate Maximization: For P1, the reward of C-DG g at slot t is

$$r_g^{(1)}(t) = R_g^{eff}(t) - \lambda_{\text{CU}}\, 1_{\{C3_g \text{ violated}\}} - \lambda_{\text{SIC}}\, 1_{\{C4_g \text{ violated}\}} \tag{4.19}$$

where $\lambda_{\text{CU}}, \lambda_{\text{SIC}} > 0$ are penalty weights associated with violations of the cellular protection and SIC decoding constraint, respectively. The effective throughput $R_g^{eff}(t)$ is given by $R_g^{eff}(t) = \begin{cases} R_{\text{g}}(t)\text{, if all constraints are satisfied,} \\ 0, \text{ otherwise} \end{cases}$, where this formulation enforces feasibility in two ways: infeasible actions yield zero effective throughput, and explicit penalties further discourage violations during learning.

P2: Fairness-Regularized Sum-Rate Maximization: Each group maintains an Exponential Moving Average (EMA) throughput estimate $\bar{R}_g(t) = \tau \bar{R}_g(t-1) + (1-\tau) R_g(t), 0 < \tau < 1$. Jain's index $J(t)$is computed from $\{\bar{R}_g(t)\}_{g=1}^{G}$, and the fairness-shaped reward is defined as

$$r_g^{(2)}(t) = r_g^{(1)}(t) - \lambda_{\text{fair}}\big(1 - J(t)\big) \tag{4.20}$$

where τ is the fairness EMA and the fairness penalty is common to all agents at time t, promoting cooperative behavior. For the Hybrid-DTDE architecture, $J(t)$ is replaced by a neighborhood fairness index $J_{\mathcal{G}_g}(t)$ computed over a subset of neighboring agents $\mathcal{G}_g \subset \{1, \dots, G\}$, with $| \mathcal{G}_g | < G$. This setting tests the robustness of decentralized resource allocation when full global coordination is impractical.

Training is conducted over multiple N_{ep} episodes, where agents interact with the environment for a fixed horizon and collect on-policy trajectories. PPO updates are performed using generalized advantage estimation and mini-batch optimization. During evaluation, the trained policies are executed deterministically, and performance metrics are averaged over independent small-scale fading realizations.

4.5.5 Training and Evaluation Procedure

For both optimization settings P1 and P2, training follows a multi-agent PPO framework. The environment parameters—including network topology, path-loss model, noise power, and SINR thresholds—are fixed prior to training, and all policy and value networks are randomly initialized. Training proceeds over N_{ep} episodes, each consisting of a fixed horizon T time slots. At each slot, agent g observes its local state, selects an action according to its stochastic policy,

and receives a reward determined by the corresponding objective and constraint satisfaction. For P1, the reward corresponds to constrained sum-rate maximization with feasibility-gated throughput and additional penalties for CR and SIC violations. For P2, a fairness-regularized reward is employed based on the exponentially weighted average throughput. The collected trajectories from all agents are stored in an on-policy buffer. PPO computes Generalized Advantage Estimates (GAE) and updates the policy and value networks using the clipped surrogate objective. In the CTDE configuration, actors operate on local observations while the critic is augmented with global state information during training. In DTDE, each agent independently updates its own policy and value networks using only locally observed trajectories. The detailed training procedure is summarized in Algorithm 4.1. During evaluation, the trained policies are executed deterministically. Performance metrics are averaged over multiple Monte Carlo trials with independent small-scale fading realizations.

P1 and P2 are non-convex MINLPs with coupled interference, binary RB reuse decisions, and SIC decoding constraints. Classical approaches (e.g., exhaustive search, MINLP solvers, or successive convex approximation) become prohibitively complex as G increases and must be repeated per slot under time-varying channels. Therefore, we adopt a model-free MADRL approach based on PPO, which learns distributed policies through interaction with the environment and naturally handles stochastic fading, multi-agent coupling, and long-term fairness objectives.

Algorithm 4.1: Multi-Agent PPO-Based Resource Allocation for Cognitive NOMA–D2D Networks

Input:

– Environment type $\phi \in$ {P1 (sum-rate), P2 (fairness)}

– Number of agents G, episode horizon T, number of episodes $\boldsymbol{N_{ep}}$

– PPO parameters: learning rate, discount factor, GAE parameter, clip ratio, entropy weight, batch size, epochs

Output:

– Trained policies $\boldsymbol{\pi_{\theta_g}}$ (DTDE/Hybrid-DTDE) or $\boldsymbol{\pi_{\theta}}$ (CTDE) for all C-D2D transmitters**.**

Initialization:

– Initialize system topology, CU RB allocation (OMA), channel/path-loss, SINR thresholds,

and CR constraints. Initialize policy and value networks (shared for CTDE, per-agent for DTDE). For P2, initialize EMA rates $\overline{\boldsymbol{R_g}}(\mathbf{0})$.

1) **for** episode $\boldsymbol{e = 1}$ **to** $\boldsymbol{N_{ep}}$ **do**

2) Reset environment; sample new small-scale fading; reset fairness as needed.

3) **for** time slot $\boldsymbol{t = 0, \ldots, T-1}$ **do**

4) Each agent g Observe local state $\boldsymbol{s_g(t)}$

5) Each agent sample action $\boldsymbol{a_g(t)}$ according to the policy.

6) Environment computes SINRs, achievable rates, and interference levels.

7) Evaluate constraint violations and apply penalties.

8) Compute reward $\boldsymbol{r_g(t)}$ according to

–For P1: $\boldsymbol{r_g^{(1)}(t) = R_g(t) - \lambda_{\text{CU}}\, \mathbf{1}_{\{\text{CU violated}\}} - \lambda_{\text{SIC}}\, \mathbf{1}_{\{\text{SIC violated}\}}}$.

–For P2: update EMA rate $\overline{\boldsymbol{R_g}}(\boldsymbol{t+1})$; compute Jain's index J(t)

$$\boldsymbol{r_g^{(2)}(t) = r_g^{(1)}(t) - \lambda_{\text{fair}}\big(1 - J(t)\big)}$$

9) Store trajectories $(\boldsymbol{s_g(t), a_g(t), r_g(t), s_g(t+1)})$ on-policy rollout buffer.

10) **End for** (time slots)

11) Compute returns and advantages using GAE

12) **for** k = 1 to K epochs **do**

13) Sample the mini-batches.

14) Update policy parameters by maximizing the PPO clipped surrogate objective with entropy regularization**.**

15) Update value function.

16) **End** for (PPO epochs)

17) **End** for (episodes)

18) **Return** trained policy network(s) for decentralized execution.

4.5.6 Computational Complexity

The optimization problems P1 and P2 involve joint RB assignment and transmit-power allocation. Classical approaches such as exhaustive search or successive convex approximation become computationally prohibitive, since the search space grows exponentially with the number

of C-D2D groups G. For K RBs and N_P transmit-power levels, the complexity scales as $\mathcal{O}((KN_P)^G)$, rendering per-slot optimization impractical in large-scale or time-varying networks.

In contrast, the proposed PPO-based framework learns a parametric policy that directly maps observed states to RB and power decisions. The primary computational cost arises from neural network updates, with per-update complexity $\mathcal{O}(BN_\theta)$ for CTDE and $\mathcal{O}(GBN_\theta)$ for DTDE, where N_θ is the number of trainable parameters, and B is the batch size. This results in polynomial scaling with G, providing a scalable alternative to classical optimization for cognitive NOMA–D2D resource allocation.

4.6 Simulation Setup

The simulation setup used to evaluate the proposed schemes is summarized in Table 4.2. In the following, the term environment refers to a specific DRL training setting corresponding to problems P1 and P2, each defined by its reward function, state variables, and system dynamics, while sharing the same underlying physical network model. We consider a single-cell uplink network where each CU is pre-assigned one RB via OMA. For each C-D2D group, the transmitter is uniformly distributed within the cell, and its two receivers are placed at random angles with link distances independently sampled from $[0, R_{\mathrm{D2D}}]$, where $R_{\mathrm{D2D}} = 50\mathrm{m}$. Node locations are generated once and fixed across experiments, while small-scale fading is independently resampled at each time slot. The transmit power is discretized into N_P levels unless otherwise specified. To evaluate scalability, the number of active C-D2D groups is varied as $G \in \{10,20,30,40,50,60\}$. For each configuration (P1/P2 and CTDE/DTDE), performance metrics are averaged over multiple Monte Carlo episodes. The following metrics are reported: average total C-D2D sum-rate, Jain's fairness index across C-D2D groups, EE, CU outage probability, and SIC success probability at strong C-DRs. All environments are implemented as Gym/Gymnasium-compatible multi-agent environments and trained using Ray RLlib [30]. The PPO agent was implemented using an actor–critic neural network architecture. Both the actor and critic were modeled as fully connected feedforward neural networks with two hidden layers, each containing 256 neurons. The tanh activation function was used in the hidden layers. The policy and value networks were optimized using the Adam optimizer. All simulations are implemented in Python 3.9 using Ray/RLlib 2.x

with a PyTorch backend. Experiments are executed on a single desktop machine (Apple iMac with an M1 processor and 16 GB RAM) using a single CPU worker and no GPU acceleration.

Table 4.2: Simulation parameters for cognitive NOMA–D2D environments

Parameter	Value
The radius of the macro base station R_{cell}	400 m
The number of CUs M	50
The number of C-D2D groups G	10-60
short-range D2D links R_{D2D}	50 m
RB bandwidth B_k	180 kHz
Noise power per RB	-104 dBm
CU transmit power P_m	23 dBm
C-D2D transmit power range P_g	$10 - 23$ dBm
Pathloss exponent	3.5
SINR thresholds $\gamma_{\text{CU,th}}$	7 dB
Minimum SIC decoding requirement $\gamma_{\text{SIC,min}}$	4.8 dB
Number of transmit power levels N_{P}	10
NOMA power-splitting ratio α_g	0.9
EMA coefficient τ	0.85
Fairness penalty coefficient λ_{fair}	1,2,5,10
PPO learning rate	10^{-5}
PPO discount factor γ	0.95
PPO GAE parameter λ	0.95
PPO entropy coefficient	0.01
PPO clipping ratio	0.3
PPO train batch size	3072 time steps

4.6.1 Baseline Learning Algorithms

To evaluate the effectiveness of the proposed multi-agent PPO framework, we compare its convergence behavior and performance against several baseline learning algorithms. In addition

to PPO**,** DQN and IMPALA are implemented as representative DRL baselines. All learning algorithms use identical environment settings, observation spaces, and reward functions. Hyperparameters for each algorithm are individually tuned to ensure stable convergence.

- DQN**:** A fully connected neural network is used to approximate the action–value function. Experience replay and a target network are employed to stabilize training [102], [104].
- IMPALA**:** A distributed actor–learner architecture with V-trace off-policy corrections is adopted to enable scalable learning with asynchronous updates [105].

In addition to convergence analysis, other performance metrics, including sum-rate**,** fairness, and EE, are evaluated for the proposed multi-agent PPO framework and compared against the following baseline schemes.

- RS: RBs and transmit power levels are selected uniformly at random within feasible ranges.
- GA: A population-based evolutionary optimization scheme is employed to jointly optimize RB allocation and power levels.

The performance of the proposed P1 scheme is evaluated under CTDE and DTDE, and P2 is evaluated under hybrid-DTDE learning architectures. All baseline schemes are tested under identical channel realizations and system parameters to ensure a fair comparison.

4.6.2 Observation Models and CSI Availability

To study the impact of CSI availability on learning performance, we consider three observation models, namely Local CSI, Partial CSI, and Full CSI. Under the Local CSI model, each C-DG agent observes only its local channel information and limited global interference statistics, which were described in Section IV. For the Partial CSI model, in addition to Local CSI, agents observe causal global indicators, including previous slot CU SINR per RB $\gamma_{m,b}^{k}(t-1)$ and RB load fraction $\frac{1}{G}\sum_{g=1}^{G} z_{g,k}\,(t-1)$. Full cross-channel matrices are not available. This represents semi-distributed coordination with limited feedback. For the Full CSI model, in addition to Local CSI, Agents additionally observe per-RB channel information related to primary links, including: CU–BS channel gains $|\,h_{m,b}^{k}\,|^2$, CU to C-D2D channel gains $|\,h_{m,w}^{k}\,|^2$and $|\,h_{m,s}^{k}\,|^2$, aggregate BS interference $I_{b}^{k}(t-1)$ and RB load fraction $\frac{1}{G}\sum_{g=1}^{G} z_{g,k}\,(t-1)$. This model

approximates near-global information and serves as a performance upper bound among distributed settings. The achievable-rate expressions remain identical across different CSI settings, as they are evaluated using the realized network state. The transmission-rate expressions in (4.5) – (4.8) are identical under all CSI settings, as they depend on the realized channel state. CSI availability affects only the information available to agents for decision-making, thereby influencing learned policies, convergence behavior, and constraint satisfaction.

4.7 Performance Evaluation

4.7.1 Convergence and DRL comparison

Figure 4.3 compares PPO, IMPALA, and DQN under DTDE for G=30. PPO achieves the highest steady-state performance with smooth and stable convergence. IMPALA converges more slowly and stabilizes at a lower reward. DQN exhibits faster initial learning but noticeable oscillations. PPO therefore provides the best tradeoff between convergence stability and achievable performance and is adopted in the remainder of this study. The reported training curves correspond to the mean episode reward, which reflects the feasibility-gated reward formulation including zero-throughput masking and penalty terms for CU and SIC violations, rather than raw sum-rate.

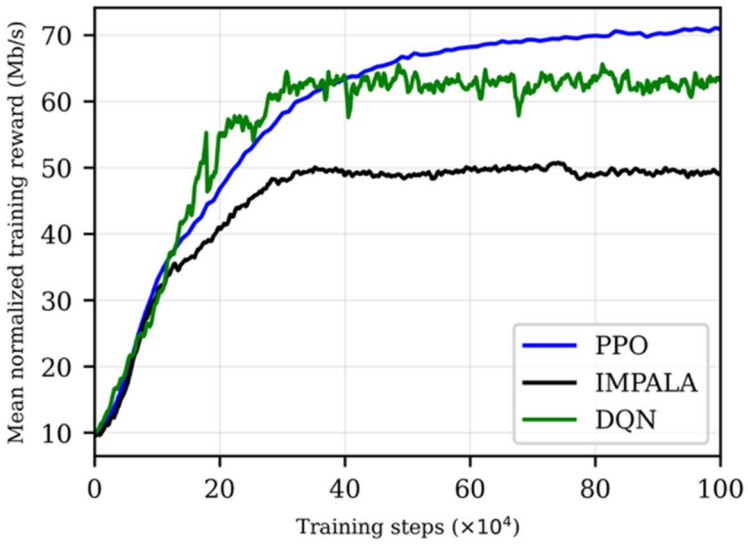


Figure 4.3: Learning convergence comparison of PPO, IMPALA, and DQN under DTDE with identical observation models and reward functions.

Figure 4.4 shows the convergence of DTDE under different CSI observation models. Local-CSI leads to slower convergence and lower steady-state performance due to limited observability, particularly as the number of C-D2D groups increases. Partial-CSI, which augments local information with CU SINR and RB load indicators, achieves faster convergence and improved performance. Full-CSI yields the fastest convergence and the highest sum-rate, serving as an upper performance bound. However, its signaling overhead makes it less practical in large-scale systems. For all observation models, increasing the number of D2D groups from $G = 30$ to $G = 60$, decreases the reward due to increased co-channel interference and RB contention. The performance gap among the three observation models highlights the impact of CSI availability on decentralized learning. Although all models eventually reach similar steady-state sum rates, the convergence speed differs significantly. These results emphasize the importance of selecting an informative observation state that captures the most relevant features, thereby reducing training complexity and accelerating convergence. As we can see in the result, partial CSI achieves performance close to the full-CSI benchmark, even at higher C-D2D density.

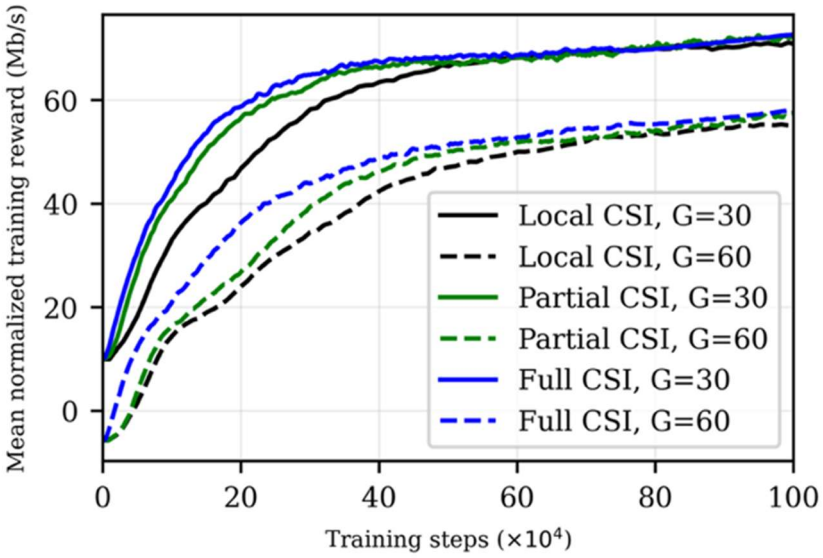


Figure 4.4: Convergence of the mean episode reward versus training steps for different observation models under DTDE. Results are shown for local CSI, partial CSI, and full CSI with $G = 30$ and $G = 60$ C-D2D groups.

4.7.2 Impact of CSI and Network Density

Figure 4.5 compares the effective sum-rate versus G. All MADRL schemes outperform GA and RS across network densities. DTDE with full CSI and CTDE with full CSI achieve the highest sum-rate, serving as performance upper bounds. Importantly, DTDE with limited CSI maintains competitive performance despite decentralized operation. For DTDE, at $\mathrm{G} = 50$, partial CSI incurs only a 2.63% reduction relative to the full-CSI configuration. Even under strictly local CSI, the performance degradation is limited to 10.62% relative to the full-CSI sum-rate. These results confirm that appropriately structured local and partial observation models can preserve near-centralized performance while significantly reducing signaling overhead.

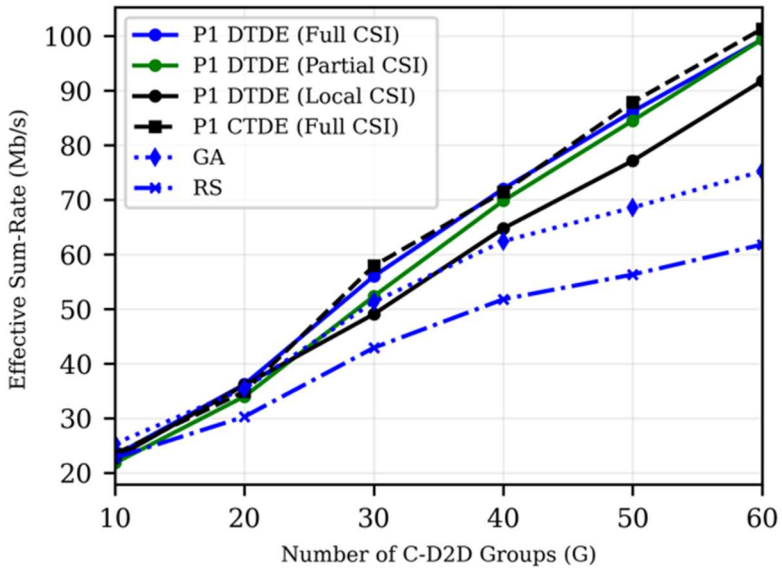


Figure 4.5: The effective sum rate versus the number of C-D2D groups

Figure 4.6 illustrates EE. MADRL methods consistently outperform GA and RS across all densities. Interestingly, DTDE with partial CSI achieves superior EE performance despite not attaining the highest sum-rate. At $G = 50$, partial CSI provides an 85% improvement over the full-CSI benchmark. These results demonstrate that observation-space design directly influences learned power allocation strategies and resulting energy efficiency. Although the reward function remains identical across all settings, modifying the observation space alters the information structure available to each agent. Consequently, the learned policies differ in coordination

capability, interference awareness, and power adaptation behavior. This confirms that observation-state design is as critical as reward engineering in multi-agent DRL systems for balancing sum-rate performance, energy efficiency, signaling overhead, and practical deployment constraints.

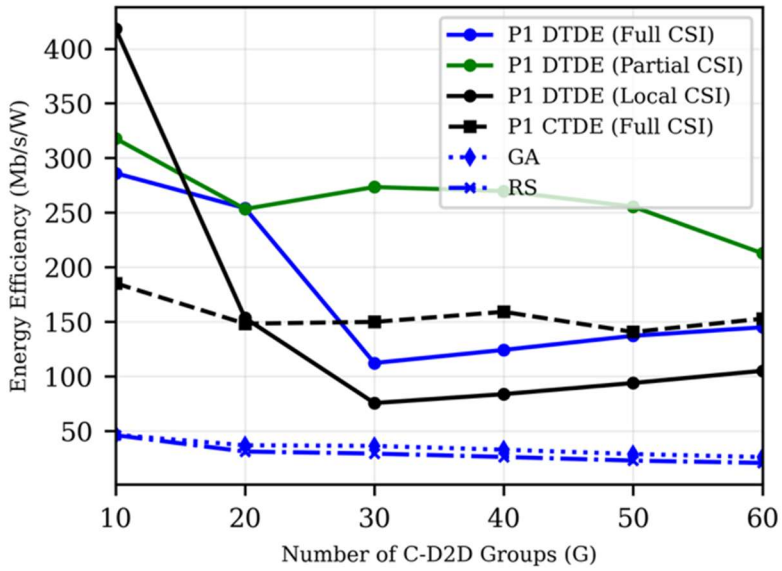


Figure 4.6: The effective EE versus the number of C-D2D groups.

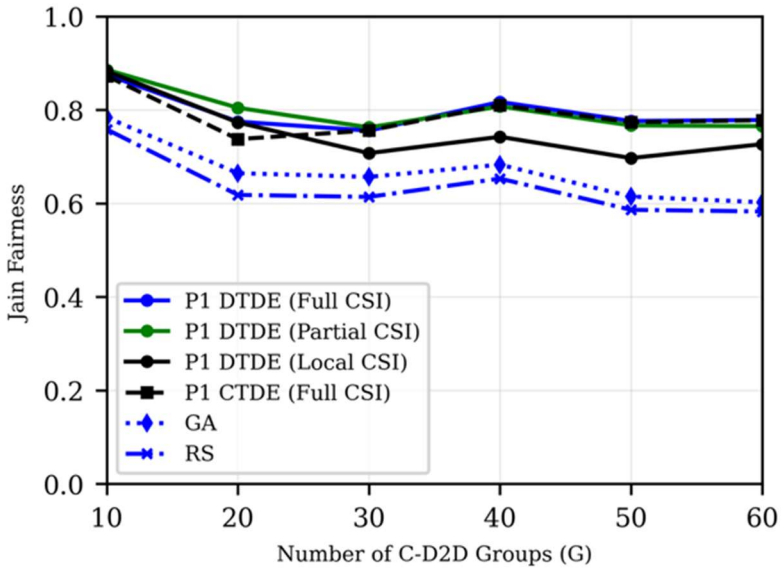


Figure 4.7: The Jain fairness versus the number of C-D2D groups.

Figure 4.7 shows Jain's fairness index versus G. MADRL schemes maintain relatively stable fairness as the network density increases. DTDE with full and partial CSI achieves the highest fairness levels, whereas local CSI shows slight performance degradation due to limited coordination. GA and RS experience significant fairness degradation at higher network densities.

Figures 4.8 and 4.9 illustrate the cellular outage probability and the SIC decoding success rate, respectively, as functions of the number of C-D2D groups G. As shown in Figure 4.8, the proposed MADRL schemes maintain a consistently low cellular outage probability across all network sizes, demonstrating effective protection of the primary CUs under the underlay CR paradigm. In contrast, GA and RS exhibit a noticeable increase in outage probability as G increases, reflecting their limited capability to adapt transmission decisions to interference constraints in dense deployments. Figure 4.9 further shows that the MADRL-based approaches achieve near-perfect SIC decoding success rates across all operating points. This confirms that the reward formulation successfully enforces the SIC constraint during learning. Although minor degradation appears for heuristic baselines at higher densities, the learning-based schemes maintain robust decoding reliability.

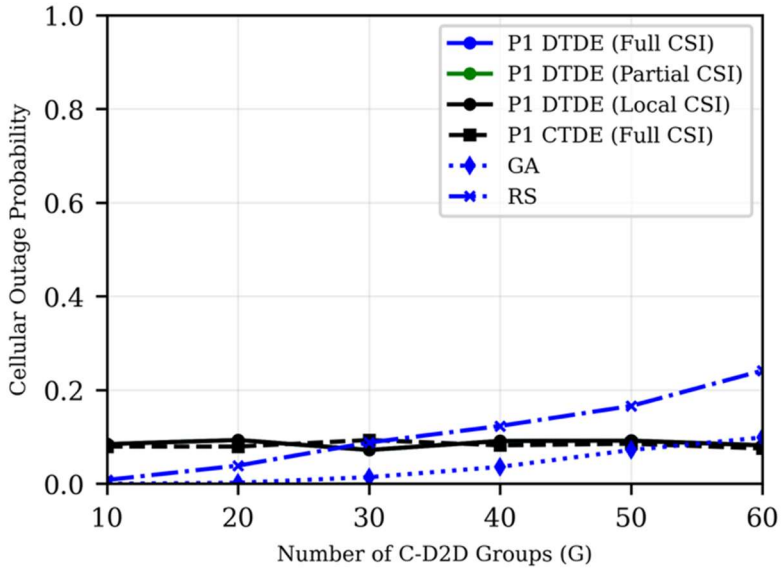


Figure 4.8: The cellular outage probability versus the number of C-D2D groups.

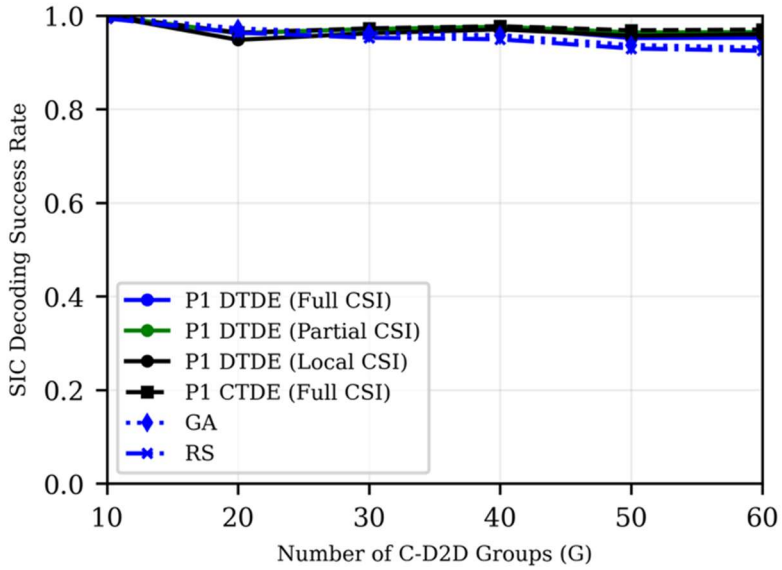


Figure 4.9: The SIC decoding success rate versus the number of C-D2D groups.

Figure 4.10 illustrates the ratio of C-D2D groups whose selected transmission configurations violate either the cellular protection or SIC decoding constraints. A group is counted as violating if its RB reuse and power allocation result in excessive interference to the CU or insufficient SINR for SIC decoding. Such violating groups are treated as infeasible and do not contribute to effective throughput. The decreasing violation ratio with training demonstrates that the learned policy progressively avoids infeasible configurations. The proposed MADRL methods consistently achieve substantially lower violation ratios compared to GA and RS, even under dense interference conditions, while partial CSI approaches the reliability of full CSI. The violation ratio increases with network density due to intensified interference and tighter spectrum reuse conditions. These results indicate that the proposed learning-based framework achieves a high level of constraint satisfaction while maintaining scalability to larger network densities.

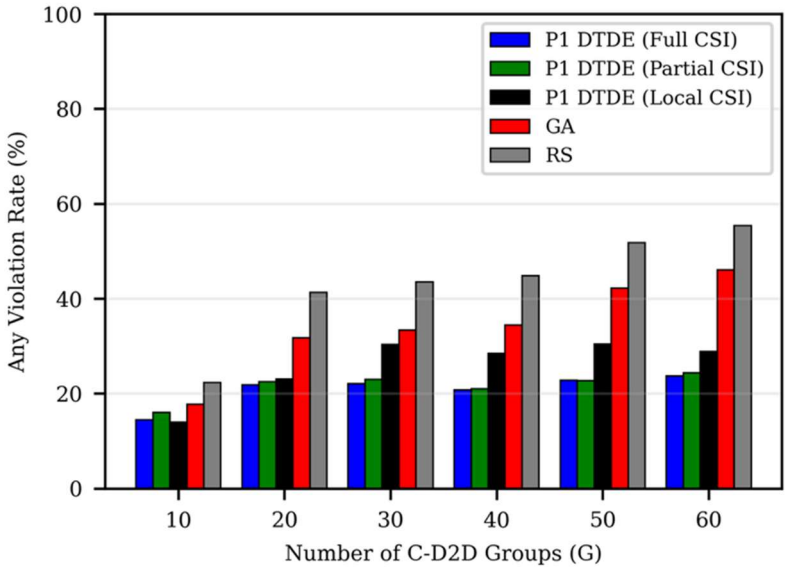


Figure 4.10: Ratio of C-D2D groups violating the cellular protection or SIC decoding constraints versus the number of C-D2D groups.

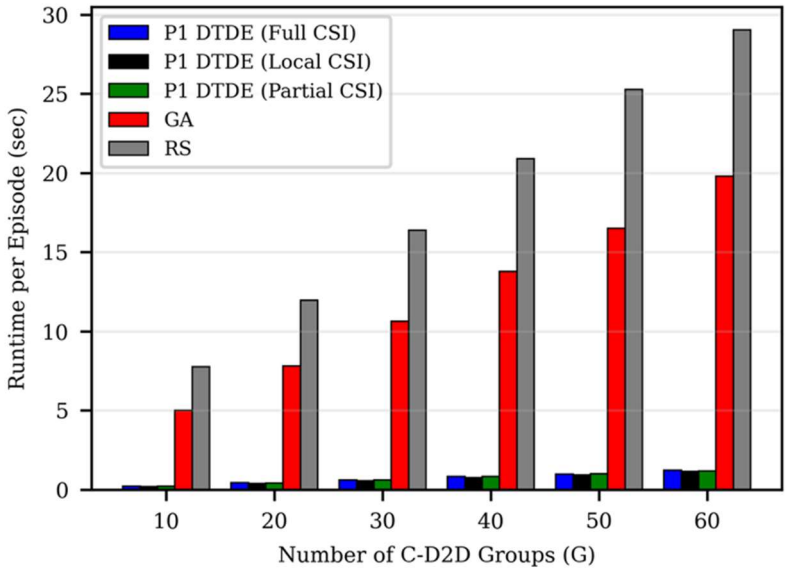


Figure 4.11: Runtime per episode versus the number of C-D2D groups.

Figure 4.11 illustrates the average runtime per episode during the evaluation phase as a function of the number of C-D2D groups G. The reported values correspond to inference time only, excluding the offline training cost. The proposed MADRL schemes exhibit significantly

lower runtime compared to GA and RS across all network sizes. The evaluation complexity of the DRL-based approaches increases moderately with G, reflecting the forward pass of the neural network for each agent. However, the growth is nearly linear and remains computationally efficient even at higher densities. In contrast, GA and RS demonstrate substantially higher runtime, with a steeper increase as G grows. This behavior is attributed to their iterative search and combinatorial exploration mechanisms, which scale poorly in dense network scenarios. These results highlight an important practical advantage of the proposed framework: although training is performed offline, the learned policies enable fast real-time decision-making with low computational overhead, making the approach suitable for dynamic spectrum management in dense C-D2D-enabled networks.

4.7.3 Fairness–Shaping Analysis

To analyze P2, the fairness coefficient λ_{fair} is varied under Hybrid-DTDE and compared with P1 DTDE (local CSI). To investigate the impact of fairness shaping, we vary the fairness coefficient λ_{fair} and evaluate system performance for P2 for Hybrid-DTDE (local CSI) with P1 DTDE (local CSI).

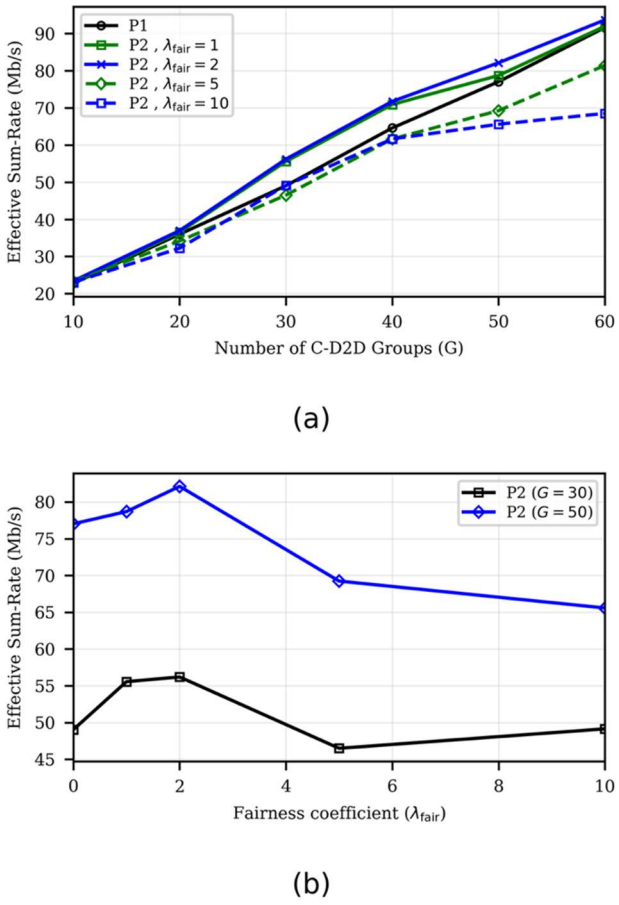
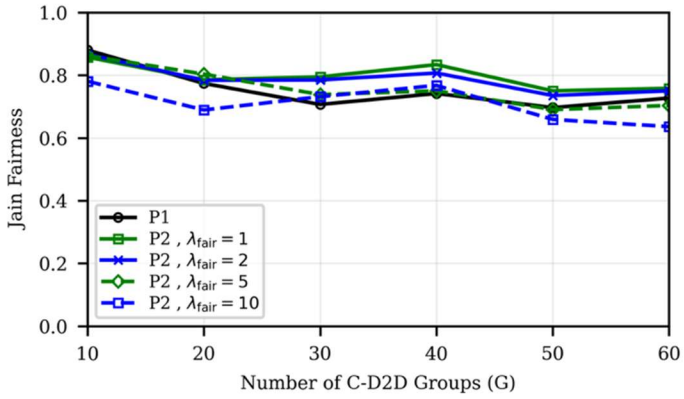
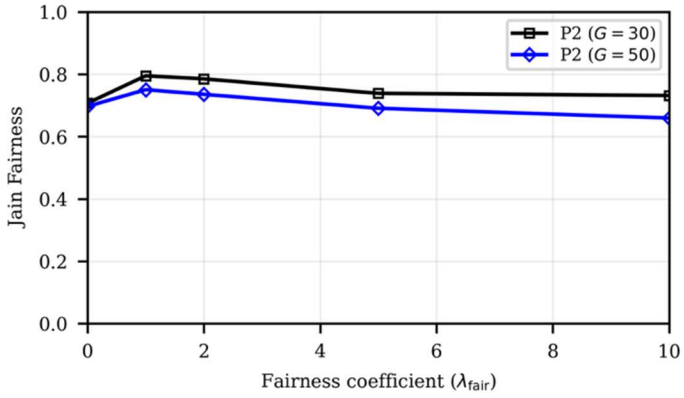


Figure 4.12: a) Effective sum-rate versus the number of C-D2D groups G for P1, $\lambda_{fair} = 0$ and P2, $\lambda_{fair} > 0$. b) Effective sum-rate for P2 versus fairness coefficient λ_{fair} for $G = 30$ and $G = 50$.

Figure 4.12 shows the effective sum-rate versus G and λ_{fair}. Moderate fairness coefficients ($\lambda_{fair} = 1,2$) achieve a higher effective sum-rate than P1 ($\lambda_{fair} = 0$). This indicates that fairness regularization improves inter-group coordination and mitigates excessive power concentration, thereby reducing interference. At $\mathrm{G} = 30$, the effective sum-rate increases by 13.8% relative to $\lambda_{\text{fair}} = 0$. At $\mathrm{G} = 50$, the corresponding gain is 6%. However, large fairness weights degrade throughput due to overly restrictive power redistribution. The resulting behavior is non-monotonic, with $\lambda_{\text{fair}} = 2$ providing the best tradeoff. This non-monotonic trend demonstrates that moderate fairness regularization enhances interference management and coordination stability, whereas excessive fairness weighting restricts throughput optimization.



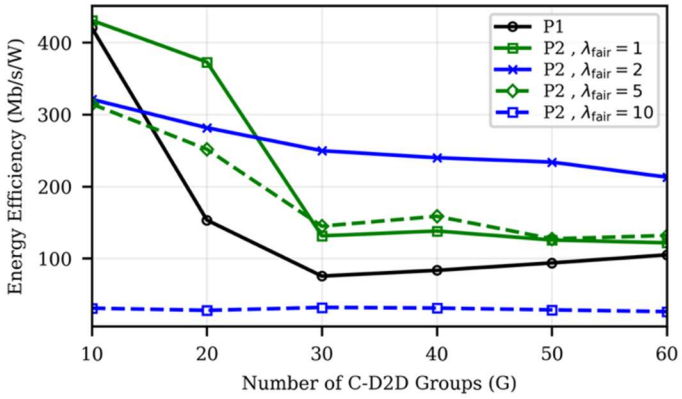
(a)



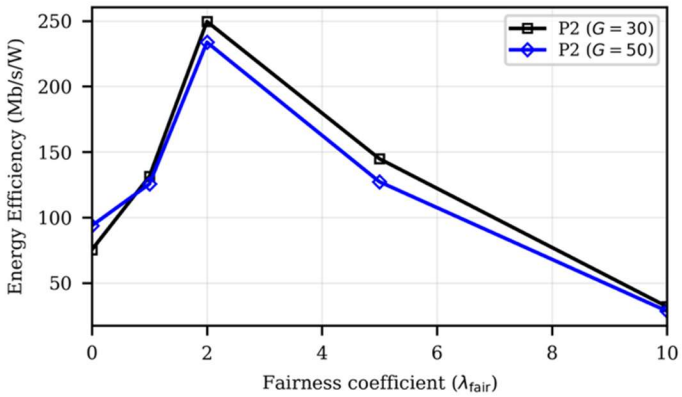
(b)

Figure 4.13: a) Jain's fairness versus the number of C-D2D groups G for P1, $\lambda_{fair} = 0$ and P2, $\lambda_{fair} > 0$. b) Jain's fairness for P2 versus the fairness coefficient λ_{fair} for $G = 30$ and $G = 50$.

Figure 4.13 shows Jain's fairness index versus G and versus λ_{fair}. Moderate fairness shaping improves fairness, particularly at higher network densities. Relative to the P1, at $G = 30$, Jain's index increases by 13.5%. At $G = 50$, the corresponding gain is 6.3%. However, excessively large λ_{fair} yields diminishing returns and may slightly reduce fairness stability due to constrained resource flexibility. Overall, these results confirm that moderate fairness shaping improves fairness without severely compromising performance, while excessive fairness weighting may negatively affect both fairness stability and overall system efficiency.



(a)



(b)

Figure 4.14: a) EE versus the number of C-D2D groups G for P1, $\lambda_{fair} = 0$ and P2, $\lambda_{fair} > 0$. b) EE for P2 versus fairness coefficient λ_{fair} for $G = 30$ and $G = 50$.

Figure 4.14 illustrates EE versus G and λ_{fair}. Moderate fairness shaping ($\lambda_{fair} = 2$) improves EE relative to the P1 by promoting more balanced power allocation and reducing aggressive transmissions. At $G = 30$, EE increases by 228%, while at $G = 50$, the corresponding gain is 148%. However, excessive fairness weighting reduces EE due to overly conservative power redistribution and reduced transmission flexibility. The resulting EE behavior is non-monotonic, with moderate fairness regularization achieving the best balance between throughput and energy consumption. Overall, these results demonstrate that fairness-aware reward shaping not only

improves rate distribution but also enhances interference coordination and EE when appropriately tuned.

4.8 Conclusion

This chapter investigated joint RB reuse and NOMA power allocation for cognitive NOMA-enabled D2D networks under CR protection and SIC decoding constraints. The resulting optimization problems are highly non-convex due to coupled interference, discrete RB reuse decisions, and power-splitting interactions. To address these challenges, a multi-agent PPO-based framework was developed to enable distributed learning of RB and power control policies without repeated real-time optimization. Two objectives were considered: constrained sum-rate maximization (P1) and fairness-regularized sum-rate maximization (P2). Both CTDE and DTDE learning architectures were evaluated under different CSI availability assumptions. The results demonstrate that decentralized learning with a structured observation design can approach centralized full-CSI performance while significantly reducing signaling overhead. These results highlight the importance of observation-state design in scalable multi-agent wireless systems. Fairness shaping exhibits a non-monotonic tradeoff: moderate fairness regularization improves sum rate, fairness, and energy efficiency, whereas excessive weighting restricts throughput optimization and degrades overall performance. Overall, the results emphasize that decentralized spectrum management in cognitive NOMA-D2D networks depends critically on information structure and constraint-aware learning design. This chapter represents the culmination of the dissertation by integrating spectrum awareness, learning-based decision-making, and advanced multiple access techniques into a unified intelligent framework for CRNs. While the proposed approach achieves significant performance gains, further extensions may consider additional practical constraints such as mobility, energy harvesting, and real-time adaptation in highly dynamic environments. These aspects are discussed in the final chapter as directions for future research.

Chapter 5

Conclusion and Future Work

This chapter concludes the dissertation by highlighting the key insights and discussing future research directions. The work presented in this dissertation addressed the problem of intelligent and adaptive resource allocation in CRNs through the integration of predictive spectrum sensing, MADRL, and NOMA.

The proposed framework was developed progressively, beginning with spectrum awareness, followed by learning-based decision-making, and culminating in a unified system-level optimization approach. This progression reflects the transition of CRNs from reactive systems toward intelligent and autonomous wireless systems.

5.1 Conclusion

This dissertation investigated intelligent and adaptive resource allocation in CRNs through a progressive integration of spectrum awareness, learning-based decision-making, and advanced multiple access techniques. The research was structured in three main chapters, reflecting the evolution of CR systems from sensing-driven operation toward fully autonomous and intelligent wireless networks.

The first chapter focused on enhancing spectrum awareness through PCSS. By integrating spectrum prediction with cooperative sensing, the proposed framework improved the accuracy of channel state estimation and enabled more efficient spectrum utilization. The results demonstrated that cooperative prediction significantly outperforms conventional sensing approaches, particularly when majority fusion is employed. This chapter also revealed an important tradeoff between SE and energy consumption, highlighting that while cooperation improves robustness and prediction accuracy, it introduces additional communication overhead.

The second chapter addressed the limitations of centralized and model-dependent approaches by introducing a MADRL framework for distributed resource allocation in C-D2D networks. By formulating the problem as a Markov decision process and employing a PPO-based learning strategy, the proposed approach enabled agents to learn efficient resource allocation

policies under dynamic and uncertain conditions. The results showed that MADRL significantly outperforms traditional optimization methods in terms of scalability, robustness, and adaptability. This chapter demonstrated that learning-based approaches eliminate the need for accurate system models and are well suited for complex wireless environments, although they introduce challenges related to reward design, training complexity, and coordination under partial observability.

The third chapter extended the framework by integrating NOMA into C-D2D communication, resulting in a unified system that jointly considers spectrum reuse, power allocation, and interference management under CR protection and SIC decoding constraints. The resulting optimization problems are highly non-convex due to coupled interference, discrete resource allocation, and power-splitting interactions. To address these challenges, a multi-agent PPO-based framework was developed to enable distributed learning of RB reuse and power control policies.

Two optimization objectives were considered: constrained sum-rate maximization (P1) and fairness-regularized sum-rate maximization (P2). A feasibility-gated reward mechanism was introduced to ensure that constraint violations, including cellular protection and SIC decoding, are properly handled during learning. In this formulation, infeasible actions result in zero effective throughput and are further penalized, ensuring consistency between achievable-rate evaluation and decoding feasibility. In addition, fairness shaping based on Jain's index was incorporated to balance throughput efficiency and fairness among users.

The results demonstrated that decentralized learning with a structured observation design can approach centralized full-CSI performance while significantly reducing signaling overhead. Furthermore, fairness shaping was shown to exhibit a non-monotonic trade-off: moderate fairness regularization improves system performance, while excessive weighting degrades throughput. These results highlight the importance of information structure and constraint-aware learning in enabling scalable and efficient decentralized resource allocation.

Overall, this dissertation presents a progressive development of intelligent CR systems, beginning with spectrum awareness, followed by learning-based decision-making, and culminating in advanced resource allocation using NOMA-enabled C-D2D communication. Each chapter addresses a critical aspect of intelligent wireless network design and collectively contributes toward the development of adaptive and autonomous communication systems.

As wireless networks evolve toward 6G and beyond, the role of AI in communication system design will continue to grow. The framework proposed in this dissertation provides a foundation for future research in intelligent and autonomous wireless systems, contributing to the development of efficient, scalable, and adaptive next-generation networks.

5.2 Future Work

Building on the contributions of this dissertation, several promising directions for future research can be identified.

5.2.1 Energy-Efficient and Energy-Harvesting Systems

Incorporating energy-harvesting techniques into CRNs can enhance sustainability and reduce reliance on conventional power sources. Future research may investigate joint optimization of energy consumption and resource allocation, where transmission decisions depend not only on channel conditions but also on available harvested energy. This is particularly relevant for IoT and large-scale wireless networks.

Energy-harvesting CRNs are particularly promising in rural or remote areas where traditional power sources are scarce, but wireless communication systems still need to be operational. By integrating energy harvesting, CRNs can reduce reliance on conventional power grids and further extend the operational lifetime of wireless devices. In the energy-harvesting phase, SUs harvest energy from various sources like radio frequency signals, solar, wind, or mechanical vibrations. This energy powers their transceivers for sensing and communication. In the spectrum sensing phase, SUs detect idle spectrum bands using the harvested energy and accurate sensing prevents interference with PUs. In the data transmission phase, if the spectrum is available, SUs transmit data using harvested energy, and the energy level is low, they may delay transmission or harvest more energy [106].

Enhancing the EE of an energy-harvesting enabled NOMA-based C-D2D communication systems, which could reduce power consumption and improve network sustainability. Moreover, the EE optimization problem can be explored. Energy-harvesting-powered D2D communication has been widely adopted in various scenarios, such as IoT services and machine-type communications. However, the dependence on harvested energy introduces a key consideration in

resource allocation: the available energy. This factor not only impacts decisions related to the transmit power but also affects spectrum assignment, which may vary depending on the energy levels [107]. With the growing focus on green communications and IoT, D2D communication in energy-harvesting scenarios has attracted significant attention. However, limited research has addressed EE during power allocation in NOMA-based D2D communication systems [108].

5.2.2 Multi-Objective Reinforcement Learning

While this dissertation considered fairness-aware optimization, future work may extend the framework to fully multi-objective reinforcement learning. This includes jointly optimizing multiple performance metrics such as spectrum efficiency, energy efficiency, fairness, and delay. This involves applying multi-objective RL to solve complex problems in C-D2D communications, where agents are trained to maximize multiple cumulative rewards—such as spectral efficiency and energy efficiency—simultaneously. In traditional RL, an agent learns to optimize a single objective or reward, such as maximizing the sum rate or minimizing delay. However, in real-world D2D networks, multiple objectives often need to be optimized simultaneously. For instance, a network might need to maximize spectral efficiency and EE at the same time. Multi-objective RL addresses this challenge by enabling agents to learn how to balance competing objectives through a single learning process. Instead of optimizing just one reward signal, the agent learns to optimize a combination of multiple rewards, which can be weighted according to their importance [109].

5.2.3 Integration with Emerging Technologies

The proposed framework can be extended by integrating emerging technologies such as Intelligent Reflecting Surfaces (IRS), 6G communication systems, and edge intelligence. IRS can enhance signal propagation and interference management, while edge intelligence can enable distributed and low-latency learning. These technologies offer new opportunities for improving system performance and adaptability. Building on the comprehensive analysis of IRS-assisted NOMA-based networks in [110], the integration of IRS in the proposed NOMA-based C-D2D network is considered a promising technology for 6G networks.

5.2.4 Secure Cognitive D2D Communication

Security is an important concern in decentralized C-D2D networks. Future work may incorporate physical layer security techniques to protect against eavesdropping and malicious interference. Integrating security considerations into the learning framework can enhance the reliability and robustness of CR systems.

In D2D communication, where devices communicate directly without going through the base station, security becomes a critical concern due to the potential for malicious nodes to interfere with communication. These malicious nodes can eavesdrop on or intentionally disrupt the signal transmission, leading to compromised data integrity and privacy. Physical layer security aims to protect communication by exploiting the physical characteristics of the wireless channel, such as signal strength, noise, and interference [111]. By integrating physical layer security into the C-D2D framework, the system can improve the confidentiality and integrity of communication, even in the presence of malicious nodes.

5.2.5 Mobile Environments

Extending the proposed framework to dynamic and high-mobility environments is an important direction. In practical wireless systems, user mobility significantly affects system performance. Future work may consider adaptive grouping strategies, dynamic role assignment in NOMA systems, and real-time learning mechanisms that can track rapid environmental changes.

Bibliography

[1] N. Siddique, S. F. Hasan, and S. M. Zabir, *Opportunistic Networking: Vehicular D2D and Cognitive Radio Networks*. Boca Raton, FL, USA: CRC Press, 2017.

[2] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, 2020.

[3] N. A. Khalek, D. H. Tashman, and W. Hamouda, "Advances in machine learning-driven cognitive radio for wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 2, pp. 1201–1237, 2024.

[4] M. Ibnkahla, *Cooperative Cognitive Radio Networks: The Complete Spectrum Cycle*. Boca Raton, FL, USA: CRC Press, 2014.

[5] N. Meghanathan and Y. B. Reddy, *Cognitive Radio Technology Applications for Wireless and Mobile Ad Hoc Networks*. Hershey, PA, USA: IGI Global, 2013.

[6] T. S. Dhope, *Cognitive Radio Networks Optimization with Spectrum Sensing Algorithms*. Aalborg, Denmark: River Publishers, 2015.

[7] X. Xing, T. Jing, W. Cheng, Y. Huo, and X. Cheng, "Spectrum prediction in cognitive radio networks," *IEEE Wireless Commun.*, vol. 20, no. 2, pp. 90–96, 2013.

[8] R. Nagarajan, D. R. Dhaya, and K. Suriyan, *Spectrum and Power Allocation in Cognitive Radio Systems*. Hershey, PA, USA: IGI Global, 2024.

[9] A. Khattab, *Cognitive Radio Networks: From Theory to Practice*. New York, NY, USA: Springer, 2013.

[10] N. Shaghluf and T. A. Gulliver, "Energy and spectrum efficiency in predictive-cooperative cognitive radio networks," *Wireless Netw.*, vol. 27, no. 8, pp. 5297–5311, 2021.

[11] A. Iqbal, A. Nauman, R. Hussain, and M. Bilal, "Cognitive D2D communication: A comprehensive survey, research challenges, and future directions," *Internet Things*, vol. 24, 2023.

[12] C. Xu, L. Song, and Z. Han, *Resource Management for Device-to-Device Underlay Communication*. New York, NY, USA: Springer, 2014.

[13] A. M. H. Alibraheemi *et al.*, "A survey of resource management in D2D communication for B5G networks," *IEEE Access*, vol. 11, pp. 7892–7923, 2023.

[14] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2133–2168, 2018.

[15] L. Nadeem *et al.*, "Integration of D2D, network slicing, and MEC in 5G cellular networks: Survey and challenges," *IEEE Access*, vol. 9, pp. 37590–37612, 2021.

[16] L. Song, *Wireless Device-to-Device Communications and Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[17] F. Hussain, S. A. Hassan, R. Hussain, and E. Hossain, "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 2, pp. 1251–1275, 2020.

[18] R. C. Qiu, Z. Hu, H. Li, and M. C. Wicks, *Cognitive Radio Communication and Networking: Principles and Practice*. Hoboken, NJ, USA: Wiley, 2012.

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

[20] O. H. Toma, M. López-Benítez, D. K. Patel, and K. Umebayashi, "Estimation of primary channel activity statistics in cognitive radio based on imperfect spectrum sensing," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2016–2031, 2020.

[21] H. Wu, F. Yao, Y. Chen, Y. Liu, and T. Liang, "Cluster-based energy efficient collaborative spectrum sensing for cognitive sensor networks," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2722–2725, 2017.

[22] X. Liu, K. Zheng, K. Chi, and Y.-H. Zhu, "Cooperative spectrum sensing optimization in energy-harvesting cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7663–7676, 2020.

[23] P. Thakur, A. Kumar, S. Pandit, G. Singh, and S. N. Satashia, "Performance analysis of high-traffic cognitive radio communication systems using hybrid spectrum access, prediction, and monitoring techniques," *Wireless Netw.*, vol. 24, no. 6, pp. 2005–2015, 2018.

[24] Y. Zhang, J. Hou, V. Towhidlou, and M. R. Shikh-Bahaei, "A neural network prediction-based adaptive mode selection scheme in full-duplex cognitive networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 540–553, 2019.

[25] S. D. Barnes, B. T. Maharaj, and A. S. Alfa, "Cooperative prediction for cognitive radio networks," *Wireless Pers. Commun.*, vol. 89, no. 4, pp. 1177–1202, 2016.

[26] N. Shaghluf and T. A. Gulliver, "Spectrum and energy efficiency of cooperative spectrum prediction in cognitive radio networks," *Wireless Netw.*, vol. 25, no. 6, pp. 3265–3274, 2019.

[27] V. D. Nguyen and O. S. Shin, "Cooperative prediction-and-sensing based spectrum sharing in cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 1, pp. 108–120, 2017.

[28] W. Zhang, C. Wang, D. Chen, and H. Xiong, "Energy–spectral efficiency tradeoff in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2208–2218, 2016.

[29] H. Shokri-Ghadikolaei *et al.*, "Green sensing and access: Energy-throughput trade-offs in cognitive networking," *IEEE Commun. Mag.*, vol. 53, no. 11, pp. 199–207, 2015.

[30] F. Haider *et al.*, "Spectral and energy efficiency analysis for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2969–2980, 2015.

[31] J. Yang and H. Zhao, "Enhanced throughput of cognitive radio networks by imperfect spectrum prediction," *IEEE Commun. Lett.*, vol. 19, no. 10, pp. 1738–1741, 2015.

[32] L. Yu *et al.*, "Spectrum availability prediction for cognitive radio communications: A DCG approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 476–485, 2020.

[33] Y. Zhao *et al.*, "Prediction-based spectrum management in cognitive radio networks," *IEEE Syst. J.*, vol. 12, no. 4, pp. 3303–3314, 2018.

[34] M. López-Benítez *et al.*, "Estimation of primary channel activity statistics in cognitive radio based on periodic spectrum sensing observations," *IEEE Trans. Wireless Commun.*, vol. 18, no. 2, pp. 983–996, 2018.

[35] A. Bhowmick *et al.*, "Throughput of an energy harvesting cognitive radio network based on prediction of primary user," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8119–8128, 2017.

[36] M. S. Kerdabadi *et al.*, "Energy consumption minimization and throughput improvement in cognitive radio networks by joint optimization of detection threshold, sensing time, and user selection," *Wireless Netw.*, vol. 25, no. 4, pp. 2065–2079, 2019.

[37] H. Hu, H. Zhang, and Y. Liang, "On the spectrum- and energy-efficiency tradeoff in cognitive radio networks," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 490–501, 2016.

[38] S. Chatterjee, S. P. Maity, and T. Acharya, "Energy-spectrum efficiency trade-off in energy harvesting cooperative cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 2, pp. 295–303, 2019.

[39] E. C. Y. Peh, Y. Liang, Y. L. Guan, and Y. Zeng, "Optimization of cooperative sensing in cognitive radio networks: A sensing-throughput tradeoff view," *IEEE Trans. Veh. Technol.*, vol. 58, no. 9, pp. 5294–5299, 2009.

[40] J. Tong, M. Jin, Q. Guo, and Y. Li, "Cooperative spectrum sensing: A blind and soft fusion detector," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2726–2737, 2018.

[41] M. Höyhtyä, S. Pollin, and A. Mammela, "Classification-based predictive channel selection for cognitive radios," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Cape Town, South Africa, 2010.

[42] G. Özcan, M. C. Gursoy, and J. Tang, "Spectral and energy efficiency in cognitive radio systems with unslotted primary users and sensing uncertainty," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4138–4151, 2017.

[43] H. Eltom, S. Kandeepan, B. Moran, and R. J. Evans, "Spectrum occupancy prediction using a hidden Markov model," in *Proc. Int. Conf. Signal Process. Commun. Syst.*, Cairns, Australia, 2015.

[44] A. Omidkar, A. Khalili, H. H. Nguyen, and H. Shafiei, “Reinforcement-learning-based resource allocation for energy-harvesting-aided D2D communications in IoT networks,” *IEEE Internet Things J.*, vol. 9, no. 17, pp. 16521-16531,2022.

[45] N. K. Jadav and S. Tanwar, “Whale optimization-based access control scheme in D2D communication underlaying cellular networks,” *IEEE Trans. Netw. Serv. Manag.,* vol. 21, no. 3, pp. 3416-3427,2024.

[46] A. H. Sakr and E. Hossain, “Cognitive and energy harvesting-based D2D communication in cellular networks: Stochastic geometry modeling and analysis,” *IEEE Trans. Commun.,* vol. 63, no. 5, pp. 1867-1880, 2015.

[47] J. Tan, Y. C. Liang, L. Zhang, and G. Feng, “Deep reinforcement learning for joint channel selection and power control in D2D networks,” *IEEE Trans. Wireless Commun*., vol. 20, no. 2, pp. 1363-1378, 2021.

[48] A. Sultana, L. Zhao, and X. Fernando, “Efficient resource allocation in device-to-device communication using cognitive radio technology,” *IEEE Trans. Veh. Technol*., vol. 66, no. 11, pp. 10024-10034, 2017.

[49] S. Guo and X. Zhao, “Deep reinforcement learning optimal transmission algorithm for cognitive Internet of Things with RF energy harvesting,” *IEEE Trans. Cognit. Commun. Netw*., vol. 8, no. 2, pp. 1216-1227, 2022.

[50] J. Si, R. Huang, Z. Li, H. Hu, Y. Jin, J. Cheng, and N. Al-Dhahir, “When spectrum sharing in cognitive networks meets deep reinforcement learning: Architecture, fundamentals and challenges,” *IEEE Netw*., vol. 37, no. 1, pp. 150-157, 2023.

[51] Y. Bokobza, R. Dabora, and K. Cohen, “Deep reinforcement learning for simultaneous sensing and channel access in cognitive networks,” *IEEE Trans. Wireless Commun*., vol. 21, no. 9, pp. 1101-1112, 2023.

[52] P. G. Ye, Y. G. Wang, and W. Tang, “S-MFRL: Spiking mean field reinforcement learning for dynamic resource allocation of D2D networks,” *IEEE Trans. Veh. Technol*., vol. 72, no. 1, pp. 1032-1047, 2022.

[53] Z. Li and C. Guo, “Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications,” *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828-1840, 2019.

[54] X. Li, Y. Zhang, H. Ding, and Y. Fang, “Intelligent spectrum sensing and access with partial observation based on hierarchical multi-agent deep reinforcement learning,” *IEEE Trans. Wireless Commun,* vol. 23, no. 4, pp. 3131-3145, 2024.

[55] S. Schneider, H. Karl, R. Khalili, and A. Hecker, “Multi-agent deep reinforcement learning for coordinated multipoint in mobile networks,” *IEEE Trans. Netw. Serv. Manag.,* vol. 21, no. 1, pp. 908-924, 2024.

[56] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y. C. Liang, and D. I. Kim, “Applications of deep reinforcement learning in communications and networking: A survey,” *IEEE Commun. Surveys Tuts.,* vol. 21, no. 4, pp. 3133-3174, 2019.

[57] J. Huang, Y. Yang, Z. Gao, D. He, and D. W. K. Ng, “Dynamic spectrum access for D2D-enabled Internet of Things: A deep reinforcement learning approach,” *IEEE Internet Things J*., vol. 9, no. 18, pp. 17793-17807, 2022.

[58] H. Xiang, Y. Yang, G. He, J. Huang, and D. He, “Multi-agent deep reinforcement learning-based power control and resource allocation for D2D communications,” *IEEE Wireless Commun. Lett*., vol. 11, no. 8, pp. 1659-1663, 2022.

[59] C. Kai, X. Meng, L. Mei, and W. Huang, “Multi-agent reinforcement learning based joint uplink–downlink subcarrier assignment and power allocation for D2D underlay networks,” *Wireless Netw.,* vol. 29, no. 2, pp. 891-907, 2023.

[60] Y. Zhi, J. Tian, X. Deng, J. Qiao, and D. Lu, "Deep reinforcement learning-based resource allocation for D2D communications in heterogeneous cellular networks," *Digit. Commun. Netw*., vol. 8, no. 5, pp. 834-842, 2022.

[61] V. Vishnoi, I. Budhiraja, S. Gupta, and N. Kumar, "A deep reinforcement learning scheme for sum rate and fairness maximization among D2D pairs underlaying cellular network with NOMA," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13506-13522, 2023.

[62] J. Huang, Y. Yang, G. He, Y. Xiao, and J. Liu, "Deep reinforcement learning-based dynamic spectrum access for D2D communication underlay cellular networks," *IEEE Commun. Lett.,* vol. 25, no. 8, pp. 2614-2618, 2021.

[63] A. Kaur, K. Kumar, A. Prakash, and R. Tripathi, "Imperfect CSI-based resource management in cognitive IoT networks: A deep recurrent reinforcement learning framework," *IEEE Trans. Cognit. Commun. Netw*., vol. 9, no. 5, pp. 1271-1281, 2023.

[64] T. Li, K. Zhu, N. C. Luong, D. Niyato, Q. Wu, Y. Zhang, and B. Chen, "Applications of multi-agent reinforcement learning in future Internet: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 1240-1279, 2022.

[65] H. Li and H. He, "Multiagent trust region policy optimization," *IEEE Trans. Neural Netw. Learn. Syst.,* vol. 35, no. 9, pp. 12873-12887, 2024.

[66] Y. Yan, B. Zhang, C. Li, and C. Su, "Cooperative caching and fetching in D2D communications: A fully decentralized multi-agent reinforcement learning approach," *IEEE Trans. Veh. Technol.,* vol. 69, no. 12, pp. 16095-16109, 2020.

[67] I. Budhiraja, N. Kumar, and S. Tyagi, "Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay D2D communication," *IEEE Internet Things J.,* vol. 8, no. 5, pp. 3143-3156, 2020.

[68] S. Jayakumar and S. Nandakumar, "Reinforcement learning-based distributed resource allocation technique in device-to-device (D2D) communication," *Wireless Netw*., vol. 29, no. 4, pp. 1843-1858, 2023.

[69] Y. Guan, S. Zou, H. Peng, W. Ni, Y. Sun, and H. Gao, "Cooperative UAV trajectory design for disaster area emergency communications: A multi-agent PPO method," *IEEE Internet Things J.*, vol. 11, no. 5, pp. 8848-8859, 2024.

[70] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, “Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications,” *IEEE Trans. Veh. Technol.,* vol. 71, no. 8, pp. 8810-8824, 2022.

[71] S. Zhou, L. Feng, M. Mei, and M. Yao, “Dynamic resource management for federated edge learning with imperfect CSI: A deep reinforcement learning approach,” *IEEE Internet Things J*., vol. 11, no. 18, pp. 30400-30412, 2024.

[72] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, “Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach,” *IEEE Access*, vol. 6, pp. 25463-25473, 2018.

[73] H. H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, “Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach,” *IEEE Internet Things J*., vol. 6, no. 2, pp. 1938-1948, 2019.

[74] X. Tan et al., “Cooperative multi-agent reinforcement-learning-based distributed dynamic spectrum access in cognitive radio networks,” *IEEE Internet Things J*., vol. 9, no. 19, pp. 19477-19488, 2022.

[75] A. Feriani and E. Hossain, “Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial,” IEEE Commun. Surveys Tuts., vol. 23, no. 2, pp. 1226-1252, 2021.

[76] R. Gupta and S. Tanwar, “A zero-sum game-based secure and interference mitigation scheme for socially aware D2D communication with imperfect CSI,” IEEE Trans. Netw. Serv. Manag., vol. 19, no. 3, pp. 3478-3486, 2022.

[77] S. Mallick, R. Devarajan, R. A. Loodaricheh, and V. K. Bhargava, “Robust resource optimization for cooperative cognitive radio networks with imperfect CSI,” IEEE Trans. Wireless Commun., vol. 14, no. 2, pp. 907-920, 2015.

[78] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv preprint arXiv:1707.06347, 2017.

[79] K. Agrawal, S. Parihar, P. Chakraborty, and S. Prakriya, "Performance of NOMA-Enabled Cooperative FD D2D Communication in Cellular Underlay," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 4, pp. 1730-1742, 2022.

[80] G. Wu, G. Chen, and G. Chen, "Energy-Efficient Utility Function-Based Channel Resource Allocation and Power Control for D2D Clusters with NOMA Enablement in Cellular Networks," *IEEE Access*, vol. 11, pp. 45001-45010, 2023.

[81] W. Chen, Y. Liu, H. Jafarkhani, Y. C. Eldar, P. Zhu, and K. B. Letaief, "Signal Processing and Learning for Next Generation Multiple Access in 6G," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 7, pp. 1146-1177, 2024.

[82] S. Zhang, J. Liu, H. Guo, M. Qi, and N. Kato, "Envisioning Device-to-Device Communications in 6G," *IEEE Network*, vol. 34, no. 3, pp. 86-91, 2020.

[83] Y. Cheng, C. Liang, Q. Chen, and F. R. Yu, "Energy-Efficient D2D-Assisted Computation Offloading in NOMA-Enabled Cognitive Networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 12, pp. 13441-13446, 2021.

[84] Y. Liu et al., "Evolution of NOMA Toward Next Generation Multiple Access (NGMA) for 6G," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 4, pp. 1037-1071, 2022.

[85] H. M. F. Noman, K. Dimyati, K. A. Noordin, E. Hanafi, and A. Abdrabou, "FeDRL-D2D: Federated Deep Reinforcement Learning-Empowered Resource Allocation Scheme for Energy Efficiency Maximization in D2D-Assisted 6G Networks," *IEEE Access*, vol. 12, pp. 109775-109792, 2024.

[86] K. Z. Shen, D. K. C. So, J. Tang, and Z. Ding, "Power Allocation for NOMA With Cache-Aided D2D Communication," *IEEE Transactions on Wireless Communications*, vol. 23, no. 1, pp. 529-542, 2024.

[87] M. K. Awad et al., "A Matching-Theoretic Approach to Resource Allocation in D2D-Enabled Downlink NOMA Cellular Networks," *Physical Communication*, vol. 54, p. 101837, 2022.

[88] Y. Pan, C. Pan, Z. Yang, and M. Chen, "Resource Allocation for D2D Communications Underlaying a NOMA-Based Cellular Network," *IEEE Wireless Communications Letters*, vol. 7, no. 1, pp. 130-133, 2018.

[89] J. Zhao, Y. Liu, K. K. Chai, Y. Chen, and M. Elkashlan, "Joint Subchannel and Power Allocation for NOMA Enhanced D2D Communications," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5081-5094, 2017.

[90] M. Sun, X. Xu, X. Tao, P. Zhang, and V. C. M. Leung, "NOMA-Based D2D-Enabled Traffic Offloading for 5G and Beyond Networks Employing Licensed and Unlicensed Access," *IEEE Transactions on Wireless Communications*, vol. 19, no. 6, pp. 4109-4124, 2020.

[91] S. Yu, W. U. Khan, X. Zhang, and J. Liu, "Optimal Power Allocation for NOMA-Enabled D2D Communication with Imperfect SIC Decoding," *Physical Communication*, vol. 46, p. 101296, 2021.

[92] M. Le, Q.-V. Pham, H.-C. Kim, and W.-J. Hwang, "Enhanced Resource Allocation in D2D Communications With NOMA and Unlicensed Spectrum," *IEEE Systems Journal*, vol. 16, no. 2, pp. 2856-2866, 2022.

[93] J. Jose et al., "Performance Analysis and Learning-Assisted Power Control for NOMA Enabled D2D-Cellular Network," *IEEE Systems Journal*, vol. 18, no. 1, pp. 278-281, 2024.

[94] I. Budhiraja et al., "An Energy Efficient Scheme for WPCN-NOMA Based Device-to-Device Communication," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 11935-11948, 2021.

[95] J. Zhao et al., "NOMA-Based D2D Communications: Towards 5G," *Proc. IEEE Global Communications Conference (GLOBECOM*), Washington, DC, USA, 2016, pp. 1-6.

[96] Y. Gao et al., "A Potential Game Approach for D2D Pairing and Channel Selection for NOMA-Enabled D2D Cellular Networks," *IEEE Wireless Communications Letters*, vol. 13, no. 9, pp. 2522-2526, 2024.

[97] R. Li et al., "Resource Allocation for Uplink NOMA-Based D2D Communication in Energy Harvesting Scenario: A Two-Stage Game Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 976-990, 2022.

[98] Z. Ding, O. A. Dobre, P. Fan, and H. V. Poor, "A New Design of CR-NOMA and Its Application to AoI Reduction," *IEEE Communications Letters*, vol. 27, no. 9, pp. 2461-2465, 2023.

[99] M. Waqas et al., "Resource Optimization of D2D-Assisted CR Network with NOMA for 5G and Beyond Systems," *IEEE Internet of Things Journal*, vol. 9, no. 21, pp. 21232-21245, 2022.

[100] J. Xu, K. Ota, and M. Dong, "AI-Centric D2D in 6G Networks," *IEEE Networking Letters*, vol. 6, no. 4, pp. 257-261, 2024.

[101] X. Wang et al., "Energy Efficiency Resource Management for D2D-NOMA Enabled Network: A Dinkelbach Combined Twin Delayed Deterministic Policy Gradient Approach," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 9, pp. 11756-11771, 2023.

[102] Y. J. Jeong, S. Yu, and J. W. Lee, "DRL-Based Resource Allocation for NOMA-Enabled D2D Communications Underlay Cellular Networks," *IEEE Access*, vol. 11, pp. 140270-140286, 2023.

[103] M. A. A. Khan et al., "Sum Throughput Maximization Scheme for NOMA-Enabled D2D Groups Using Deep Reinforcement Learning in 5G and Beyond Networks," *IEEE Sensors Journal*, vol. 23, no. 13, pp. 15046-15057, 2023.

[104] L. Guo et al., "Resource Allocation for Multiple RISs Assisted NOMA Empowered D2D Communication: A MAMP-DQN Approach," *Ad Hoc Networks*, vol. 146, p. 103163, 2023.

[105] S. Huang, Q. Gallouédec, F. Felten, A. Raffin, R. F. J. Dossa, Y. Zhao, R. Sullivan *et al.*, "Open RL Benchmark: Comprehensive Tracked Experiments for Reinforcement Learning," *arXiv preprint arXiv:2402.03046*, 2024.

[106] X. Huang, T. Han and N. Ansari, "On Green-Energy-Powered Cognitive Radio Networks," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 827-842, 2015.

[107] R. Li, P. Hong, K. Xue, M. Zhang, and T. Yang, "Resource Allocation for Uplink NOMA-Based D2D Communication in Energy Harvesting Scenario: A Two-Stage Game Approach," *IEEE Transactions on Wireless Communications*, vol. 21, no. 2, pp. 976-990, 2022.

[108] A. Omidkar, A. Khalili, H. H. Nguyen, and H. Shafiei, "Reinforcement-Learning-Based Resource Allocation for Energy-Harvesting-Aided D2D Communications in IoT Networks," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 16521-16531, 2022.

[109] A. A. Al-Habob, H. Tabassum, and O. Waqar, "Non-Orthogonal Age-Optimal Information Dissemination in Vehicular Networks: A Meta Multi-Objective Reinforcement Learning Approach," *IEEE Transactions on Mobile Computing*, vol. 23, no. 10, pp. 9789-9803, Oct. 2024.

[110] D. Sarkar, Yogita, S. S. Yadav, V. Pal, N. Kumar, and S. K. Patra, "A Comprehensive Survey on IRS-Assisted NOMA-Based 6G Wireless Network: Design Perspectives, Challenges and Future Directions," *IEEE Transactions on Network and Service Management*, vol. 21, no. 2, pp. 2539-2562, 2024.

[111] B. Q. Zhao, H. M. Wang, and H. Deng, "Does D2D Communication Always Benefit Physical-Layer Security?, " *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 224-240, 2023.

[112] J. Chen, Y. Deng, J. Jia, M. Dohler and A. Nallanathan, "Cross-Layer QoE Optimization for D2D Communication in CR-Enabled Heterogeneous Cellular Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 719-734, 2018.

[113] H. Yang, W. D. Zhong, C. Chen, A. Alphones, and X. Xie, "Deep-Reinforcement-Learning-Based Energy-Efficient Resource Management for Social and Cognitive Internet of Things," *IEEE Internet of Things Journal*, vol. 7, no. 6, pp. 5677-5689, 2020.

[114] H. Pennanen, T. Hänninen, O. Tervo, A. Tölli, and M. Latva-Aho, "6G: The Intelligent Network of Everything," *IEEE Access*, vol. 13, pp. 1319-1421, 2024