Edge Intelligent Multi-User Uplink Radio Access Methods for Beyond 6G Networks

Xueyu Wu

Doctor of Philosophy

University of York
School of Physics, Engineering and Technology

June 2025

Abstract

Uplink radio access is a critical component of future 6G wireless networks, enabling simultaneous data transmissions from multiple users to access points (APs). Due to dynamic wireless environments, users must adaptively adjust modulation and coding schemes (MCS) and random access strategies. The proliferation of machine-type communication devices, characterized by sporadic traffic and limited resources, underscores the need for distributed, low-complexity algorithms to optimize uplink access.

This thesis develops distributed learning-driven solutions for these challenges in 6G uplink access, offering three main contributions. First, a federated learning method is proposed for adaptive orthogonal frequency division multiplexing with index modulation (OFDM-IM), utilizing k-means clustering. This approach aggregates the learning outcomes of local devices into a global model, reducing local training requirements and enhancing throughput while introducing reduced federation overhead. The effectiveness of this federated learning assisted adaptive modulation inherently relies on robust and efficient multiple access methods to manage the anticipated massive connectivity in future networks. Second, a learning-based two-step non-orthogonal random access (NORA) strategy is developed for massive connectivity, allowing users to independently select transmission slots and power levels without channel state information (CSI). The base station (BS) applies successive interference cancellation (SIC) to decode overlapping transmissions. This joint slot-power selection is modeled as a Markov decision process (MDP) and solved with tailored multi-state and confidence-aided Q-learning algorithms, significantly improving throughput and fairness, particularly under high congestion. Finally, the thesis presents a decentralized deep reinforcement learning-based NORA framework for multi-AP machine-type communications. Users autonomously select APs, transmission slots, and power levels without CSI, while APs decode packets using SIC. Custom deep Q-network algorithms address the AP-slot-power selection, showing substantial improve-

ments in throughput, fairness, and scalability under diverse, high-traffic scenarios, emphasizing the framework's potential for intelligent random access in 6G ecosystems.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Youngwook Ko and Prof. Andy Tyrrell. Academic research is often difficult and challenging, filled with false starts and mistakes that can easily derail progress. Their expertise, experience, and patient guidance helped me navigate these obstacles. They taught me how to conduct rigorous and impactful research, and I would not have come this far without their supervision.

I would also like to thank my colleagues and friends at the University of York—especially Tianyuan, Junbo, and Jie—for their generous support in both my studies and daily life. I am also grateful to Zhihao, my best gym buddy, for keeping me motivated and balanced.

To my partner, Yu Wang, who is with me all the time. You are the person who gives me the confidence and courage to face every challenge. I cannot imagine this journey without you.

Finally, and most importantly, I would like to express my heartfelt gratitude to my loving parents, Huiping Lv and Yanyun Wu, and to my stepfather, Jinsheng Zhai—thank you for your unconditional support. To my late grandmother, Shuzhen Li, I deeply appreciate your lifelong care and selfless love.

Xueyu Wu,
York,
June 2025.

Declaration and Related publications

I hereby declare that this thesis represents my own original work, and I am the sole author. This work has not been submitted previously for any award at his or any other University. All sources are properly acknowledged in the References section.

Chapters 3 is based on the following journal paper.

> X. Wu, A. M. Tyrrell and Y. Ko, "Federated K-Means Clustering for Adaptive OFDM-IM," in *IEEE Communications Letters*, vol. 27, no. 10, pp. 2648-2651, Oct. 2023, doi: 10.1109/LCOMM.2023.3308334.

Chapter 4 is based on the following journal paper and conference paper.

> X. Wu, Y. Ko and A. M. Tyrrell, "Distributed Multi-Agent Reinforcement Learning for Heterogeneous NOMA-ALOHA Systems," in IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 3, pp. 1902-1912, June 2025, doi: 10.1109/TCCN.2024.3474709.

> X. Wu, Y. Ko and A. M. Tyrrell, "Decentralized Multi-State Q-Learning for NOMA-ALOHA Systems," *2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Valencia, Spain, 2024, pp. 1-6, doi: 10.1109/PIMRC59610.2024.10817213.

Chapter 5 is based on the following paper to be submitted to IEEE Transactions on Cognitive Communications and Networking.

> X. Wu, Y. Ko and A. M. Tyrrell, "Distributed Deep Q Network based Multi-Agent Multi-Point NORA Networks,".

Contents

List of Tables

List of Abbreviations

2G/3G/4G/5G/6G 2nd/3rd/4th/5th/6th Generation (mobile networks).

AAOI Average Age of Information.

AI Artificial Intelligence.

AMC Automatic Modulation Classification.

AP Access Point.

APM Amplitude–Phase Modulation.

APs Access Points.

AR/VR Augmented Reality / Virtual Reality.

AWGN Additive White Gaussian Noise.

BER Bit Error Rate.

BS Base Station.

CBRA Contention-Based Random Access.

CNN Convolutional Neural Network.

CP Cyclic Prefix.

CSI Channel State Information.

DDQN Double Deep Q-Network.

DNN Deep Neural Network.

DQN Deep Q-Network.

DRL Deep Reinforcement Learning.

EI Edge Intelligence.

FL Federated Learning.

GB-RA Grant-Based Random Access.

GF-NOMA Grant-Free Non-Orthogonal Multiple Access.

GF-RA Grant-Free Random Access.

gNB Next-Generation NodeB (5G base station).

IoT Internet of Things.

LSTM Long Short-Term Memory.

MAB Multi-Armed Bandit.

MAP Mode Activation Pattern.

MARL Multi-Agent Reinforcement Learning.

MCS Modulation and Coding Scheme.

MDP Markov Decision Process.

MEC Multi-Access Edge Computing.

MIMO Multiple-Input Multiple-Output.

ML Machine Learning.

Msg1/Msg2/Msg3/Msg4 Conventional 4-Step RA messages.

MsgA 2-Step RA combined preamble+payload uplink message.

MsgB 2-Step RA contention-resolution downlink message.

MTC Machine-Type Communication.

NGMA Next-Generation Multiple Access.

NOMA Non-Orthogonal Multiple Access.

NORA Non-Orthogonal Random Access.

OFDM Orthogonal Frequency Division Multiplexing.

OFDMA Orthogonal Frequency Division Multiple Access.

OFDM-IM OFDM with Index Modulation.

OMA Orthogonal Multiple Access.

PDCCH Physical Downlink Control Channel.

PDSCH Physical Downlink Shared Channel.

PRACH Physical Random Access Channel.

PUSCH Physical Uplink Shared Channel.

QoS Quality of Service.

RA Random Access.

RAN Radio Access Network.

RAR Random Access Response.

RB Resource Block.

RIS Reconfigurable Intelligent Surface.

RL Reinforcement Learning.

SA Slotted ALOHA.

SIC Successive Interference Cancellation.

SINR Signal-to-Interference-plus-Noise Ratio.

SM Spatial Modulation.

SNR Signal-to-Noise Ratio.

TDD Time Division Duplexing.

UAV Unmanned Aerial Vehicle.

UE User Equipment.

UL/DL Uplink / Downlink.

URLLC Ultra-Reliable Low-Latency Communication.

List of Mathematical Symbols

α Learning rate.

B System bandwidth.

$\binom{N}{k}$ Binomial coefficient (combinations).

$\mathcal{D}$ Replay buffer / memory.

δ Temporal-difference (TD) error.

d_n, $d_{n,m}$ Distance of user n to BS / AP m.

ε Exploration probability in ε-greedy.

Flag_n, FB_m Uplink feedback indicator (BS / AP).

$\lfloor \cdot \rfloor$ Floor operator.

γ Discount factor, $0 \leq \gamma < 1$.

G_t Return: $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$.

h_k Channel gain of user k.

$\{i_1, \ldots, i_k\}$ Indices of active subcarriers.

$\mathbf{I}$ Identity matrix.

k Number of active subcarriers in OFDM-IM.

K Number of time slots (in slotted RA) or users per scenario.

L Number of discrete power levels (NORA).

$\log_2(\cdot)$ Base-2 logarithm.

$L(W)$ Loss function for function approximation.

M APM constellation order (e.g., QAM size).

M Number of access points (multi-AP scenarios).

μ Reward for fail transmission.

N_0 Noise power spectral density.

n_k Number of training samples at client k.

$\|\cdot\|$ Vector / matrix norm.

N_t Number of transmit antennas (spatial modulation).

p Total input bits per OFDM-IM subblock.

p_1, p_2 Index bits and symbol bits, with $p = p_1 + p_2$.

P_k Transmit power of user k.

$\pi(a \mid s)$ Policy: probability of taking action a in state s.

$P^a_{ss'}$ Transition probability from s to s' under action a.

P_t Total (max) transmit power / constraint.

P_{tol} Minimum power difference required for SIC.

$Q_\pi(s, a)$ Action-value under policy π.

R_k Throughput / achievable rate of user k.

R^a_s Expected immediate reward at state s under action a.

$\mathbf{s} = [s_1, \ldots, s_k]$ APM symbols mapped to active subcarriers.

$\mathcal{S}, \mathcal{A}$ State space and action space.

$\mathbb{R}, \mathbb{C}$ Real and complex number sets.

S_t, A_t, R_t State, action, and reward at time t.

θ Trainable parameter vector (e.g., policy / network).

θ_{Fair} Fairness-related control parameter.

$(\cdot)^\top$ Transpose.

$V_\pi(s)$ State-value under policy π.

w_t^k Local model parameters of client k at iteration t.

w_t Global model parameter vector at iteration t.

$\mathbf{x} \in \mathbb{C}^N$ Frequency-domain transmit vector.

1 Introduction

Since Shannon introduced information theory in 1948, digital communication technologies have advanced rapidly, revolutionizing how information is transmitted and processed. The introduction of packet switching in the 1960s laid the foundation for modern data networks, culminating in the creation of ARPANET in 1969, the precursor to the internet. During the 1970s, the development of the Transmission Control Protocol/Internet Protocol (TCP/IP) enabled seamless communication across different networks. Notably, the first transatlantic email was transmitted in 1973 from the Richmond Building at the University of Sussex, marking a significant milestone in global digital communication.

The 1980s and 1990s saw a transition from analog to digital systems, with the advent of 2G mobile networks, the World Wide Web (WWW), and the rapid expansion of broadband internet. The turn of the century introduced 3G networks, significantly enhancing mobile data transmission, while the proliferation of Wi-Fi and smartphones transformed personal and professional communication. The 2010s brought the widespread adoption of 4G LTE, enabling high-speed mobile internet, and saw the emergence of the Internet of Things (IoT), integrating billions of connected devices.

Today, the rollout of 5G networks is enabling ultra-low latency and high-

bandwidth applications, such as autonomous vehicles, edge computing, and AI-driven communication systems. Looking ahead, the development of 6G networks, quantum communication, and intelligent network optimization is expected to drive the next generation of digital connectivity. According to IHS Markit, the number of smart devices is projected to reach 125 billion by 2030 [1], reflecting the continued exponential growth of digital communication technologies.

However, upcoming massive connectivity and edge intelligence will bring challenges to uplink transmissions in 6G radio access network (RAN).

1.1 Motivation

The advent of 6G networks is poised to transform wireless communication by delivering ultra-reliable, low-latency, and high-capacity connectivity. A key component of this evolution is Edge Intelligence (EI), which integrates artificial intelligence (AI) and machine learning (ML) at the network edge to facilitate real-time decision-making, intelligent resource management, and adaptive modulation strategies [2]. These advancements are particularly critical for multi-user uplink transmission scenarios.

1.1.1 The Role of Multi-User Uplink Transmission in Edge Intelligence Applications

Emerging edge intelligence applications are driving innovation across various industries. These applications demand sophisticated multi-user uplink transmission technologies to facilitate seamless data exchange and real-time

decision-making. Key domains include:

- **Autonomous Transportation**: Edge intelligence enables self-driving vehicles to process sensor data in real time, ensuring safe and efficient navigation in dynamic environments [3].
- **Industrial Automation**: AI-driven predictive maintenance and robotic process automation enhance operational efficiency, minimize downtime, and optimize production workflows [4].
- **Smart Healthcare**: Edge intelligence facilitates remote patient monitoring, real-time medical imaging analysis, and AI-assisted diagnostics, significantly improving healthcare accessibility and quality [5].
- **Augmented Reality (AR) and Virtual Reality (VR)**: These applications necessitate ultra-low latency and high-bandwidth uplink transmission to provide immersive experiences for gaming, remote collaboration, and training simulations [6].
- **Urban Infrastructure**: Smart cities leverage edge intelligence for intelligent traffic management, energy-efficient smart grids, and real-time environmental monitoring, fostering sustainable urban development [7].

To support these applications, multi-user uplink transmission must address the growing demand for efficient data aggregation, low-latency response times, and high-throughput communication. AI-driven approaches, including Federated Learning (FL), Reinforcement Learning (RL), and Learning-Driven Adaptive Modulation (LDAM), provide promising solutions to optimize spectrum utilization, reduce transmission delays, and enhance network resilience. These advancements ensure that critical edge intelligence

applications function optimally, even in complex and resource-constrained environments.

1.1.2 Challenges in Multi-User Uplink Transmission for 6G

Despite its potential, multi-user uplink transmission in 6G networks faces several challenges:

- **Scalability and Dynamic Resource Allocation**: The exponential growth of edge devices complicates spectrum access and transmission scheduling, necessitating adaptive, decentralized resource management techniques [8].

- **Latency Constraints**: Many edge intelligence applications demand near-instantaneous data processing, a requirement that current wireless architectures may struggle to meet, particularly in densely populated network environments [9].

- **Interference and Reliability Issues**: The dense deployment of intelligent devices sharing the spectrum introduces significant interference, necessitating advanced mitigation strategies to maintain robust and reliable uplink transmission [10].

- **Security and Privacy Concerns**: Federated Learning and distributed AI methodologies introduce new vulnerabilities, such as model poisoning and data privacy breaches, necessitating robust encryption and secure transmission protocols [11].

- **Energy Efficiency and Computational Constraints**: Given the limited power resources of edge devices, implementing energy-efficient communication protocols and lightweight AI models is essential to ensuring long-term sustainability and operational continuity [12].

1.1.3 Leveraging Edge Intelligence to Enhance 6G Uplink Transmission

To address these challenges, Edge Intelligence offers a transformative framework for optimizing multi-user uplink transmission. Key benefits include:

- **AI-Driven Adaptive Resource Allocation**: Machine learning models at the network edge dynamically optimize spectrum allocation, power control, and transmission scheduling [13].
- **Federated Learning for Uplink Optimization**: Federated Learning facilitates distributed learning across multiple devices while preserving data privacy, improving uplink modulation schemes and access control strategies [14].
- **Reinforcement Learning-Based Uplink Access**: AI-driven decision-making mechanisms enhance contention-based and grant-free random access protocols, reducing collisions and increasing system throughput [15].
- **Energy-Efficient AI Strategies**: The integration of lightweight AI models tailored for edge computing minimizes computational overhead and energy consumption, ensuring optimal performance for resource-constrained devices [16].

The convergence of edge intelligence and advanced uplink transmission technologies represents a fundamental shift from centralized architectures to autonomous, distributed frameworks. By integrating AI-driven methodologies with next-generation wireless technologies, 6G uplink transmission can achieve enhanced resilience, adaptability, and efficiency. This fusion of AI-powered edge computing and multi-user uplink transmission will play a critical role in developing self-optimizing and self-healing wireless networks. These advancements will underpin a new era of high-performance wireless communication, supporting a broad range of applications, including smart cities, industrial IoT, and mission-critical autonomous systems.

Hypothesis

Taking advantage of emerging edge intelligence, the critical challenges in uplink radio access can be addressed by deploying adaptive and learning-driven algorithms at the network edge.

Novel Contributions

This thesis advances the field of intelligent wireless communication by developing learning-driven strategies for adaptive modulation and scalable random access in 6G networks. Key novel contributions include:

- The design of a **federated k-means clustering framework** for adaptive OFDM-IM systems, enabling distributed learning across devices with reduced training cost and improved throughput.
- The proposal of **model-free Q-learning algorithms** for a 2-step NORA system under CSI uncertainty, allowing users to make decentral-

ized decisions on slot and power level selection, significantly enhancing throughput and fairness in congested environments.

- The development of a **deep reinforcement learning-based NORA framework** for multiple access point (AP) scenarios, where users autonomously determine optimal AP, slot, and power configurations. This approach demonstrates strong scalability and robustness in high-density 6G networks.

Collectively, these contributions introduce scalable and decentralized machine learning techniques tailored for next-generation wireless systems, addressing the challenges of massive connectivity, limited channel information, and heterogeneous user environments.

1.2 Structure and Contributions of Thesis

The structure and contributions of this thesis are summarized as follows, including six chapters.

- Chapter 2 introduces the theory and recent literature of machine learning, federated learning, reinforcement learning, and how these learning algorithms can assist adaptive modulation and random access.

- Chapter 3 demonstrates a new federated learning strategy for k-means clustering assisted adaptive orthogonal frequency devision multiplexing with index modulation (OFDM-IM) system, which develops a global model using local learning outcomes aggregated from distributed devices. The proposed strategy aims to efficiently leverage the computing

power of all the devices in a distributed system. Simulation results show that the proposed strategy reduces the training steps required at each device and simultaneously improve the throughput, with a federation cost for model aggregation and broadcast.

- Chapter 4 presents a learning-aided non-orthogonal random access (NORA) system where 2-step random access is adopted. Specifically, each user independently selects a slot and a power level for uplink packet transmission without any information about other users' selection and channel state information (CSI); and the base station (BS) performs successive interference cancellation (SIC) to decode packets from multiple users with the use of power differences on the same slot. To design a model-free multiple access under growing complexity and CSI uncertainty, the joint slot and power level selecting problem is modelled as a Markov decision process (MDP) where actions are slot-power pairs. Multi-state Q-Learning algorithms and a confidence-aided Q-Learning method are tailored for the NORA system to solve the MDP under heterogeneous environments. Simulation results show that the three proposed algorithms help the distributed users to find their strategies for slot and power level selections, improving system throughput and fairness simultaneously. The proposed algorithms are shown to make superior performance compared to the benchmarks in high congestion traffics scenarios. This is crucial for achieving massive connectivity in 6G ecosystems, which requires intelligent random access designs to accommodate the growing number of machine type users in diverse conditions.

- Chapter 5 investigates a learning-driven non-orthogonal random access (NORA) framework for distributed agents in multiple access point (AP) environments. In a decentralized manner, users autonomously select an AP, transmission slot, and power level for uplink communication without prior knowledge of other users' decisions and channel state information (CSI). The APs employ successive interference cancellation (SIC) to decode multiple data packets on the same slot by leveraging power differences. To address the challenges of designing robust multiple access mechanisms under CSI uncertainty and growing system complexity, the problem of selecting APs, slots, and power levels is formulated as a Markov decision process (MDP), where each action corresponds to a specific AP-slot-power combination. Tailored deep Q network algorithms are developed for each agent to train its own random access strategy in heterogeneous networking scenarios. Simulation results demonstrate that the proposed algorithms enable distributed users to effectively determine their access strategies, enhancing system throughput and fairness. Furthermore, leveraging the presence of multiple APs and decentralized decision-making at agents, the proposed approach provides scalable solutions that perform well in high-traffic scenarios. These findings highlight the potential of intelligent random access schemes to achieve large connectivity in 6G networks, accommodating diverse operational conditions and a growing number of machine type devices.

- Chapter 6 concludes the research findings and points out future directions.

This thesis is structured to highlight the close interplay between adaptive modulation and efficient random access mechanisms in addressing the anticipated challenges of 6G uplink transmissions. Specifically, the federated k-means clustering adaptive OFDM-IM framework developed in Chapter 3 underscores the critical requirement for robust multiple access methods, motivating the subsequent exploration of advanced Non-Orthogonal Random Access (NORA) solutions in Chapters 4 and 5.

2

Background

This chapter provides an essential foundation for understanding the methods and technologies explored in this thesis, particularly focusing on edge intelligent solutions for multi-user uplink radio access in future beyond-6G networks. It begins by introducing federated learning, emphasizing its capacity for distributed intelligence and highlighting the importance of managing heterogeneity among connected devices. Following this, the chapter discusses reinforcement learning, delineating both value-based and policy-based methods, and their relevance to adaptive decision-making in dynamic wireless environments. Subsequently, the principles and advancements of index modulation are reviewed, showcasing its potential for efficient spectrum utilization. Finally, the chapter outlines non-orthogonal random access techniques, crucial for managing massive connectivity and interference mitigation in densely populated network environments. Collectively, these foundational concepts set the stage for the innovative learning-driven strategies developed in subsequent chapters.

2.1 Federated Learning

Benefiting from the increasing number of connected devices and the enhanced computational resources within each device, Multi-Access Edge Computing (MEC) enables decentralization of intelligence from the cloud to the edge. This environment provides ideal conditions for applying federated learning (FL).

Federated learning involves two primary entities: the FL server and the clients. The clients store their own datasets, which remain inaccessible to other clients. Initially, the FL server creates a global model and distributes it to all clients. Each client then updates its local model using its own data. After performing several local updates, selected clients upload their updated models to the FL server. Subsequently, the FL server aggregates these local models to update the global model, which is then redistributed to all clients, marking the beginning of a new iteration. Fig.2.1 illustrates the system architecture of federated learning.

Compared to centralized learning strategies, FL shares only the model updates rather than the entire dataset, significantly reducing spectral and energy resource consumption. This advantage is particularly beneficial for real-time applications, such as learning-driven modulation in wireless communications. There are three types of FL: Horizontal FL, Vertical FL, and Federated Transfer Learning (FTL) [17]. In Horizontal FL, each client has datasets with identical feature spaces but different sample spaces. Conversely, Vertical FL involves clients with identical sample spaces but different feature spaces. FTL applies to scenarios where both feature and sample spaces differ.

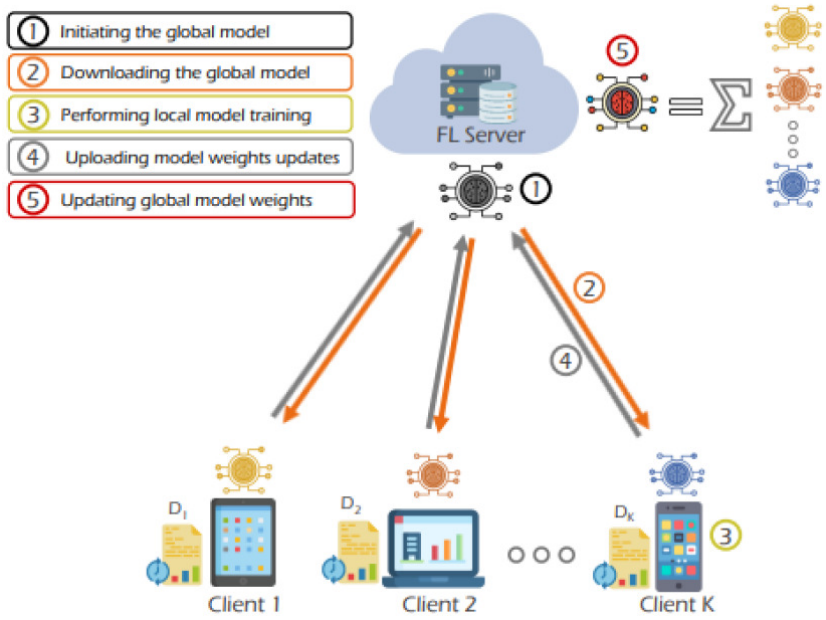


Figure 2.1: Typical architecture of a Federated Learning (FL) system. The process begins with the FL server initializing a global model, which is then distributed to multiple clients. Each client performs local model training on its private data, and the resulting model updates are uploaded back to the server. The server aggregates these updates to refine the global model. Image from the paper [17].

The uplink speed is typically much lower than the downlink speed, making communication costs during aggregation a significant concern in FL [18]. Additionally, the channel conditions among clients can vary greatly. Data quality, which significantly affects convergence rates, also differs across clients. Due to client heterogeneity and data heterogeneity, selecting clients with high-quality local data, favorable channel conditions, and sufficient computational resources for each learning iteration becomes critical. This selection strategy helps reduce communication costs and improves the convergence rate of the global model. Furthermore, compressing the local updates exchanged between the FL server and clients is another method for reducing communication overhead. Additionally, data distribution across clients is typically

non-independent and identically distributed (non-IID), whereas many existing algorithms assume IID data [17]. Consequently, optimization schemes capable of ensuring convergence under non-IID data conditions need to be developed.

There are two primary aggregation schemes in FL: FedSGD and FedAvg. In FedSGD, selected clients upload randomly sampled data points from their local datasets to the FL server, where the global model parameters are updated. However, this approach results in substantial communication costs. To mitigate this issue, FedAvg was developed by [19]. In FedAvg, clients perform local model updates and subsequently upload only their local parameters to the FL server. The server then computes the global parameters by averaging these local parameters, weighted according to the size of each client's dataset, as expressed in the following equation.

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \tag{2.1}$$

where w_{t+1} and w_t^k denote the parameters of global model and k-th client's local model at time step t, respectively. n and n_k are the total number of training data samples of all the clients and the number of training data samples of k-th client, respectively.

[20] proposed a Federated Stochastic Variance Reduced Gradient (FSVRG) optimization algorithm, demonstrating strong performance for non-IID data distributions. In SVRG, an optimal parameter, w^*, is updated after a fixed number of parameter updates. The gradient of the loss function is estimated using the difference between the gradient evaluated at a randomly sampled point in the current parameter and the optimal parameter w^*. In

the FSVRG algorithm, each client performs several parameter updates based on the same w^*, after which all clients' current parameters are aggregated, and their average is adopted as the new optimal parameter w^* for the subsequent iteration. The experimental results presented in Fig.2.2 confirm that the proposed FSVRG algorithm converges more rapidly than CoCoA+ and simple distributed gradient descent, requiring only 30 iterations.

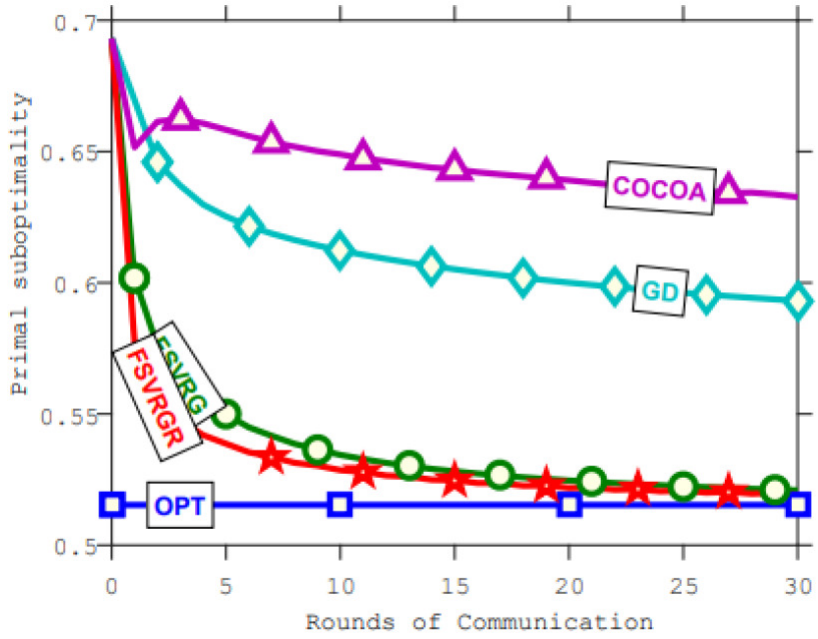


Figure 2.2: Cost function convergence curve. The results demonstrate that the proposed FSVRG algorithm converges faster than CoCoA+ and gradient descent (GD), requiring only a few rounds of communication. Image from the paper [20].

[18] address communication inefficiencies in federated learning (FL), focusing on scenarios where data is distributed across numerous mobile clients with limited network capabilities. They propose two primary strategies: structured updates, restricting model updates to low-rank or sparse formats, and sketched updates, involving compression methods such as subsampling, probabilistic quantization, and structured random rotations. Empirical re-

sults on CIFAR-10 (CNNs) and Reddit datasets (LSTMs) demonstrate that these methods substantially reduce uplink communication—by up to two orders of magnitude—while minimally affecting model accuracy, highlighting their practical applicability for efficient, large-scale, privacy-preserving federated training.

[21] proposed a federated k-means clustering algorithm, where the FL server initializes and distributes a global model. Clients update their centroids locally by assigning new samples to the nearest centroid and recalculating centroid positions based on cumulative data. Clients send their local models to the FL server after a defined number of updates. The server aggregates local models by averaging the centroids. Although this method demonstrated improved or comparable performance against baselines, it has drawbacks, including potential divergence between corresponding centroids from different clients and neglect of varying data volumes per centroid. To address these limitations, a refined federated clustering algorithm is proposed in Chapter 3.

[22] proposed a distributed Q-learning algorithm, in which each agent maintains its local Q-table and independently selects actions. The global Q-value for each state-action pair is assumed to be the highest value from local Q-tables.

[23] propose an adaptive federated learning (FL) scheme employing gradient compression to enhance communication efficiency specifically for uplink transmission using non-orthogonal multiple access (NOMA). Recognizing the limitations posed by wireless fading channels, the authors integrate adaptive gradient quantization and sparsification methods into the FL update pro-

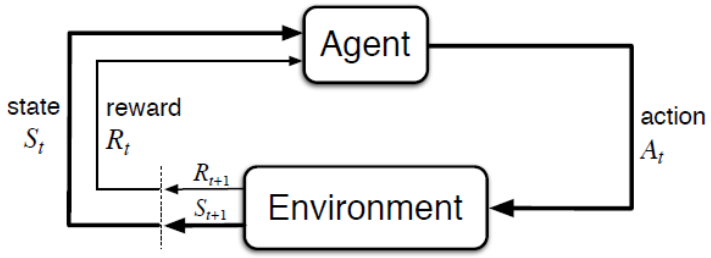


Figure 2.3: Markov decision process (MDP). At each time step t, the agent observes a state S_t, takes an action A_t, and receives a reward R_t from the environment, which transitions to a new state S_{t+1}. Image from [24]

cess. Simulation results across multiple datasets (MNIST, FEMNIST, and Sent140) demonstrate that this adaptive compression scheme significantly reduces communication latency—achieving up to a seven-fold improvement compared to traditional TDMA-based FL—while maintaining comparable model accuracy. The work highlights the practicality and efficiency of NOMA combined with gradient compression techniques in FL applications with constrained wireless communication resources.

2.2 Reinforcement Learning

Reinforcement learning is an emerging unsupervised machine learning method based on Markov decision process (MDP). A MDP involves the environment and agents, and there are three elements in a MDP, called state, action, and reward. Fig.2.3 shows a typical MDP workflow.

At each step, the agent takes an action based on its current state to interact with the environment and observes the resulting reward and the next state from the environment. Denoted by $\pi(\mathbf{a} \mid \mathbf{s})$ the probability of

selecting action $\mathbf{a}$ with given state $\mathbf{s}$, the action strategy is given by

$$\pi(\mathbf{a} \mid \mathbf{s}) = \mathbb{P}[\mathbf{A}_t = \mathbf{a} \mid \mathbf{S}_t = \mathbf{s}] \tag{2.2}$$

where $\mathbf{S}_t$ and $\mathbf{A}_t$ denote the state and action in MDP, respectively. The agent aims at maximizing the return in a MDP episode, which is defined as

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k} \tag{2.3}$$

where γ denotes the discount factor which weights the rewards at different time instance. $V_\pi(\mathbf{s})$ denotes the state-value function that measures the expected return with respect to a given state at the start and a given policy π, which is defined as

$$\begin{aligned} V_\pi(\mathbf{s}) &= \mathbb{E}_\pi[G_t \mid \mathbf{S}_t = \mathbf{s}] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(\mathbf{S}_{t+1}) \mid \mathbf{S}_t = \mathbf{s}] \\ &= \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a} \mid \mathbf{s}) \Big(\mathcal{R}_{\mathbf{s}}^{\mathbf{a}} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}} V_\pi(\mathbf{s}') \Big) \end{aligned} \tag{2.4}$$

where $\mathcal{R}_{\mathbf{s}}^{\mathbf{a}}$ and $\mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}}$ denote the average reward at state $\mathbf{s}$ and the transition probability from state $\mathbf{s}$ to state $\mathbf{s}'$ when taking action $\mathbf{a}$, respectively. To further quantize the expected return with respect to a given state-action pair

at the start and a given policy π, action-value function $Q_\pi(\mathbf{s}, \mathbf{a})$ is defined as

$$\begin{aligned} Q_\pi(\mathbf{s}, \mathbf{a}) &= \mathbb{E}_\pi[G_t \mid \mathbf{S}_t = \mathbf{s}, \mathbf{A}_t = \mathbf{a}] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(\mathbf{S}_{t+1}) \mid \mathbf{S}_t = \mathbf{s}, \mathbf{A}_t = \mathbf{a}] \\ &= \mathcal{R}_\mathbf{s}^\mathbf{a} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^\mathbf{a} V_\pi(\mathbf{s}'). \end{aligned} \tag{2.5}$$

Since

$$V_\pi(\mathbf{s}) = \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a} \mid \mathbf{s}) Q_\pi(\mathbf{s}, \mathbf{a}) \tag{2.6}$$

$Q_\pi(\mathbf{s}, \mathbf{a})$ can be calculated by

$$Q_\pi(\mathbf{s}, \mathbf{a}) = \mathcal{R}_\mathbf{s}^\mathbf{a} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^\mathbf{a} \sum_{\mathbf{a}' \in \mathcal{A}} \pi(\mathbf{a}' \mid \mathbf{s}') Q_\pi(\mathbf{s}', \mathbf{a}') \tag{2.7}$$

(2.4) and (2.7) are known as the Bellman expectation equations. Denoted by $V_*(\mathbf{s})$ and $Q_*(\mathbf{s}, \mathbf{a})$ the optimal state-value function and optimal action-value function, respectively, the optimal policy π_* that achieves $V_*(\mathbf{s})$ can be found by

$$\mathbf{a}_n(t) = \begin{cases} 1 & \text{, if } \mathbf{a} = \arg\max_{\mathbf{a} \in \mathcal{A}} Q_*(\mathbf{s}, \mathbf{a}) \\ 0 & \text{, otherwise.} \end{cases} \tag{2.8}$$

Thus $V_*(\mathbf{s})$ is given by

$$\begin{aligned} V_*(\mathbf{s}) &= \max_\mathbf{a} Q_*(\mathbf{s}, \mathbf{a}) \\ &= \max_\mathbf{a} \mathcal{R}_\mathbf{s}^\mathbf{a} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^\mathbf{a} V_*(\mathbf{s}') \end{aligned} \tag{2.9}$$

and $Q_*(\mathbf{s}, \mathbf{a})$ is given by

$$\begin{aligned} Q_*(\mathbf{s}, \mathbf{a}) &= \mathcal{R}_{\mathbf{s}}^{\mathbf{a}} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}} V_*(\mathbf{s}') \\ &= \mathcal{R}_{\mathbf{s}}^{\mathbf{a}} + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}_{\mathbf{ss}'}^{\mathbf{a}} \max_{\mathbf{a}'} Q_*(\mathbf{s}', \mathbf{a}') \end{aligned} \tag{2.10}$$

(2.9) and (2.10) are known as the Bellman optimality equations. However, since transition probabilities are unknown in many scenarios, it is impossible to adopt the dynamic programming method. Therefore, the following discussions focus on model-free methods such as Q-learning.

2.2.1 Value-Based Method

There are two types of value-based RL algorithms, called on-policy learning and off-policy learning, depending on whether the adopted policy is the same as the policy to be optimized. Sarsa is an on-policy RL algorithm, and the Q-value updating is given by

$$Q(\mathbf{s}, \mathbf{a}) \leftarrow Q(\mathbf{s}, \mathbf{a}) + \alpha Q_{Target} \tag{2.11}$$

where Q_{Target} denotes the target Q-value and is given by

$$Q_{Target} = R + \gamma Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}) \tag{2.12}$$

where $\mathbf{a}'$ is selected by the policy being optimized. Q-learning is one of the commonly used off-policy RL algorithms. The Q-value updating is same as

that for Sarsa in (2.11) but the Q_{Target} in Q-learning is given by

$$Q_{Target} = R + \gamma \max_{\mathbf{a}'} Q(\mathbf{s}', \mathbf{a}') - Q(\mathbf{s}, \mathbf{a}) \tag{2.13}$$

The Q-value in value-based methods can be estimated by function approximation, which is a mapping between states and values. Neural networks are commonly used in function approximation. Denoted by $\mathbf{W}$ the neural network parameter, the cost function of the function approximation is given by

$$L(\mathbf{w}) = \mathbb{E}\left[\Big(Q_{Target}(\mathbf{s}, \mathbf{a}) - Q_{\mathbf{w}}(\mathbf{s}, \mathbf{a})\Big)^2\right] \tag{2.14}$$

2.2.2 Policy-Based Method

Policy gradient methods directly parameterize a policy rather than a value function, which makes it possible to learn a stochastic policy. The policy objective function is defined as the expected reward at a random time step and is given by

$$\begin{aligned} J(\theta) &= \mathbb{E}_{\pi_\theta}[r] \\ &= \sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) \sum_{\mathbf{a}\in\mathcal{A}} \pi_\theta(\mathbf{s}, \mathbf{a}) \mathcal{R}_{\mathbf{s},\mathbf{a}} \end{aligned} \tag{2.15}$$

Where $d(\mathbf{s})$ denotes the stationary state distribution under policy π_θ, i.e., the probability of visiting state $\mathbf{s}$ when following the policy. The gradient with

respect to the learned model parameter θ is calculated by

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) \sum_{\mathbf{a}\in\mathcal{A}} \pi_\theta(\mathbf{s},\mathbf{a}) \nabla_\theta \log \pi_\theta(\mathbf{s},\mathbf{a}) \mathcal{R}_{\mathbf{s},\mathbf{a}} \\ &= \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(\mathbf{s},\mathbf{a}) Q_{\pi_\theta}(\mathbf{s},\mathbf{a})\right] \end{aligned} \tag{2.16}$$

The action-value function $Q_{\pi_\theta}(\mathbf{s},\mathbf{a})$ can be estimated using a separate learned model called critic so that the policy can be updated after every step rather than every episode. $Q_{\mathbf{w}}(\mathbf{s},\mathbf{a})$ denotes the action-value function estimated by the critic, and the cost function of the critic is the same as that in value-based method, which is presented in (2.14)

To reduce variance, a baseline function $B(\mathbf{s})$ can be adopted, which is subtracted from $Q_{\pi_\theta}(\mathbf{s},\mathbf{a})$, so that the policy gradient is given by

$$\begin{aligned}\nabla_\theta J^*(\theta) &= \nabla_\theta J(\theta) - \mathbb{E}_{\pi_\theta} \left[\nabla_\theta \log \pi_\theta(\mathbf{s},\mathbf{a}) B(\mathbf{s})\right] \\ &= \nabla_\theta J(\theta) - \sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) B(\mathbf{s}) \nabla_\theta \sum_{\mathbf{a}\in\mathcal{A}} \pi_\theta(\mathbf{s},\mathbf{a}) \\ &= \nabla_\theta J(\theta). \end{aligned} \tag{2.17}$$

Note that the second summation in (2.17) is zero since $\sum_{\mathbf{a}\in\mathcal{A}} \pi_\theta(\mathbf{s},\mathbf{a}) = 1$ for all $\mathbf{s}$, and thus its gradient with respect to θ vanishes. By adopting the baseline function, the expectation of the policy gradient does not change while the variance of the policy gradient is reduced. Value function is a potential candidate for the baseline function. The policy gradient when using

value function $V(s)$ as baseline function is given by

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{s},\mathbf{a})\Big(Q(\mathbf{s},\mathbf{a}) - V(\mathbf{s})\Big)\right] \\ &= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{s},\mathbf{a})\delta\right] \end{aligned} \tag{2.18}$$

where δ denotes the TD error and is given by

$$\delta = R + \gamma V(\mathbf{s}') - V(\mathbf{s}) \tag{2.19}$$

2.3 Index Modulation

Building on the distributed learning capabilities explored in Section 2.1 and the adaptive decision-making methods presented in Section 2.2, this section introduces index modulation as a promising technology for enhancing efficiency and adaptability in future wireless communication systems. Federated learning's decentralized model aggregation and reinforcement learning's capacity for dynamic, environment-responsive optimization both complement index modulation's intrinsic flexibility. By leveraging these advanced learning techniques, index modulation schemes can intelligently adapt to heterogeneous and resource-limited conditions, thus significantly improving spectral efficiency and resilience in multi-user wireless environments typical of beyond-6G scenarios.

Index modulation has been considered as an energy efficient method for the mMTC network in 6G [25], [26]. In conventional modulation schemes, all data bits are conveyed by the amplitude-phase modulation (APM) symbols, whereas in index modulation, part of data bits are conveyed by the in-

dices of active subcarriers or antenna. Exploiting the index modulation concept in the multi-carrier systems, various OFDM-IM schemes that have been proposed. [27] proposed a spread-OFDM-IM scheme achieving high transmit diversity, which applies a precoding matrix to the transmit signal. [28] proposed a scheme called coordinate interleaved OFDM-IM (CI-OFDM-IM) which seperately transmit the real and imaginary parts of the data symbol using different active subcarriers. [29] proposed a super-mode OFDM-IM (SuM-OFDM-IM) which forms index symbol via both mode activation patterns (MAPs) and subcarrier activation patterns (SAPs) at the same time to maximize the number of data bits transmitted by index symbol. Mainly focusing on coding gain and diversity gain, such existing OFDM-IM schemes have neglected to discuss new insights into learning-driven adaptive modulation signals, which could improve the spectral efficiency and reliability in multi-user heterogeneous environments.

Fig.2.4 shows a typical structure of OFDM-IM transmitters. It is assumed that the OFDM transmission bandwidth is divided into multiple sub-bands, where each sub-band is allocated to an individual user for data transmission. In orthogonal frequency-division multiplexing with index modulation (OFDM-IM), let N denote the total number of subcarriers, k denote the number of active subcarriers, and M denote the order of APM. The p incoming data bits are divided into two parts, p_1 and p_2:

$$p = p_1 + p_2 \tag{2.20}$$

where p_1 bits are employed by index symbols. Let $\mathbf{i}$ denote the index vector

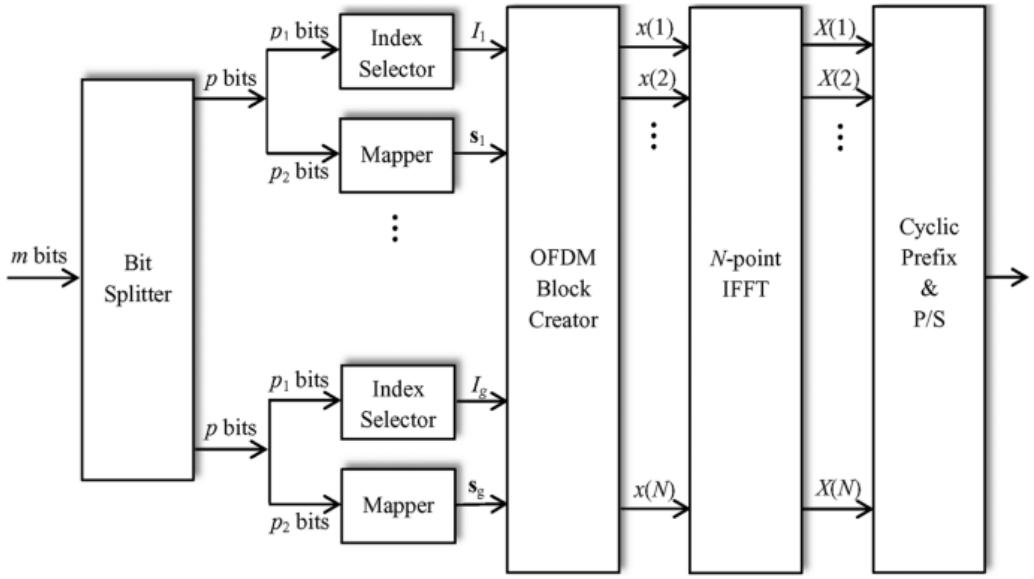


Figure 2.4: A typical structure of an OFDM-IM transmitter. The input bit stream is first split into sub-blocks. Each sub-block is divided into bits for index selection and bits for symbol mapping. The selected indices and mapped symbols form the input to the OFDM block, which undergoes an N-point IFFT followed by cyclic prefix addition and parallel-to-serial (P/S) conversion. Image from the paper [30]

containing the indices of active subcarriers in an OFDM symbol:

$$\mathbf{i} = \{i_1, ..., i_k\}, \tag{2.21}$$

where $i_k \in [1, ..., N]$ and

$$p_1 = \lfloor \log_2 C(N, k) \rfloor \tag{2.22}$$

p_2 bits are conveyed by APM symbols. The vector of APM symbols, denoted by s, is given by:

$$\mathbf{s} = [s_1, ..., s_k] \tag{2.23}$$

and

$$p_2 = k \log_2 |M| \tag{2.24}$$

The transmitted signal is:

$$\mathbf{x} = [x_1, ..., x_N] \tag{2.25}$$

where

$$\mathbf{x_i} = \mathbf{s}. \tag{2.26}$$

The unused subcarriers are set to zero. The transmitter will apply IFFT and add cyclic prefix (CP) to $\mathbf{x}$ and then transmit the time domain signal. The receiver will remove the CP and apply FFT to the received signal. The received signal in the frequency domain can be represented by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} \tag{2.27}$$

where $\mathbf{H}$ is the fading channel matrix, and $\mathbf{n}$ is the additive white Gaussian noise (AWGN) vector. At the receiver side, the detector will recover the data bits from y. Fig.2.4 shows a typical structure of OFDM-IM transmitter.

In spatial modulation (SM), index bits are conveyed by the index of selected transmit antenna (TA). Fig.2.5 shows a typical structure of SM transceiver. N_t denotes the number of antenna in transmitter. The income bits are splited into two parts, p_1 and p_1. For each transmission, one out of N_t antenna are selected to transmit the APM symbol. p_1 bits, which are conveyed by the index of selected antenna, satisfies

$$p_1 = \lfloor \log_2 N_t \rfloor \tag{2.28}$$

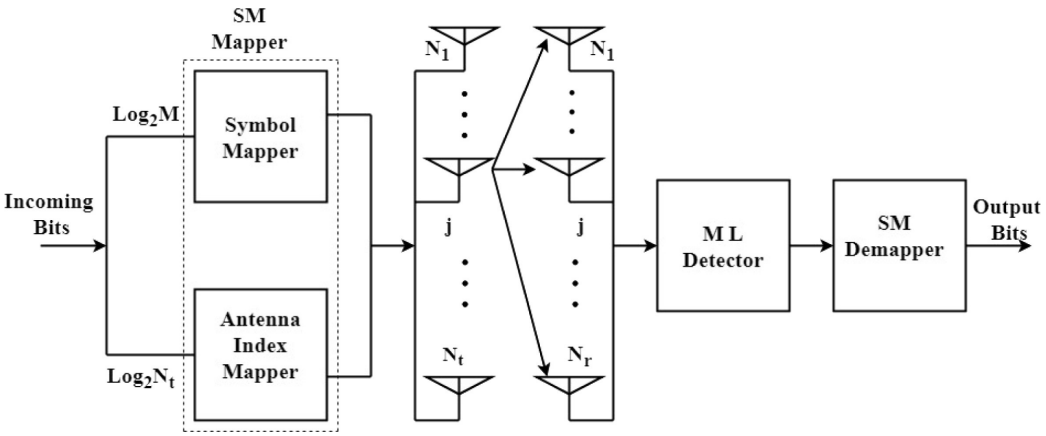


Figure 2.5: A typical structure of a Spatial Modulation (SM) transceiver. The incoming bit stream is divided into two parts: one part selects the transmit antenna via the Antenna Index Mapper, while the other is mapped to modulation symbols. These symbols are transmitted over the selected antennas. At the receiver, the ML detector identifies the active antenna and the transmitted symbol, and the SM Demapper recovers the original bits. Image from the paper [31]

The p_2 bits are conveyed by APM symbols, which is similar to that in OFDM-IM scheme.

Conventional adaptive modulation schemes require channel state information (CSI) at the transmitter while learning driven schemes can learn the pattern from the environment. In time division duplexing (TDD) applications, the CSI of downlink and uplink are the same so that downlink signal can be adopted as the indicator of CSI for uplink. K-means clustering can efficiently extract the implicit pattern of multi-dimensional vectors by clustering them according to the Euclidean distance between them. [32] proposed an OFDM-IM adaptation with the use of single user k-means clustering. However, the process of constituting a sufficient size of training dataset is expensive for resource limited MTC devices. Centralized learning is a possible solution leveraging the resources of central servers but it leads to high communication cost for sharing the data.

Federated learning (FL) is a potential enabler for connected intelligence in 6G [33]. Compared to centralized learning strategy, FL only shares the model updates instead of all the data, which reduces the consumption of spectral and energy resources. For example, [34] proposed a federated deep learning strategy for automatic modulation classification (AMC), which avoids data leakage while the performance loss is within 2% compared to the centralized algorithm.

The model aggregation strategy is a crucial challenge in federated learning. In FedSGD, a part of clients upload samples randomly selected from their local dataset and the model parameters will be updated in the FL server, which causes expensive cost in terms of communications. To optimize this weakness, an algorithm named FedAvg was developed [19]. In FedAvg, clients update their local parameters locally, and only upload the parameters to the FL server. The FL server will average the local parameters with appropriate weights to get global parameters. In [34], FedSGD and FedAvg based algorithms are proved to have similar performance when solving AMC problem. [20] proposed a federated stochastic variance reduced gradient (FSVRG) optimization algorithm, which improved the performance for non-independent, and identical data distribution. Such strategies focus on using deep learning, which may not suit to constrained MTC devices. Only a few paper developed federation strategy for k-means clustering. [21] proposed a federated k-means clustering algorithm based on FedAvg for image recognition. [35] proposed a federated k-means scheme for proactive caching, where the training data are shared for model aggregation at the high cost. Potential of effectively federating k-means clustering has been overlooked in

OFDM-IM variants at MTC devices.

Adaptive OFDM-IM schemes, such as the federated k-means clustering method discussed in Chapter 3, rely heavily on efficient multiple access and random access technologies to manage the simultaneous connectivity of numerous distributed users. Hence, advanced random access methods, like Non-Orthogonal Random Access (NORA), which efficiently manage resources and reduce congestion, are critically required to complement and fully realize the benefits of adaptive OFDM-IM systems.

2.4 Non-Orthogonal Random Access

Following the discussion on index modulation, this section introduces another emerging technology in wireless networks, Non-Orthogonal Random Access (NORA), an advancement combining Non-Orthogonal Multiple Access (NOMA) and random access protocols. NORA leverages the simultaneous resource-sharing capability of NOMA with the flexibility of random access procedures, effectively addressing the massive connectivity and sporadic traffic challenges anticipated in future 6G wireless networks. In this context, two foundational random access methods, the 4-step and 2-step random access protocols, are introduced, providing a clear understanding of their operational differences, strengths, and limitations. The combination of these technologies underpins the effectiveness of NORA in managing high-density device environments, offering enhanced throughput, reduced latency, and improved scalability.

2.4.1 Non-Orthogonal Multiple Access

Multiple access is a kind of technology which allows multiple users to share the same resources. There are two main types of multiple access schemes, orthogonal multiple access (OMA) and non-orthogonal multiple access (NOMA).

OFDMA is a main kind of OMA technologies, which is based on OFDM. In OFDMA, users are allocated with different subblock of adjacent subcarriers, which are orthogonal to each others, in frequency domain. According to the shannon capacity equation, the maximum throughput of the k-th user OFDMA system is given by

$$R_k = \frac{B}{N} \log_2(1 + \frac{P_k \mid h_k \mid^2}{N_{0,k}}) \tag{2.29}$$

Where B and N are the total bandwidth and total number of subcarriers, respectively. Note that in this equation, it is assumed that only one subcarrier is allocated to each user.

Conventional orthogonal random access (ORA) protocols, where each RB is allocated to a single user per transmission interval, restrict spectral efficiency and user connectivity. Recently, non-orthogonal random access (NORA) has gained attention for its potential to enhance spectral efficiency [36]. Power-domain NORA allows multiple users to access a single RB simultaneously by transmitting at different power levels [37]. A NOMA-ALOHA protocol combined with multichannel access was proposed in [38], achieving higher throughput than traditional multichannel ALOHA. However, such approaches often require channel state information (CSI) for channel selection and inversion, increasing the computational burden on MTC devices, and throughput remains constrained by user collisions.

In NOMA, all the users use the same frequency resources but with different transmit power. At the receiver side, a SIC detector is adopted to recover the signal of each user. The received signal is decoded by the order of decreasing received signal power [39]. The signal of user with highest power will be recover firstly, and the signals of other users, which has lower power than this one, will be treat as interference. After the detection. this signal will be subtracted from the total received signal to remove the interference from that user. In downlink NOMA, the transmit power order is in contrast to the channel gain order in order to achieve similar received signal power level for different users. The maximum throughput of the k-th user in an downlink NOMA system is

$$R_k = B\log_2(1 + \frac{P_k \mid h_k \mid^2}{\sum_{j=1}^{k-1} P_j \mid h_k \mid^2 + \bar{P}_{n,k}}) \tag{2.30}$$

Where P_k and $\bar{P}_{n,k}$ denotes the transmit power and noise power of k-th user, respectively [40]. Note that $h_1 > h_2 > \cdots > h_K$. The maximum throughput of the k-th user in an uplink NOMA system is

$$R_k = B\log_2(1 + \frac{P_k \mid h_k \mid^2}{\sum_{j=k+1}^{K} P_j \mid h_j \mid^2 + \bar{P}_n}) \tag{2.31}$$

[41] conclude the conditions need to be satisfied in NOMA power allocation. P_{tol} is defined as the minimum power difference between the signal to be recovered and the signal treated as inteference. For downlink NOMA, the conditions are given by

$$P_k h_{k-1} - \sum_{j=1}^{k-1} P_j h_{k-1} \geqslant P_{tol}, \quad k = 2, 3, \ldots, K \tag{2.32}$$

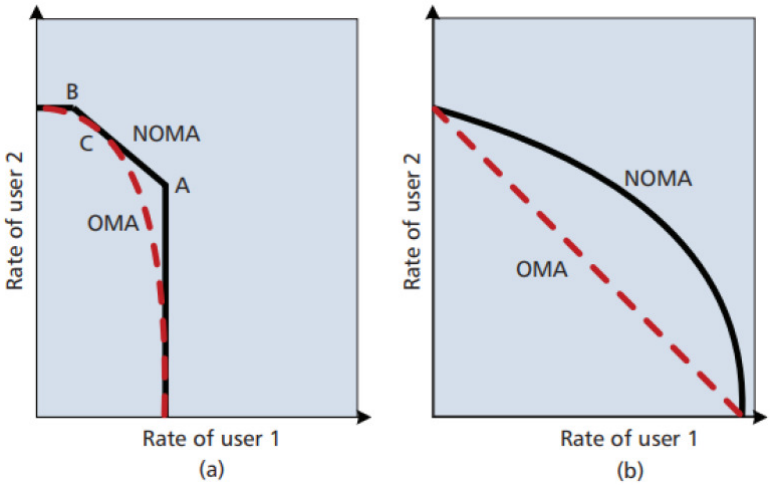


Figure 2.6: Throughput comparison between NOMA and OMA in (a)uplink and (b)downlink AWGN channel. Image from the paper [42].

$$\sum_{k=1}^{K} P_k \leqslant P_t \tag{2.33}$$

Where P_t is the maximum transmit power. For uplink NOMA, the conditions are given by

$$P_k h_k - \sum_{j=k+1}^{K} P_j h_j \geqslant P_{tol}, \quad k = 1, 2, \ldots, (K-1) \tag{2.34}$$

$$P_k \leqslant P_t \tag{2.35}$$

From (2.29), (2.30) and (2.31), it can be found that the interference NOMA suffered has a logarithmic level degradation on the capacity, while NOMA has N times OFDMA bandwidth. Fig.2.6 shows the throughput comparison between NOMA and OMA considering a two-user system with AWGN channel. It has been proved that NOMA can achieve higher capacity than OMA if the power allocation and user clustering are well designed [42].

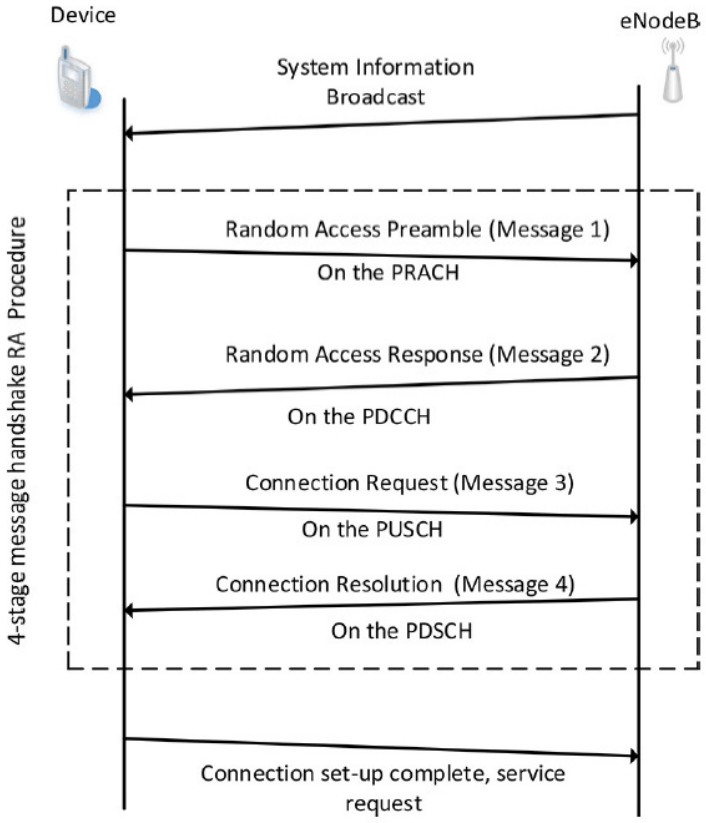


Figure 2.7: Signaling procedure in the 4-step Random Access (RA) process. The device initiates communication by sending a Random Access Preamble (Message 1) on the PRACH. The eNodeB responds with a Random Access Response (Message 2) on the PDCCH. The device then sends a Connection Request (Message 3) on the PUSCH, and the eNodeB completes the handshake with a Connection Resolution (Message 4) on the PDSCH. This process enables initial access and uplink synchronization. Image from the paper [1].

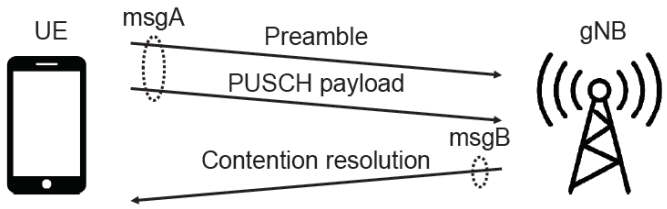
2.4.2 4-Step Random Access

Random access (RA) poses a significant challenge in future Internet of Things (IoT) networks due to the substantial number of connected devices, which leads to severe congestion in the radio access network (RAN) [43]. In traditional four-step grant-based random access (GB-RA) protocols, users must establish a handshake with the base station (BS) before transmitting any

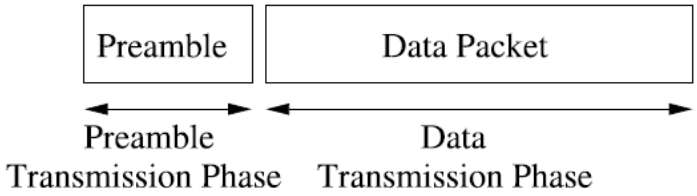
uplink data [1].

Fig.2.7 shows the 4-step RA procedure. An user starts the RA process by sending a preamble (Msg1) to the BS through the physical random access channel (PRACH), and then waits for random access response (RAR) for a period of time defined by the RAR window parameter. The BS broadcasts the index of identified preambles through the physical downlink control channel (PDCCH). If the user finds the preamble index it sent to the BS, it will perform the first data payload transmission (Msg3) using the resource block (RB) associated with the selected preamble on the physical uplink shared channel (PUSCH). After that, the user monitors for contention resolution (Msg4) from the BS on the physical downlink shared channel (PDSCH), and the RA process will be considered successful if the contention resolution identity in the received Msg4 is same as the transmitted one in Msg3.

However, grant-based random access (GB-RA) methods face limitations, including excessive signaling overhead and latency, particularly in machine-type communication (MTC) scenarios characterized by sporadic traffic and small payloads. This highlights the need to develop more efficient random access protocols. Two optimization strategies can be considered to mitigate latency associated with the 4-step RA procedure. These include reducing or entirely eliminating the delay experienced by the user equipment (UE) in receiving Msg2, as well as minimizing or fully removing the waiting period imposed by the network before the transmission of Msg4.



(a) Signaling in 2-step RA. Image from the paper [44].



(b) Two phases of MsgA in 2-step RA. Image from the paper [45]

Figure 2.8: 2-step Random Access (RA) procedure. (a) The User Equipment (UE) sends a combined preamble and payload (MsgA) to the gNB, followed by a contention resolution message (MsgB) from the gNB. (b) The MsgA comprises a preamble transmission phase and a data transmission phase.

2.4.3 2-Step Random Access

Grant-free random access (GF-RA) has emerged as a promising alternative due to its reduced signaling overhead and improved spectral efficiency. Unlike GB-RA, GF-RA enables users to transmit data by randomly selecting an RB without a handshake with the BS. A foundational GF-RA protocol, ALOHA [46], allows users to transmit their packets over a shared channel with an identical access probability. Slotted ALOHA (SA), a widely utilized variation of ALOHA, synchronizes user transmissions into predefined time slots. Analytical studies such as [47] and [48] examined the utilization and fairness of SA using Markov models. Further improvements, such as the protocol in [49], focused on minimizing the time-average age of information (AAOI) in slotted ALOHA systems.

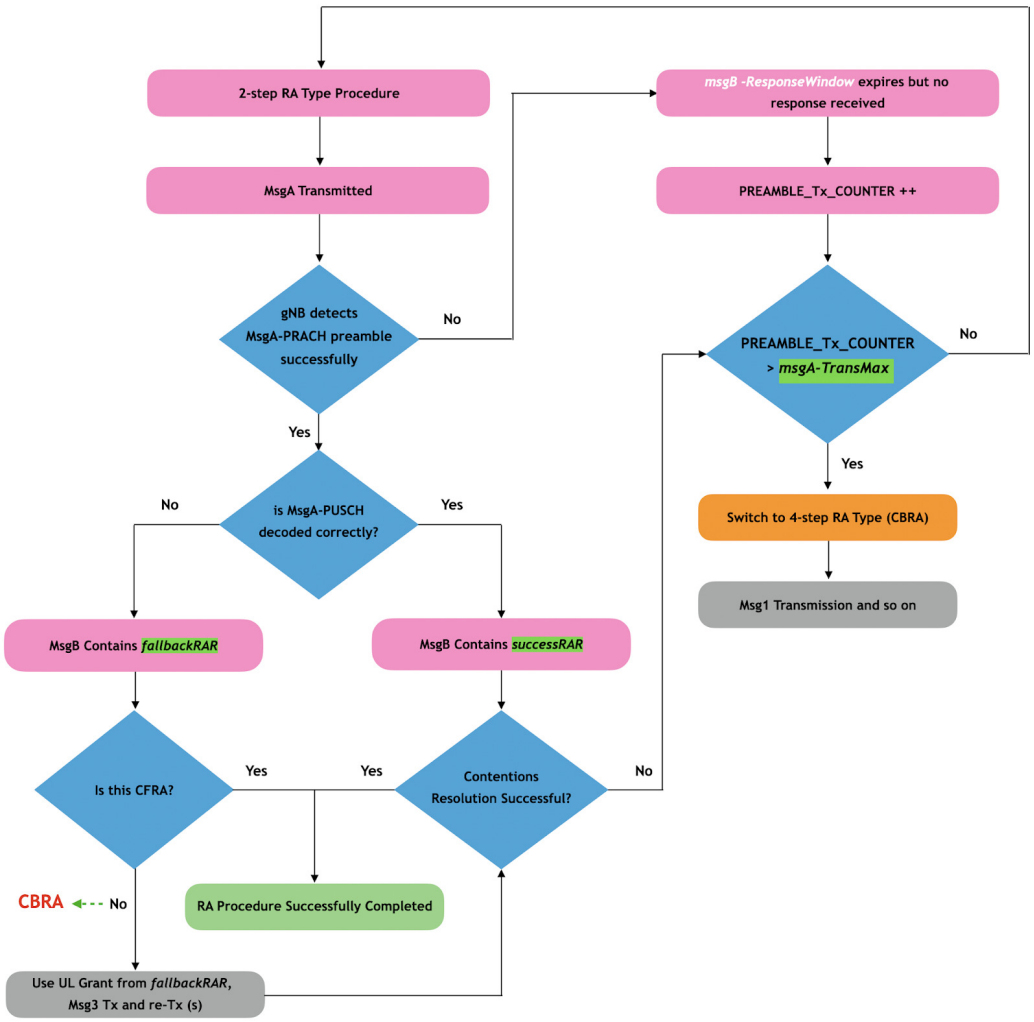


Figure 2.9: Workflow of the 2-step Random Access (RA) procedure. After transmitting MsgA, the gNB attempts to detect the preamble and decode the PUSCH. Depending on success or failure, it responds with either `successRAR` or `fallbackRAR`. If the maximum number of transmission attempts is exceeded without response, the procedure switches to the 4-step RA type (CBRA). Image from [50]

The 2-step Random Access (RA) procedure in 5G NR involves a streamlined approach compared to the conventional 4-step procedure, combining preamble and payload transmission into a single message (MsgA), which is shown in Fig.2.8. The detailed workflow is presented in Fig.2.9, and is sum-

marized as follows:

1. **Initiation:** The User Equipment (UE) initiates the procedure by sending MsgA, which integrates both the preamble (MsgA-PRACH) and payload data (MsgA-PUSCH).

2. **Preamble Detection:** The gNB attempts to detect the MsgA-PRACH preamble:

   a. If there is not any preamble collision and the gNB successfully detects the preamble, the gNB proceeds to decode the MsgA-PUSCH payload.

   b. otherwise, the UE awaits a response during the configured *msgB-ResponseWindow*.

3. **Payload Decoding:** Upon successful preamble detection, the gNB attempts to decode the MsgA-PUSCH:

   a. If decoding succeeds, the gNB responds with *successRAR* in MsgB.

   b. If decoding fails, the gNB issues a *fallbackRAR* in MsgB, prompting a fallback to the 4-step RA.

4. **UE Response to MsgB:**

   a. Upon receiving *successRAR*, the UE attempts contention resolution:

      - Successful resolution concludes the RA procedure.
      - Failed resolution triggers fallback procedures or contention-based random access (CBRA).

   b. Upon receiving *fallbackRAR*, the UE utilizes the provided uplink grant to transmit Msg3, transitioning into the traditional 4-step RA.

5. **Handling Non-response:** If no response (MsgB) is received within the *msgB-ResponseWindow*, the UE increments the *PREAMBLE_Tx_COUNTER*:

   a. If the counter exceeds the configured maximum (*msgA-TransMax*), the UE reverts to the 4-step RA (CBRA).

   b. Otherwise, the UE retransmits MsgA.

The above procedural flow highlights critical decision points that determine the success or fallback of the streamlined 2-step random access mechanism in 5G NR. Note that preamble collision or payload decoding failure cause RA failure, which should be alleviated by optimizing the users' resource selection strategies.

2.4.4 Related Works

Recent advances in learning-driven algorithms have demonstrated significant potential in addressing random access (RA) collision challenges, particularly in optimizing resource allocation and enhancing quality of service (QoS) without relying on explicit system models. Early contributions have explored various centralized and supervised learning frameworks. For example, centralized actor-critic deep reinforcement learning (DRL) was employed by [51] to jointly optimize unmanned aerial vehicle (UAV) altitudes and channel access probabilities under energy constraints. Similarly, supervised deep neural networks (DNNs) have been utilized effectively for power optimization in Non-Orthogonal Random Access (NORA) systems [52] and collision detection in adaptive Physical Uplink Shared Channel (PUSCH) resource allocation [53]. Cooperative multi-agent approaches using double

deep Q-networks (DDQN), originally introduced by [54], were also adopted to optimize grant-free NOMA (GF-NOMA) parameters, addressing critical ultra-reliable low-latency communication (URLLC) scenarios [55].

Subsequent research has shifted toward distributed, scalable, and robust methodologies to accommodate massive device connectivity and dynamic network environments typical of emerging wireless networks. Notably, game-theoretic approaches have been proposed to enhance spectral efficiency in NOMA-ALOHA systems [56]. Concurrently, data-driven frameworks have significantly advanced adaptive slotted ALOHA schemes, demonstrating improved throughput in multi-cell NOMA-based IoT networks [57]. Comprehensive reviews have underscored the pivotal role of artificial intelligence (AI)-driven signal processing techniques, advocating for these approaches as viable replacements for traditional orthogonal access methods in next-generation multiple access (NGMA) frameworks [58].

In the context of supporting dense IoT deployments, cell-free and massive multiple-input multiple-output (MIMO) architectures have become focal points. A hybrid DRL-based clustering mechanism proposed by [59] improved distributed antenna system operations, specifically in beamforming and access point (AP) selection. Additionally, clustering-based maximum likelihood detection [60], user equipment (UE) detection thresholds [61], and adaptive AP selection methods using log-likelihood ratios [62] have further enhanced performance in grant-free scenarios. Moreover, pilot-hopping strategies [63], analytical bounds for success probabilities [64], and exploration of space diversity for diverse services [65] have significantly contributed to efficient access coordination.

Research has also tackled resource efficiency under communication constraints through innovative approaches such as budget-constrained multi-player multi-armed bandit (MAB) protocols for cognitive-radio IoT [66] and simplified NOMA frameworks enabling autonomous device-side power and sub-carrier selection [67]. Multi-sequence spreading RA techniques [68] and stochastic geometry-based analyses for spreading-based NOMA [69] have also played a critical role in addressing interference and improving spectral efficiency.

Advancements in multi-agent reinforcement learning (MARL) have further propelled the field, particularly in mitigating collisions and managing interference. The QMIX algorithm applied by [70] demonstrated notable improvements in decoding success rates within GF-NORA contexts. MARL strategies for interference management in ultra-dense networks [71], sparsity-aware RL algorithms for massive MIMO user detection [72], and contextual MAB methods for NOMA-based scheduling [73] have substantially advanced distributed multi-agent capabilities. Additionally, distributed Q-learning approaches for satellite-terrestrial networks [74] and insights into the evolution from 5G random access to unsourced multiple access (UMAC) in 6G [75] further underscore the practical benefits of distributed learning frameworks.

Federated and distributed reinforcement learning has increasingly attracted attention due to its potential for coordinated yet decentralized decision-making. Federated DRL protocols enhancing fairness and throughput in WiFi networks [76] and MARL frameworks optimizing spectrum sharing in vehicular communication scenarios [77] illustrate the diverse applications of these methods. DRL-based uplink resource allocation tailored for URLLC

was developed by [78], significantly reducing latency. Moreover, deep neural network-assisted double-contention RA schemes [79] and device-to-device clustering integrated with Q-learning [80] have provided effective means to handle collision and scalability challenges.

Distributed learning strategies specifically targeting GF-RA have demonstrated efficacy, although challenges remain regarding scalability and channel state information (CSI) dependency. Notable contributions include MARL frameworks employing centralized training with decentralized execution [81, 82], POMDP-based accelerated Q-learning [83], and congestion-aware Q-learning approaches [84, 85]. While efforts addressing user heterogeneity in NOMA-ALOHA have made significant strides [86], limitations persist regarding scalability and CSI requirements in distributed Q-learning environments [15, 87–90].

To overcome these scalability and adaptability challenges, recent works have extensively explored dynamic multichannel access frameworks using advanced DRL methods. Deep Q-networks (DQN) developed by [91] effectively addressed the challenges of unknown channel dynamics and extensive state spaces, significantly surpassing heuristic methods. Further advances through deep actor-critic reinforcement learning frameworks by [92] demonstrated improved scalability and computational efficiency, particularly in decentralized multi-agent contexts, dramatically reducing collision rates and optimizing spectrum utilization.

Despite substantial progress, existing literature still lacks a comprehensive, unified, and scalable distributed DRL framework capable of concurrently managing throughput, fairness, collision avoidance, and robustness against

fading in heterogeneous multi-AP NORA environments.

2.5 Summary

This chapter reviewed recent advancements in two major areas relevant to this thesis: Federated Learning (FL)-driven Index Modulation (IM) and Non-Orthogonal Random Access (NORA). In the context of FL-driven IM, research has demonstrated that FL enables the development of a global model by aggregating local learning outcomes from distributed devices. This approach efficiently leverages the computing power of all devices in a distributed system, thereby enhancing adaptive modulation strategies without requiring large training datasets. Such frameworks are particularly well-suited for multi-user uplink scenarios in heterogeneous and resource-constrained environments. Additionally, multiple access (MA) is a critical component for enabling local model uploading in FL. Non-Orthogonal Random Access (NORA), as an emerging MA technique, has gained significant attention for improving grant-free access and uplink performance in massive machine-type communication (mMTC) scenarios. Key focus areas include collision resolution, successive interference cancellation, and power-domain multiplexing strategies that support simultaneous transmissions. Together, these insights provide a solid foundation for the learning-driven, decentralized communication frameworks proposed in the following chapters. The next chapter introduces a novel federated k-means clustering framework tailored for adaptive OFDM-IM, serving as a practical implementation of the principles discussed herein. This transition marks the beginning of applying these theories to develop efficient, distributed modulation and access strategies for

next-generation wireless networks.

3 Federated K-Means Clustering for Adaptive OFDM-IM

3.1 Introduction

In this chapter, a federated k-means clustering called Fed-k-means is developed to obtain the precise adaptive OFDM-IM model enhancing the system throughput with less training data at devices. The main contributions of this work are: (i) to develop the multi-user adaptive OFDM-IM system with the use of federated k-means clustering; (ii) to develop a novel weighting strategy for the federated k-means clustering to provide the accurate adaptation strategy of OFDM-IM signals; (iii) to evaluate the effective throughput of the system by simulations and the simulated results clearly present that the proposed system can outperform the benchmarks, in terms of the effective throughput.

3.2 System Model

Consider a distributed multi-carrier system where B users send uplink data packets to the base station (BS) by employing OFDM-IM in TDD mode. Fig. 3.1 shows the system diagram. For the OFDM-IM transmission, assume that

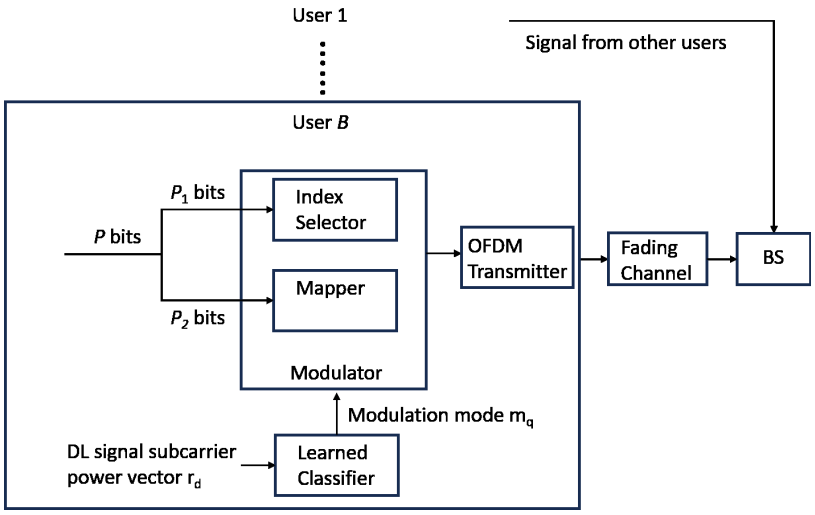


Figure 3.1: Learning-driven adaptive OFDM-IM system for uplink transmission. $\mathbf{m}_q$ and $\mathbf{r}_d$ denote the selected modulation mode and the downlink signal subcarrier power vector, respectively.

the sub-block size for each user is N. Since the communication and learning process run independently across users, for simplify, the following discussion will focus on one user. The communication process are introduced here with the learning process in Section 3.3.

At every transmission, each user intends to adjust its modulation mode. The KT modulation modes are combinations of T modulation orders and K different numbers of active subcarriers. Denote the t-th modulation order and the k-th number of active subcarriers by $M_t \in \{M_1, \cdots, M_T\}$ and $k \in \{1, \cdots, K\}$, respectively. $\mathbf{m}_q \in \{\mathbf{m}_0, \cdots, \mathbf{m}_{KT}\}$ is the q-th modulation mode, where $\mathbf{m_0}$ means no transmission and

$$\mathbf{m}_q \triangleq \begin{cases} (0,0) & \text{if } q = 0 \\ (k, M_t) & \text{if } q \in \{1, \cdots, KT\} \end{cases} \tag{3.1}$$

Given $\mathbf{m}_q$, there are p bits to be transmitted. Denote $C(N,k)$ the number of possible combinations of k active subcarriers over N subcarriers, p is given by

$$\begin{aligned} p &= p_1 + p_2 \\ &= \lfloor \log_2 C(N,k) \rfloor + k \log_2 M_t \end{aligned} \tag{3.2}$$

$\mathbf{s} = [s_1, ..., s_k]^T$ is the modulated symbol vector, $s_k \in \mathcal{S}_t$, where $\mathcal{S}_t$ is the constellation of M_t-ary modulation. Notice that for a given k, the index subset is $\mathbf{i} = \{i_1, \cdots, i_k\}$ where $i_a \in \{1, \cdots, N\}$ for $a = 1, \cdots, k$. The OFDM-IM signal vector of each user in the frequency domain is $\mathbf{x} = [x_1, ..., x_N]^T$, where

$$x_n = \begin{cases} s_m & \text{if } n = i_m \in \mathbf{i} \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

The $N \times N$ Zadoff-Chu precoding matrix by [27] is adopted, which is given by

$$\mathbf{G} = \frac{1}{\sqrt{N}} \begin{bmatrix} c_1 & c_2 & \cdots & c_N \\ c_N & c_1 & \cdots & c_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ c_2 & c_3 & \cdots & c_1 \end{bmatrix}, \tag{3.4}$$

where the root Zadoff–Chu sequence is defined as

$$c_n = \begin{cases} \exp\left(-j\frac{2\pi m}{N}\left(\frac{n^2}{2}+qn\right)\right), & \text{if } N \text{ is even,} \\ \exp\left(-j\frac{2\pi m}{N}\left(\frac{n(n+1)}{2}+qn\right)\right), & \text{if } N \text{ is odd,} \end{cases} \quad n = 1, \ldots, N, \tag{3.5}$$

with m an integer relatively prime to N and q an integer shift parameter.

The Zadoff–Chu precoding matrix $\mathbf{G}$ is unitary ($\mathbf{G}^{-1} = \mathbf{G}^{H}$), ensuring that each non-zero symbol is spread with constant magnitude across all N subcarriers. This provides frequency diversity, preserves Euclidean distances, and enables low-complexity detection through simple de-spreading.

Denote by $\mathbf{H} = \text{diag}\{h_1, ..., h_N\}$, the frequency domain channel matrix whose elements h_i are complex Gaussian random variables with $h_i \sim \mathcal{CN}(0, 1)$. In the frequency domain, the uplink received signal is

$$\mathbf{y} = \sqrt{\bar{\alpha}}\mathbf{HGx} + \mathbf{n} \tag{3.6}$$

where $\mathbf{n}$ is the Additive White Gaussian Noise (AWGN) vector whose elements follow $\mathcal{CN}(0, 1)$, and $\bar{\alpha}$ denotes the average channel gain. Since $\mathbf{H}$ has diagonal elements with unit variance, $\mathbf{G}$ is unitary, and $\mathbf{x}$ is power-normalised with unit mean square amplitude, the average power of $\mathbf{HGx}$ equals that of $\mathbf{n}$. Therefore, the average SNR reduces to $\bar{\alpha}$ under these assumptions. At the BS side, maximum likelihood (ML) detector is adopted to recover the signal of each user.

$\omega(\mathbf{x}, \hat{\mathbf{x}})$ is the number of error bits when $\mathbf{x}$ is decoded as $\hat{\mathbf{x}}$. The upper bound of conditional bit error probability with $\mathbf{m}_q$ and channel matrix $\mathbf{H}$ is

given by [27]

$$P_b(\mathbf{m}_q, \mathbf{H}) \leq \frac{1}{pCM_t^k} \sum_{\mathbf{x}} \sum_{\hat{\mathbf{x}}} \omega(\mathbf{x}, \hat{\mathbf{x}}) Q\left(\sqrt{\frac{\sqrt{P\bar{\alpha}}\|\mathbf{HG}(\mathbf{x} - \hat{\mathbf{x}})\|^2}{2N_0}}\right) \tag{3.7}$$

The lower bound of instantaneous effective throughput when using modulation mode $\mathbf{m}_q$ is given by

$$T_E(\mathbf{m}_q, \mathbf{H}) \geq p(1 - P_b(\mathbf{m}_q, \mathbf{H})) \tag{3.8}$$

The effective throughput is measured by

$$\begin{aligned} T_{Eavg} &= \mathbb{E}[\sum_{q=0}^{KT} T_E(\mathbf{m}_q, \mathbf{H}) Z(\mathbf{m}_q)] \\ &= \sum_{q=0}^{KT} (\mathbb{E}[T_E(\mathbf{H}|\mathbf{m}_q)] \Pr(\mathbf{m} = \mathbf{m}_q|\mathbf{F})) \end{aligned} \tag{3.9}$$

where $Z(\mathbf{m}_q)$, the indicator on whether modulation mode $\mathbf{m}_q$ is chosen in i-th transmission, is given by

$$Z(\mathbf{m}_q) \triangleq \begin{cases} 1 & \text{if } \mathbf{m}_q \text{ is chosen} \\ 0 & \text{otherwise,} \end{cases} \tag{3.10}$$

and $\Pr(\mathbf{m} = \mathbf{m}_q|\mathbf{F})$ is the probability of choosing modulation mode $\mathbf{m}_q$ with a given learned model $\mathbf{F}$. In the simulations presented in this chapter, the effective throughput is measured by the average number of correctly decoded bits per transmission.

It can be seen that the T_{Eavg} is effected by the learned-model $\mathbf{F}$. The

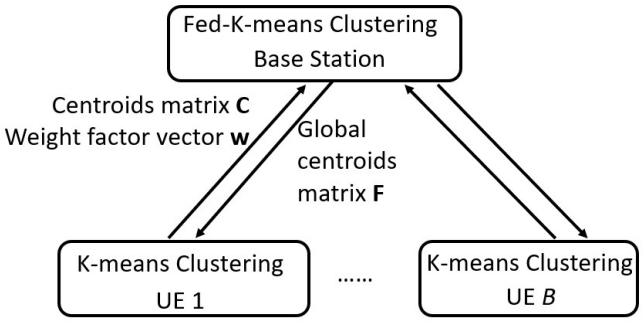
main focus of this chapter is to develop a federated learning assisted adaptive OFDM-IM algorithm to find the best modulation mode, enhancing the throughput without CSI.

3.3 Federated Clustering Adaptive OFDM-IM

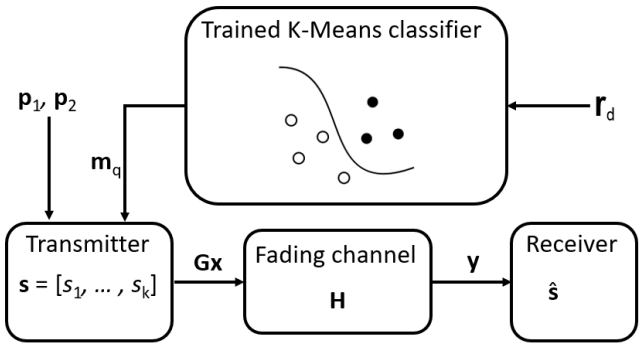
This section introduces the learning process of the proposed system. Fig.3.2(a) shows the structure of federated learning. In this phase, the BS uses the local models collected from B users to develop a global model and send it to B users. Fig.3.2(b) illustrates the structure of the learning-driven adaptive OFDM-IM, where $\mathbf{r}_d$ is the vector of downlink signal energy which is acquired by observing the downlink signal of each subcarrier at the user. In this stage, each user uses $\mathbf{r}_d$ as the input of the learned model to predict the modulation mode maximizing the effective throughput.

Consider that each user has V different data vectors to train itself over the learning process. The v-th training data vector contains $(|h_{1v}|^2, ..., |h_{Nv}|^2, q)$, $v \in \{1, \cdots, V\}$, where the first N entries, $|h_{iv}|^2$, represent uplink subcarrier gains. The $(N+1)$-th element is the best modulation mode, which achieves maximum effective throughput for the uplink channel $\mathbf{H}$ over the learning process. The best modulation mode estimation q for a given channel matrix $\mathbf{H}$ is obtained by

$$\begin{aligned} q &= \arg\max_{j} T_E(\mathbf{m}_j, \mathbf{H}) \\ s.t. \quad & P_e(\mathbf{m}_j, \mathbf{H}) \leq \mu \end{aligned} \tag{3.11}$$



(a) Federated learning structure



(b) Learning-driven adaptive OFDM-IM framework

Figure 3.2: The structure of federated clustering adaptive OFDM-IM. (a) shows the federated K-means clustering process where the base station aggregates local centroids and weight factors from users to update a global model. (b) illustrates the learning-driven adaptive OFDM-IM framework, where each user utilizes the trained model with the observed downlink signal vector $\mathbf{r}_d$ to predict the optimal modulation mode.

where μ is the BEP threshold.

To this end, denote, first, b-th user exploits its local dataset obtained from experienment in advance to get local centroids matrix $\mathbf{C}^b$ by using Algorithm 1. The superscripts represent user index.

The BS collects local centroids from the users after the local training process, and performs another clustering algorithm with the local centroids to develop a global model. In particular, computing the federated centroids

Algorithm 1 Fed-k-means local model training at the b-th user

Denote $\mathbf{Z}^b$ the matrix whose columns containing the training data, $\mathcal{S}_k^b$ the index set of the training data points assigned to the k-th cluster, and $|\mathcal{S}_k^b|$ the number of elements in $\mathcal{S}_k^b$
Input(s): $(N+1) \times V$ training data matrix $\mathbf{Z}^b$, number of local clusters K_C
Output(s): $(N+1) \times K_C$ local centroids matrix $\mathbf{C}^b$, $1 \times K_C$ weights vector $\mathbf{w}^b$
Initialization: Randomly generate initial local centroids matrix $\mathbf{C}^b$ and then assign each data point, $\mathbf{Z}_{1:N,v}^b$, $v \in [1, V]$, to its closest local cluster
Repeat until $\mathbf{C}^b$ does not change:
(1) Update $\mathbf{C}^b$
for $k = 1, ...K_C$ **do**
$\mathbf{C}_{1:N,k}^b \leftarrow \frac{1}{|\mathcal{S}_k^b|} \sum_{v \in \mathcal{S}_k^b} \mathbf{Z}_{1:N,v}^b$
Set $\mathbf{C}_{N+1,k}^b$ to the modulation mode which appears most frequserntly in the k-th cluster
end for
(2) Assign each data point, $\mathbf{Z}_{1:N,v}^b$, $v \in [1, V]$, to its closest cluster and then calculate the weight for each centroid
for $k = 1, ...K_C$ **do**
$\mathbf{w}_k^b \leftarrow |\mathcal{S}_k^b|$
end for

in the global model, the sum of weighted local centroids are iteratively considered as shown in Algorithm 2. The weight coefficients of local centroids represent the number of data points in the local clusters, which can be used to indicate the relative importance of local centroids with large data points against those with small data points. Based on these, the global model is computed. The global model accuracy increases when the number of users increases by implicitly leveraging more training data.

Such federated clustering is designed to decrease the loss function, which is given by

$$\mathcal{L}(\mathbf{F}) = \sum_k \sum_{l \in \mathcal{A}_k} \mathbf{w}_l \|\mathbf{F}_{1:N,k} - \mathbf{C}_{1:N,l}\|_2^2 \tag{3.12}$$

Algorithm 2 Fed-k-means global model updating

Denote $\mathcal{A}_k$ the index set of the local centroids assigned to the k-th global cluster
Input(s): $\mathbf{C} = [\mathbf{C}^1, \cdots, \mathbf{C}^B]$, $\mathbf{w} = [\mathbf{w}^1, \cdots, \mathbf{w}^B]$, number of global clusters K_G
Output(s): $(N+1) \times K_G$ global centroids matrix $\mathbf{F}$
Initialization: Randomly generate initial global centroids matrix $\mathbf{F}$ and then assign each data point, $\mathbf{C}_{1:N,l}$, $l \in [1, L]$, $L = BK_C$, to its closest global cluster
Repeat until F does not change:
(1) Update $\mathbf{F}$
for $k = 1, ...K_G$ **do**
$\mathbf{F}_{1:N,k} \leftarrow \frac{1}{\sum_{l \in \mathcal{A}_k} \mathbf{w}_l} \sum_{l \in \mathcal{A}_k} \mathbf{w}_l \mathbf{C}_{1:N,l}$
Set $\mathbf{F}_{N+1,k}$ to the modulation mode which appears most frequserntly in the k-th global cluster
end for
(2) Assign each local centroid, $\mathbf{C}_{1:N,l}$, $l \in [1, L]$, to its closest global cluster

Once the global model is updated, each user is assumed to access the global model and predict their best modulation mode to be used at each adaptive transmission. During the process of predicting the modulation mode maximizing the effective throughput, the downlink signal energy vector $\mathbf{r}_d$ is adopted as the input of the online prediction model. The prediction scheme is shown in Algorithm 3.

Algorithm 3 Prediction algorithm

Input(s): Global centroids matrix $\mathbf{F}$, downlink signal energy vector $\mathbf{r}_d$
Output(s): Index of modulation mode q
for every transmission **do**
Predict q (assign $\mathbf{r}_d$ to its closest global centroid)
Find $l = \arg\min_k \|\mathbf{F}_{1:N,k} - \mathbf{r}_d\|_2^2$
$q \leftarrow \mathbf{F}_{N+1,l}$
end for

3.4 Simulation Results

Simulation results of the proposed algorithms in the distributed adaptive OFDM-IM systems are presented. To measure their efficacy, the focus is on two simulation scenarios: (i) the effective throughput per user and BER performance in different average SNRs of the Fed-k-means and the single user k-means strategy; and (ii) the sensitivity of the federated OFDM-IM adaptation with different number of users in terms of effective throughput. For all the simulations, Rayleigh fading channel is applied to each subcarrier, the number of subcarriers $N = 4$, the number of active subcarriers $k \in \{1, 2\}$, the cardinality of possible modulation constellations $M_t \in \{0, 2, 4\}$ since BPSK and QPSK are considered. The BEP threshold $\mu = 0.01$. By using the well known elbow method, the number of clusters of the local model, K_C, are found to be 10, 20, 40, 100 for the training dataset sizes of 50, 100, 200, 500, respectively, and the number of clusters of the global model, K_G, is chosen to 100.

In Fig.3.3 and Fig.3.4, the effective throughput and BER of the four schemes, (i) Classical k-means with 200 training data; (ii) Fed-k-means with 200 training data at each user; (iii) Classical k-means with 500 training data; and (iv) Non-adaptive modulation, are presented in a 70 users scenario. The theoretical lower bounds of effective throughput and the theoretical upper bounds of BER, with a given learned model, are also depicted for validation. In addition, the optimal adaptive modulation based on CSI is simulated as well. Note that the performance of the optimal adaptive modulation based on CSI is not available at extreme low SNR (<4dB) since the majority of the decisions will be the no transmission mode, which makes it unfeasible

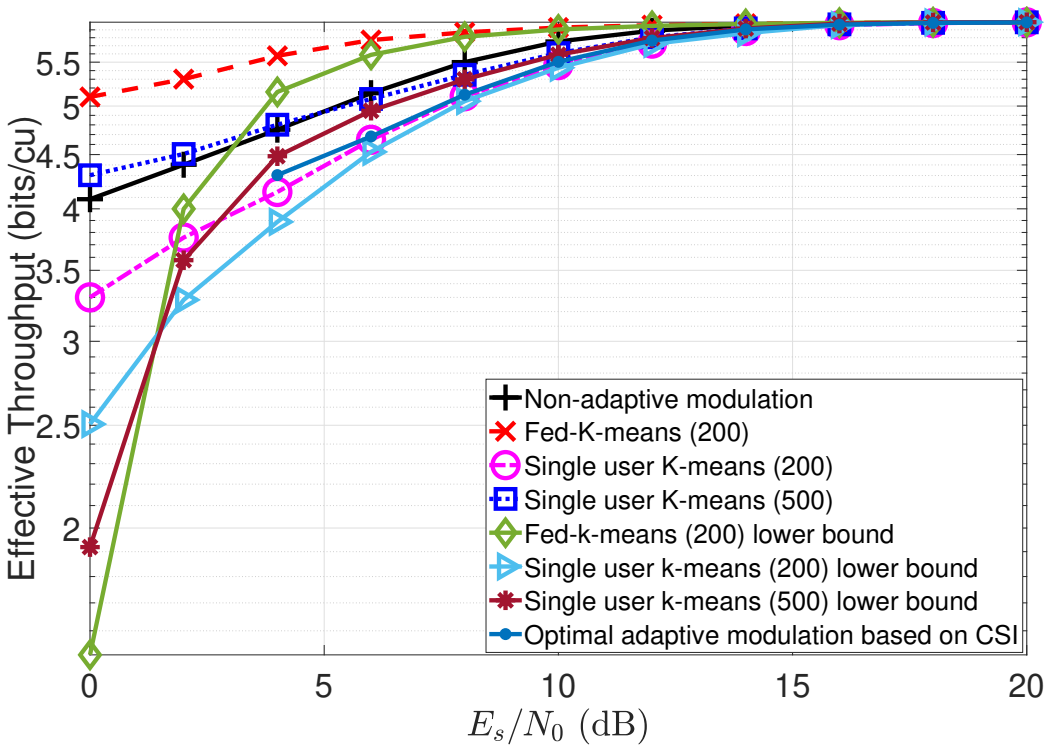


Figure 3.3: Effective throughput per user versus average SNR of Fed-k-means adaptive OFDM-IM with 70 users.

to collect enough samples for simulations. At mid and low SNRs, the Fed-k-means has higher effective throughput than the single user k-means with either 200 training data or 500 training data. The BER of the Fed-k-means and single user k-means are similar, which are lower than the non-adaptive modulation. When the SNR is greater than 14dB, the effective throughput of all the four schemes are similar because the mode with highest data rate becomes the majority choice. These results show that the proposed Fed-k-means algorithm can improve the effective throughput with an even smaller training dataset than the single user k-means algorithm.

In Fig.3.5, the effects of number of users on the effective throughput of the Fed-k-means schemes with different size of training dataset are depicted.

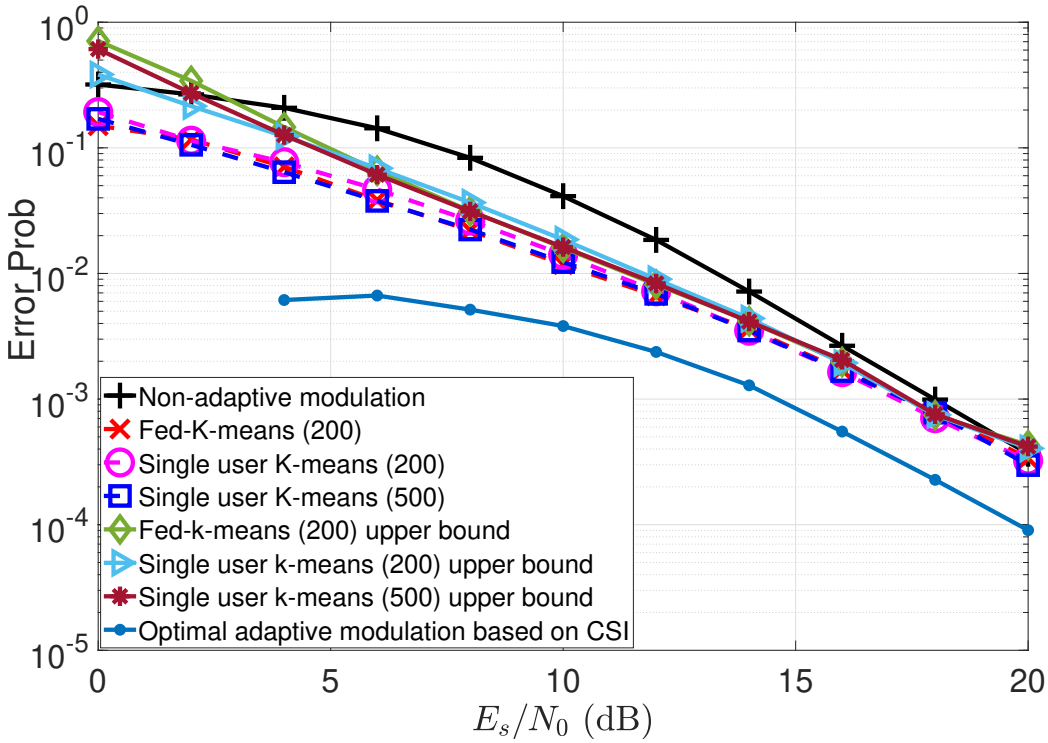


Figure 3.4: Average BER versus average SNR of Fed-k-means adaptive OFDM-IM with 70 users.

Table 3.1: Computational Complexity

Local training complexity	Global aggregation complexity
$\mathcal{O}(K_C V)$	$\mathcal{O}(K_G B)$

Note that all the results in this part are average values of 50 simulations, and the SNRs for all the three settings are 4dB. The effective throughput increases when the number of users increases. The plots with 200 and 100 training data points reaches their highest effective throughput, at 5.5 bits/cu, at 50 users and 100 users, respectively. This result indicates that the required number of users for achieving the best performance decreases when the size of training dataset in each user increases. Table 3.1 shows the computational complexity of the local training and the global aggregation process. The

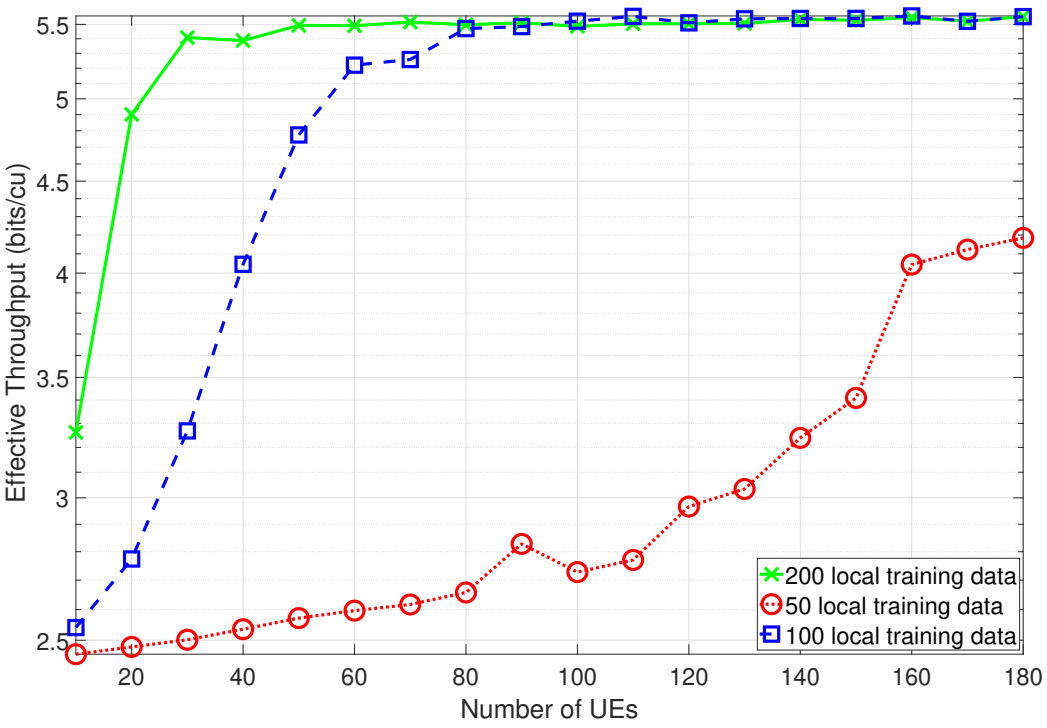


Figure 3.5: Effective throughput per user versus number of users of Fed-k-means adaptive OFDM-IM.

local computational complexity increases with the number of local centroids and the size of local training dataset. However, the increase of local training data leads to reduced number of required users for achieving the same performance, which reduces the complexity of the global model aggregation. Therefore, it is a trade off between local complexity and global complexity according to scenarios.

3.5 Summary

This chapter proposed the federated k-means clustering strategy for adaptive OFDM-IM. By aggregating the learning outcome of distributed users, the adaptation strategy developed at the BS improved the accuracy of the

global learning model, requiring less training data from individual devices. With the global adaptation model, distributed users were able to reliably adjust OFDM-IM signals to their local conditions. The simulation results showed that the Fed-k-means OFDM-IM improved the throughput through the multi-user federation. Heterogeneous training features across users such as asymmetric sets of modulation modes will be investigated in the future.

Although the federated k-means clustering adaptive OFDM-IM significantly improves modulation adaptability and throughput, its optimal performance heavily depends on efficient multiple access schemes. Given the anticipated massive connectivity in 6G networks, exploring robust random access methods is essential. This motivates the subsequent investigation into learning-driven Non-Orthogonal Random Access (NORA) techniques, discussed in detail in Chapter 4.

4

Distributed Multi-Agent Reinforcement Learning for Heterogeneous NOMA-ALOHA Systems

4.1 Introduction

Building on the federated k-means clustering adaptive OFDM-IM presented in Chapter 3, this chapter addresses the complementary need for efficient random access strategies. The enhanced adaptive modulation strategies of Chapter 3 necessitate robust random access mechanisms capable of handling the increased demands of simultaneous multi-user transmissions. This Chapter introduces distributed reinforcement learning frameworks specifically designed for Non-Orthogonal Random Access (NORA), which effectively manage massive user connectivity and resource allocation challenges in 6G networks.

In this chapter, distributed Q-Learning algorithms for heterogeneous NOMA-ALOHA systems are proposed to optimize the slot and power level selections of each user without any information sharing between users. In this context, both action collision and fading are considered, and there is no CSI availability at users' transmitters due to the limited spectrum and energy resources of

MTC devices. More importantly, a network in which users have asymmetric conditions in terms of average SNR is considered, which makes it challenging for each user to independently learn a strategy maximizing the throughput. Moreover, many existing papers [85], [87], [93], [84] only consider the total throughput of the system while this chapter additionally measures the fairness between users in terms of the average number of users achieving the minimum desired throughput. Details are introduced in Section 4.2. The main contributions of this chapter are summarized as following.

- A heterogeneous NOMA-ALOHA system where users under different average channel gains send packets by dynamically exploiting one of channel slots and power differences is proposed. To detect and avoid both collisions and fading in the NOMA-ALOHA system, a multi-agent Q-Learning framework is designed. This framework incorporates a new reward function, which influences the exploitation and exploration of action selections.

- Within the multi-agent reinforcement learning framework for NOMA-ALOHA, three algorithms are developed for each user to find a strategy of selecting both channel slots and power levels, towards the enhanced throughput. They are multi-state Q-Learning with state definitions 1 and 2, and confidence-aided Q-Learning. For this, insights into the benefits of multiple state-action values and confidence-aided action values are discussed.

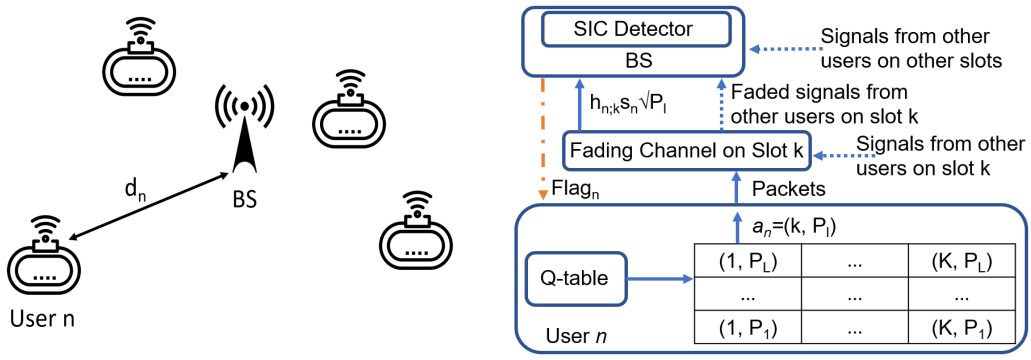
- Through simulative analysis, the impact of hyper parameters such as the numbers of users and slots, as well as heterogeneous average channel gains among users are investigated. In addition, the proposed al-

gorithms are compared to the benchmarks in terms of both the packet throughput and the number of users under the desired performance, over several congestion scenarios. In particular, when the number of users are greater than the number of possible actions, the proposed algorithms are shown to increase the throughput over trials, while the benchmarks suffer from the performance degradation at high congestion level.

- Based on these observations, it is found that under a medium congestion level, the multi-state Q-Learning with state definition 1 may perform the best in terms of average number of users with desired throughput while the confidence-aided algorithm is the best candidate for the system throughput. The algorithm with state definition 2 can be chosen as the best with the consideration of a trade-off between system throughput and fairness, under medium congestion level. When it comes to extreme congestion condition, the confidence-aided algorithm performs best on both system throughput and fairness.

4.2 System Model

Suppose that N users are randomly distributed to transmit packets over K slots to the BS, as shown in Fig.4.1(a). R denotes the radius of the coverage area, $d_n(\leq R)$ denotes the distance between the n-th user and the BS, and $d_n \neq d_m, \forall m \neq n$. Denote by $h_{n;k}$ channel coefficient from user n to the BS over slot $k \in \{1, \cdots, K\}$, where K is the number of slots. Assume that given d_n, each user experiences Rayleigh fading channel, considering $h_{n;k}$



(a) Geographic layout of the NORA system (b) Uplink transmission process of the n-th user

Figure 4.1: System diagram of the Q-Learning NORA system. (a) shows the geographic layout of the users and BS. Provided that a distance between the n-th user and BS is $d_n \in (0, R]$. (b) shows the process of that the n-th user selects a slot-power pair using the learned Q-table, and transmit the packet using the selected slot-power pair. $Flag_n \in \{0, 1\}$ denotes the transmission feedback, $n \in \{1, \ldots, N\}$.

under a complex Gaussian distribution with zero mean and variance $\bar{g}_{n;k}$, i.e., $h_{n;k} \sim \mathcal{CN}(0, \bar{g}_{n;k})$, where $\mathcal{CN}(\cdot)$ denotes complex Gaussian distribution, and $\bar{g}_{n;k} = A_0 d_n^{-\kappa}$, κ is the pathloss exponent and A_0 is a shadowing coefficient. The instantaneous channel gain of the n-th user is $g_{n;k} = |h_{n;k}|^2$, where $g_{n;k} \sim \mathbf{Exp}(\frac{1}{\bar{g}_{n;k}})$, $\mathbf{Exp}(\cdot)$ denotes exponential distribution.

A distributed user randomly selects one out of K slots and delivers its packet over the chosen slot, without interaction among them. In this situation, the users are motivated to find their own strategies of grant-free random access, under heterogeneous condition. Inspired by the concept of GF-NORA [38] [94], it is required to eliminate the signalling overhead and improve the spectrum efficiency such that N users transmit packets with the use of power differences. Details of NORA systems are presented in the following sub-section.

4.2.1 NORA Process

Each user randomly choose an action at each transmission interval without any CSI at the transmitter. The action of the n-th user, $a_n(t) \in \mathcal{A}$, is defined as a combination of choosing channel slot k and transmit power P_l, which is given by

$$a_n(t) = \begin{cases} (0,0) & \text{no transmit at step } t \\ (k, P_l) \end{cases} \tag{4.1}$$

where $k \in \{1, \ldots, K\}$ and $l \in \{1, \ldots, L\}$ are the slot index and the transmit power level index, respectively, L is the number of power levels, and $\mathcal{A}$ is the set of all actions. $P_l \in (0, 1)$ denotes the normalized power level, $P_1 < \ldots < P_L$, $\sum_l P_l = 1$. $\mathbb{1}(\cdot)$ denotes the binary indicator function. The action selection is indicated by

$$Z_{n;k,l} = \mathbb{1}(a_n = (k, P_l)). \tag{4.2}$$

The received signal at the BS on slot k is given by

$$y_k = \sum_{l=1}^{L} \sum_{n=1}^{N} \sqrt{P_l} h_{n;k} S_n Z_{n;k,l} + w_k \tag{4.3}$$

where $w_k \in \mathcal{CN}(0, N_0)$ denotes the additive white gaussian noise (AWGN) on slot k, N_0 is the noise power spectrum density, S_n denotes the modulated symbol. As shown in (4.3), when more than two users randomly compete the same slot k, the NOMA transmission may allow to decode the signals, through successive interference cancellation (SIC) steps.

Given $a_n = (k, P_l)$ from the n-th user, the received signal SINR at the

BS on slot k with power level P_l is given by

$$SINR_{n;k,l} = \frac{P_l g_{n;k}}{\sum_{i=1}^{l-1} \sum_{n'=1,n' \neq n}^{N} P_i g_{n';k} Z_{n';k,i} + N_0}. \tag{4.4}$$

Particularly, the $SINR_{n;k,l}$ will become $SNR_{n;k,l}$ if there is no interference ($\sum_{i=1}^{l-1} \sum_{n'=1}^{N} Z_{n';k,i} = 0$).

The criteria for successful decoding for action (k, P_l) is

Con1) $\sum_{n=1}^{N} Z_{n;k,l'} \leq 1$, for $l' \geq l$ (no action collision)

Con2) $SINR_{n;k,l'} \geq \Gamma \sum_{n=1}^{N} Z_{n;k,l'}$, for $l' \geq l$ (SIC success)

where Γ is the SINR threshold. Assume that packets are successfully decoded only when meeting both $Con1$) and $Con2$). $Con1$) indicates a no-collision event that there are no more than two users choosing the same action. For example, given an action (k, P_l), at most one user (if exist) is allowed to choose this action, which means, if exist, $Con1$ allows only one user choosing $P_{l'}$ for $l' \geq l$ at a given slot k. In other words, since the power domain NOMA technology enables multiple users to simultaneously send their packets through the same RB using different transmit power levels, an action (k, P_l) can be chosen at most by one user. Otherwise, packet decoding is assumed to fail due to random collision (more than one user chooses the same power level at same RB) because the capturing effect is not considered in this work. $Con2$) represents an event associated with channel fading. That is, packet decoding can be successful only if the SINR after the SIC is greater than or equal to the desired threshold. In addition, the packets transmission will fail if the decoding of any higher power level signal at the same RB fail since the SIC decoding order is from high power level to low power level. The transmission feedback of the n-th user at time step t is indicated by $Flag_n(t)$, which is

given by

$$Flag_n(t) = \begin{cases} 1 & \text{, for successful decoding at step } t \\ 0 & \text{, for failure.} \end{cases} \tag{4.5}$$

4.2.2 Problem formulation

A distributed grant-free NORA algorithms is developed in order to maximize the system throughput while maintaining the fairness among users. Denote ASR_n the average success rate (ASR) [15] of the n-th user, which is viewed as

$$ASR_n = \mathbb{E}[Flag_n]. \tag{4.6}$$

where $\mathbb{E}[\cdot]$ denotes expectation. ASR measures the average number of packets successfully conveyed by the n-th user for given users' strategies. In addition, the algorithm design needs to monitor the fairness among users such that each intends to make ASR_n at minimum the throughput threshold. Based on these, the performance of the algorithms is analysed through two metrics: average number of users with desired ASR and average packet throughput. For the case of fairness-sensitive systems, the fairness is measured by counting the number of users whose ASR_n is greater than the throughput threshold. For the case of fairness-tolerant systems, the average packet throughput introduced by [86] is used to measure the system throughput only with no fairness. They are defined as:

- *Average number of users per slot with desired ASR:* Given N users and K slots, the average number of users per slot with $ASR > ASR_0$ is

calculated by

$$N_{users} = \frac{1}{K}\sum_{n=1}^{N} \mathbb{1}(ASR_n > ASR_0). \tag{4.7}$$

Each indicator function in (4.7) can be approximated by the sigmoid function, $(1 + e^{-\theta(ASR_n - ASR_0)})^{-1}$, where the steepness parameter θ is chosen to be sufficiently large to mimic the sharp transition of the indicator function [95]. Note the strategies of all the users change over steps, ASR_n is a random variable which is effected by the users' strategies. By taking expectation, it becomes

$$\mathbb{E}[N_{users}] \approx \mathbb{E}[\frac{1}{K}\sum_{n=1}^{N} \frac{1}{1 + e^{-\theta(ASR_n - ASR_0)}}] \tag{4.8}$$

where θ is the slop parameter of the sigmoid function.

- *Average packet throughput per slot:* Given N users and K slots, the average packet throughput per slot is

$$\mathbb{E}[N_{packet}] = \frac{1}{K}\sum_{n=1}^{N} \mathbb{E}[ASR_n]. \tag{4.9}$$

Notice that the probability density function (PDF) of ASR_n is not trackable due to the time-varying strategies of other users. Denote $\mathbf{x}_n \in \mathcal{X}$ the mixed strategy of the n-th user,

$$\mathbf{x}_n = [x_{n;0,0}\ x_{n;1,1} \cdots x_{n;K,1} \cdots x_{n;k,l} \cdots x_{n;1,L} \cdots x_{n;K,L}]^T \tag{4.10}$$

where $x_{n;k,l}$ denotes the probability that the n-th user takes action (k, P_l),

and

$$x_{n;0,0} = 1 - \sum_{k=1}^{K} \sum_{l=1}^{L} x_{n;k,l}. \tag{4.11}$$

The mixed strategies of the other users are given by

$$\mathbf{x}_{-n} = [\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N]^T. \tag{4.12}$$

With given $\mathbf{x}_{-n}$, $ASR_{n;k,l}$ denotes the ASR of the n-th user being associated only to action (k, P_l). The analytical expression for $ASR_{n;k,l}$ is derived in Appendix A. The expectation of the average number of users per slot with $ASR > 0.1$ is given by

$$\mathbb{E}[N_{users}] \geq \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{x_{n;k,l}}{1 + e^{-\theta(\mathbb{E}[ASR_{n;k,l}] - ASR_0)}}. \tag{4.13}$$

The expectation of the average packet throughput is given by

$$\mathbb{E}[N_{packet}] \geq \frac{1}{K} \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{L} x_{n;k,l} \mathbb{E}[ASR_{n;k,l}]. \tag{4.14}$$

Due to the fact that $\mathbf{x}_n$ is influenced by $\mathbf{x}_{-n}$, which are unknown to the n-th user, each user is desired to learn from its own trials in making actions in a distributed manner. The optimization problem is given by

$$\max_{\mathbf{x}_n} ASR_n, \ \forall n \in \{1, \cdots, N\}. \tag{4.15}$$

In order to enhance the average number of users with desired ASR and average packet throughput, distributed Q-Learning aided NORA algorithms optimizing $\mathbf{x}_n$, the action strategy of individual user, are investigated in Sec-

tion 4.3.

4.3 Proposed Reinforcement Learning Algorithms

The slot and power level selection task is modeled as a MDP in this chapter. In a MDP, the agent interacts with the environment by taking action according to its state and receiving reward at each step. One of the widely used reinforcement learning algorithms resolving MDP is Q-Learning [24]. Although deep reinforcement learning algorithms, such as DDQN and actor-critic [96], [97], are more efficient than Q-Learning algorithms, those algorithms are too complex to be implemented on resource limited MTC devices. Consequently, Q-Learning is a competitive candidate in this application scenario. The adopted Q-Learning model in this chapter considers each user as an agent, and individual users select one of the actions (k, P_l) according to the action value function, which is the Q table in this chapter. The Q-table of each user is updated following (4.16).

$$Q_n(s_n(t), a_n(t)) \leftarrow Q_n(s_n(t), a_n(t)) + \alpha\delta_n(t) \tag{4.16}$$

where $s_n(t)$, $a_n(t)$ and α denote the state, action and learning rate, respectively. $\delta_n(t)$ is the temporal difference (TD) error,

$$\begin{aligned}\delta_n(t) &= G_n(t) - Q_n(s_n(t), a_n(t)) \\ &= R_n(t) + \gamma \max_a Q_n(s_n(t+1), a) - Q_n(s_n(t), a_n(t))\end{aligned} \tag{4.17}$$

where $R_n(t)$ and γ denote the immediate reward and discount factor, respectively.

Three state definition and one reward definition are proposed in this chapter, which are used to develop three novel Q-Learning algorithms for the NORA problem formulated in Section 4.2.2. The Q-table of each user is designed to be updated in every coherence NORA step to adapt to the dynamic environment. The algorithms demonstrate the independent behaviour of each user and thus are identical for all users.

4.3.1 Reward

In a number of papers [86], [15], [81], [93], the reward of fail transmission is set to -1 or 0, which is straightforward and can result in a fast convergence in their system model. However, in the proposed system model, the difference between the average channel gains of individual users are relatively large, which may increase the probability of failed decoding caused by interferences from other users. In this case, part of the Q-values will be underestimated if the rewards for successful transmission and fail transmission have the same absolute value. The absolute value of the reward for failed transmission, μ, is made smaller than that for successful transmission. The reward function of the MDPs is considered as

$$R_n(t) \triangleq \begin{cases} 1 & \text{if } Flag_n(t) = 1 \\ 0 & \text{if } a_n(t) = (0,0) \\ -\mu & \text{if } a_n(t) \neq (0,0) \text{ and } Flag_n(t) = 0 \end{cases} \tag{4.18}$$

where $\mu > 0$, such that

$$\mathbb{E}[R_n] = \mathbb{E}[ASR_n] - \mu(1 - \mathbb{E}[ASR_n]). \tag{4.19}$$

It can be seen from (4.19) that the mean reward received by the n-th user, $\mathbb{E}[R_n]$, increases with the ASR_n so that the algorithm maximizing the reward enhances average number of users with desired ASR and average packet throughput. Using (4.18), the learning models and algorithms are presented below.

Algorithm 4 Q-Learning Assisted NORA with State Def.1

Output: Updated Q-values $Q_n(s, a), \forall s, a$
Initialization: Initialize all the Q-values with zeros, $t = 0$, $a_n(0) = (0, 0)$, $s_n(1) = [(0, 0), 0, 0]$, $T = 5000$, $\alpha = 0.1$, $\gamma = 0.05$
while $t < T$ **do**
  $t \leftarrow t + 1$
  $a_n(t) \leftarrow \arg\max_a Q_n(s_n(t), a)$
  if $a_n(t) = (0, 0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t) = 1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $\sigma_{n;ch}(t) \leftarrow \mathbb{1}(a_n(t)_1 \neq a(t-1)_1)$
  $\sigma_{n;pow}(t) \leftarrow \mathbb{1}(a_n(t)_2 \neq a(t-1)_2)$
  $s_n(t+1) \leftarrow (a_n(t), \sigma_{n;ch}(t), \sigma_{n;pow}(t))$
  $Q_n(s_n(t), a_n(t)) \leftarrow Q_n(s_n(t), a_n(t)) + \alpha\delta_n(t)$
end while

4.3.2 State Definition 1

Algorithm 5 Q-Learning Assisted NORA with State Def.2

Output: Updated Q-values $Q_n(s,a), \forall s,a$
Initialization: Initialize all the Q-values with zeros, $t=0$, $a_n(0)=(0,0)$, $s_n(1)=(0,0)$, $T=5000$, $\alpha=0.1$, $\gamma=0.05$
while $t<T$ **do**
  $t \leftarrow t+1$
  $a_n(t) \leftarrow \arg\max_a Q_n(s_n(t),a)$
  if $a_n(t)=(0,0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t)=1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $s_n(t+1) \leftarrow a_n(t)$
  $Q_n(s_n(t),a_n(t)) \leftarrow Q_n(s_n(t),a_n(t)) + \alpha\delta_n(t)$
end while

In the first learning model, the process is modelled as a MDP. The state of the MDP at step t is defined as the action taken by the individual user at last step $t-1$, and two indicators on whether the slot and power level selection at last step are the same as those at $t-2$.

$$s_n(t) \triangleq (a_n(t-1), \sigma_{n;ch}(t-1), \sigma_{n;pow}(t-1)) \tag{4.20}$$

where

$$\sigma_{n;ch}(t) \triangleq \mathbb{1}(a_n(t)_1 \neq a_n(t-1)_1) \tag{4.21}$$

$$\sigma_{n;pow}(t) \triangleq \mathbb{1}(a_n(t)_2 \neq a_n(t-1)_2) \tag{4.22}$$

and the subscript j in $a_n(t)_j$ is the index for j-th element of $a_n(t)$. By this definition, the users can leverage the history information to find a dynamic strategy rather than a static action selection. This allows each slot-power pair to alternately serve multiple users. Algorithm 4 illustrates the multi-state Q-Learning assisted NORA algorithm with the state definition 1.

4.3.3 State Definition 2

Algorithm 6 Confidence-aided Q-Learning Assisted NORA

Output: Updated Q-values $Q_n(a), \forall a$
Initialization: Initialize all the Q-values with zeros, $t = 0$, $T = 5000$, $\alpha = 0.1$, $\gamma = 0$, $W_n(a) = 0, \forall a \in \mathcal{A}$
while $t < T$ **do**
  $t \leftarrow t + 1$
  $a_n(t) \leftarrow \arg\max_a (Q_n(a) + \sqrt{\frac{2 \ln t}{W_n(a)}})$
  $W_n(a_n(t)) \leftarrow W_n(a_n(t)) + 1$
  if $a_n(t) = (0, 0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t) = 1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $Q_n(a_n(t)) \leftarrow Q_n(a_n(t)) + \alpha \delta_n(t)$
end while

Although Algorithm 4 allows the users to leverage history information, the large state space might lead to low convergence speed and insufficient exploration. To address this issue, another state definition with a smaller state space is proposed. This only consists of the action taken at the last

step $t-1$.

$$s_n(t) \triangleq a_n(t-1). \tag{4.23}$$

Algorithm 5 illustrates the multi-state Q-Learning assisted NOMA-ALOHA algorithm with the state definition 2.

4.3.4 Stateless with confidence-aided actions

In this learning model, the slot and power level selecting process is modelled as a multi-arm bandit problem. Since the strategies of users keep changing at the early stage of the learning process and are unknown to each other, high quality exploration at the early stage are crucial for potentially converging to near optimal strategy. However, the widely used ϵ-greedy results in a linear increase on an accumulated error between optimal action values and estimated action values. To address this issue, the confidence-aided algorithm [98] was known to provide logarithmic increase on the accumulated error. This confidence concept is motivated for the proposed algorithm to better balance exploration and exploitation.

In the confidence-aided algorithm, the agent maintains a Q-table consisting of the estimated reward of each action, and a counter $W_t(a)$ recording how many times action a has been chosen. According to Hoeffding's inequality, the probability that the true Q-value exceeds its upper confidence bound is

$$\Pr\Bigg(Q^*(a) > Q(a) + U_t(a)\Bigg) \leq e^{-2W_t(a)U_t^2(a)} \tag{4.24}$$

where $U_t(a)$ denotes the difference between the estimated Q-value and its upper confidence bound. Since the probability in (4.24) is desired to converge

to 0 (confidence level equals to 100%) as $t \to +\infty$, the right hand side of (4.24) is designed as equalling to t^{-4}. Consequently, the $U_t(a)$ is given by

$$U_t(a) = \sqrt{\frac{2 \ln t}{W_t(a)}}. \tag{4.25}$$

The action policy of individual users at every transmission interval are given by

$$a_n(t) = \arg\max_a (Q(a) + U_t(a)). \tag{4.26}$$

By adopting this action policy, the agent selects the action with highest upper confidence bound under a dedicate confidence level of the moment, which makes the agent always take the action with biggest potentials, helping the agent to explore the unknown environment at the early stage of the learning process. Moreover, since $\lim_{t \to +\infty} U_t(a) = 0$, the confidence-aided algorithm actually becomes a greedy action policy when $t \to +\infty$. After each packet transmission, each user updates its Q-table according to the transmission result. Algorithm 6 illustrates the confidence-aided NOMA-ALOHA algorithm.

The three proposed algorithms have different properties (e.g., convergence speed, scalability, complexity), suitable for meeting different requirements of various application scenarios. To improve fairness between users, Algorithm 4 with the largest state space is able to produce diversity among users. However, it may not be suitable for applications where the devices have limited memory. Moreover, it may suffer slower convergence speed compared with the other two algorithms. Algorithm 6 adopts an unique exploration strategy aimed to improve the exploration quality, which makes it a potential candi-

date under high congestion traffics. Algorithm 5 is a multi-state algorithm with a simplified state space, expected to achieve trade-off between fairness and system throughput.

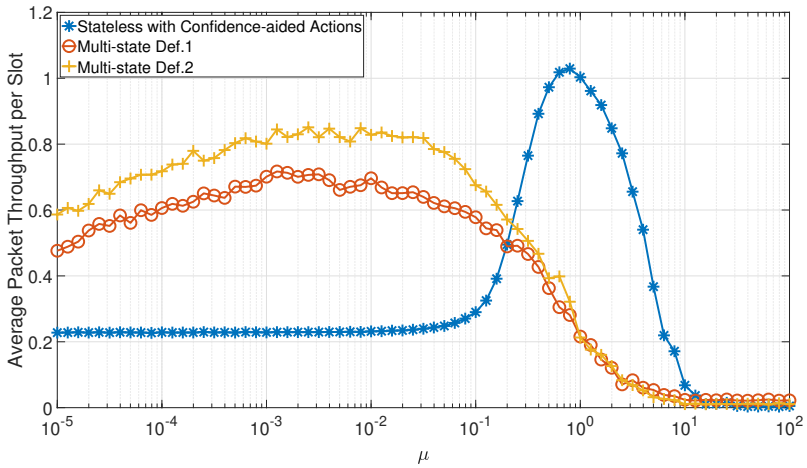
4.4 Simulations and Discussions

The simulation results and numerical analysis of the three proposed algorithms in the distributed heterogeneous NOMA-ALOHA system are presented in this section. Four benchmark schemes are adopted, which are

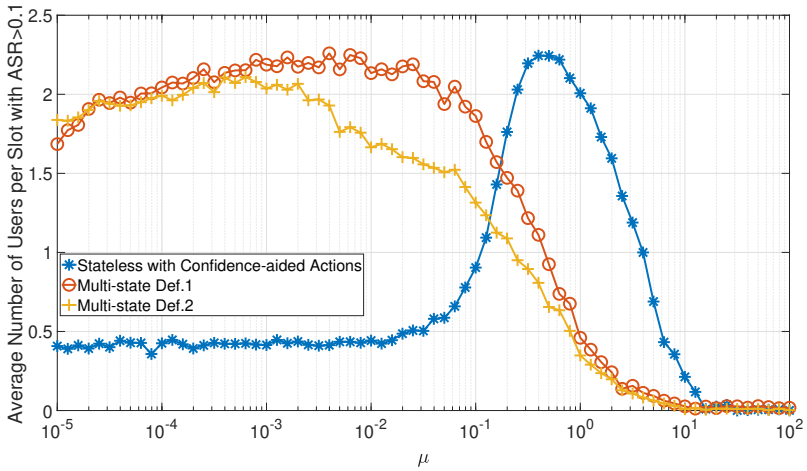
- Slotted NORA [38].
- RL-NORA Acceleration-GA [15].
- RL-NORA Acceleration-ϵ-GA [15].
- woSDC-BAP-QL [86].

For all simulations, $A_0 = 1$, $\kappa = 3$, $\Gamma = -3dB$, $L = 3$ with $P_1 = 0.04$, $P_2 = 0.16$, $P_3 = 0.8$. It is assumed that each successful slot-power pair can transmit one packet. The focus is on five simulation scenarios

- The sensitivity of the proposed algorithms with different μ, the absolute value of the reward for failed transmission, in terms of average packet throughput and average number of users with $ASR > 0.1$.
- The convergence properties.
- Average packet throughput and average number of users with $ASR > 0.1$ for different number of users of the proposed algorithms and the benchmarks.



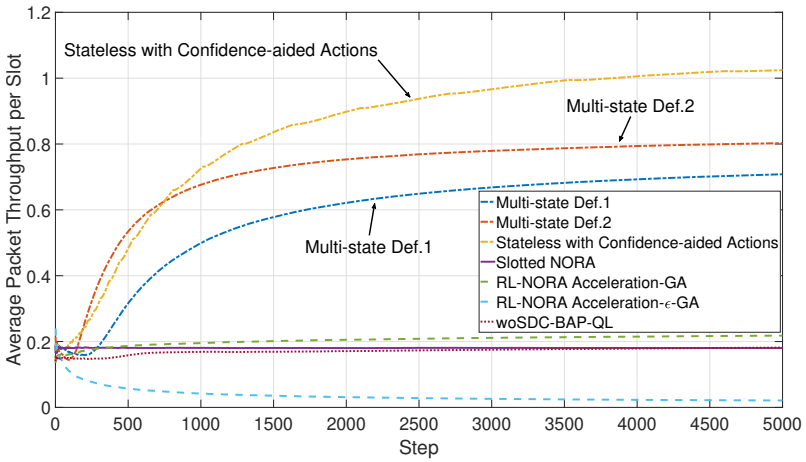
(a) Average packet throughput per slot versus μ



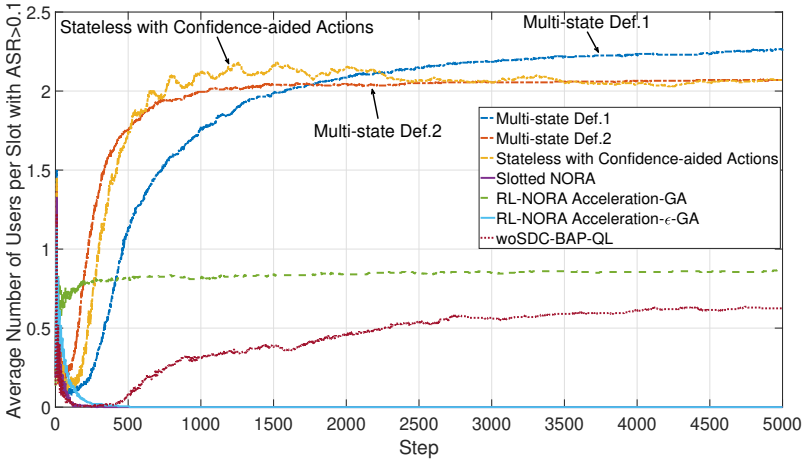
(b) Average number of users per slot with $ASR > 0.1$ versus μ

Figure 4.2: Effect of μ on average packet throughput (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, when $N = 24$, $K = 4$, $L = 3$.

- The sensitivity of the proposed algorithms and the benchmarks with different number of slots.
- The sensitivity of the proposed algorithms and the benchmarks with different minimum channel gain.



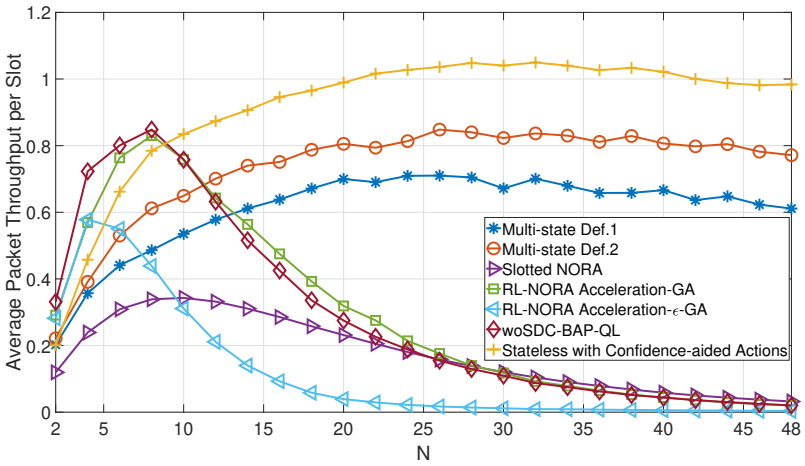
(a) Average packet throughput per slot versus step



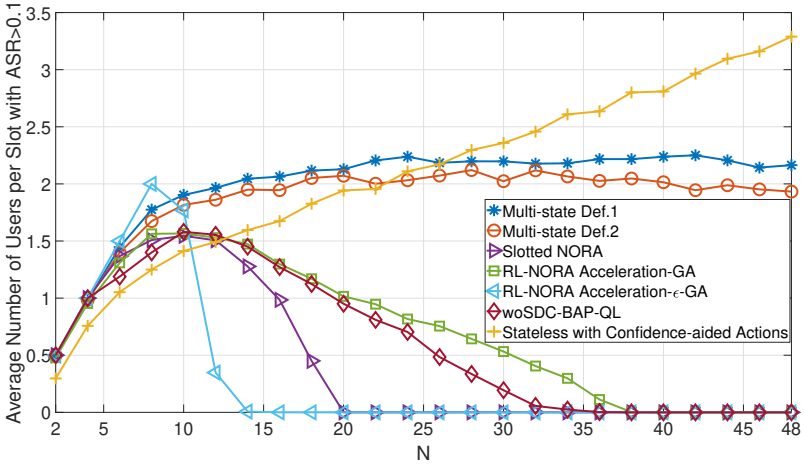
(b) Average number of users per slot with $ASR > 0.1$ versus step

Figure 4.3: Performances related to average packet throughput (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmark schemes are depicted, when $N = 24$, $K = 4$, $L = 3$.

In Fig.4.2, the average packet throughput and average number of users with desired ASR of the three proposed algorithms are presented in an over-distributed case ($N = 2LK$). Fig.4.2(a) illustrates that the confidence-aided algorithm performs best in terms of average packet throughput, and it reaches its best performance around 1.029 packets/slot when $\mu = 10^{-0.1}$. This is due



(a) Average packet throughput per slot versus N



(b) Average number of users per slot with $ASR > 0.1$ versus N

Figure 4.4: Effect of N on average packet throughput (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $K = 4$, $L = 3$.

to the agents not being able to balance the exploration and exploitation well when the ratio of rewards in fail and success actions is not set properly, which leads to a convergence to local optimal rather than global optimal. The two multi-state algorithms reach their best performance when $10^{-3} < \mu < 10^{-1.9}$.

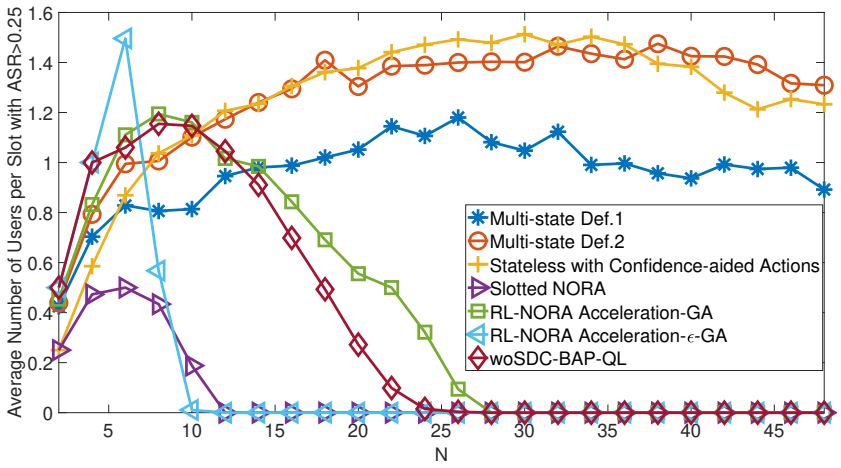


Figure 4.5: Average number of users per slot with $ASR > 0.25$ versus N. (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $K = 4$, $L = 3$.

Fig.4.2(b) indicates that the Multi-state Def.1 achieves the highest average number of users with desired ASR when $\mu = 10^{-2.4}$ while the confidence-aided algorithm and Multi-state Def.2 reaches their best performance, which is slightly lower than the Multi-state Def.1, when $\mu = 10^{-0.4}$ and $\mu = 10^{-3.2}$, respectively.

The convergence behaviour of the proposed algorithms are shown in Fig.4.3. The average number of users per slot have a negative trend in the first iterations. This is because the number of steps is very small so that the ASR hasn't been accurate enough to approximate the real value (much higher than the real value). The Multi-state Def.2 has the fastest convergence speed, and Multi-state Def.1 is the slowest. The confidence-aided algorithm converges just slightly slower than the Multi-state Def.2 while it achieves higher average packet throughput.

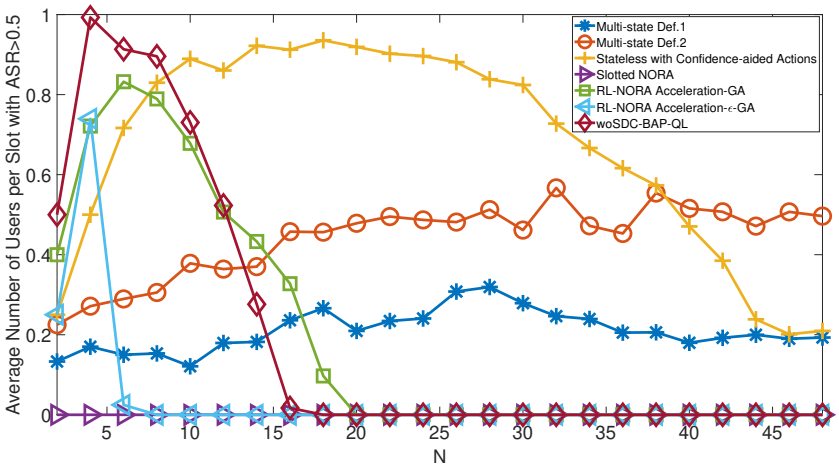


Figure 4.6: Average number of users per slot with $ASR > 0.5$ versus N. (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $K = 4$, $L = 3$.

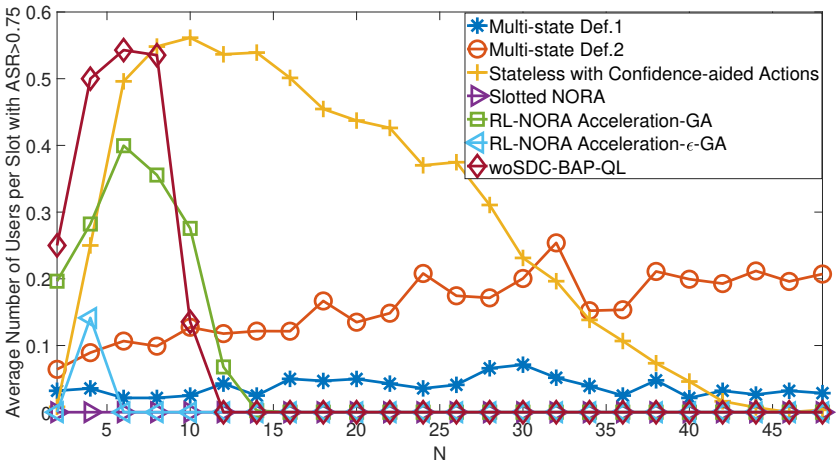
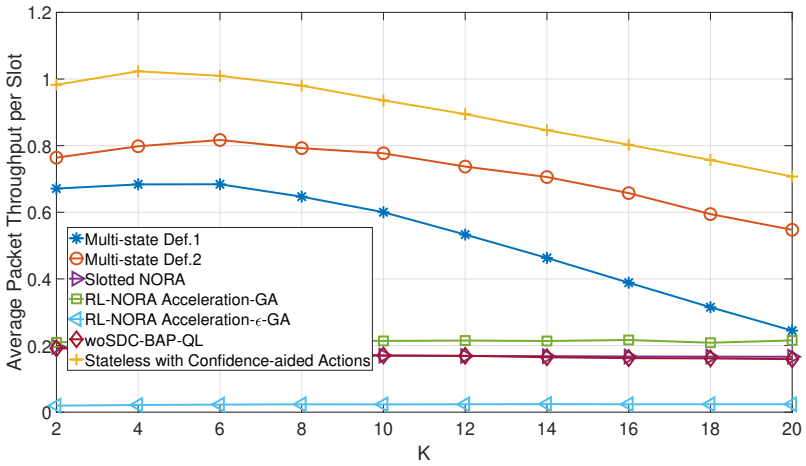
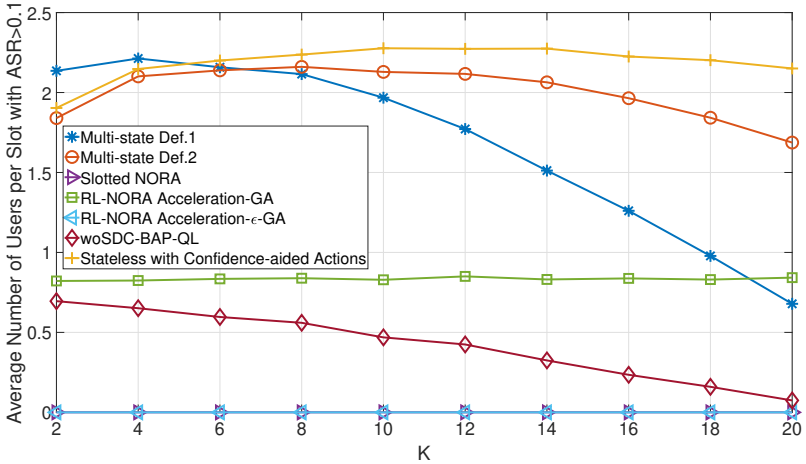


Figure 4.7: Average number of users per slot with $ASR > 0.75$ versus N. (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $K = 4$, $L = 3$.

In Fig.4.4, the average packet throughput and average number of users with desired ASR for different number of users of the three proposed algorithms and the benchmark schemes are compared. The performance of all



(a) Average packet throughput per slot versus K



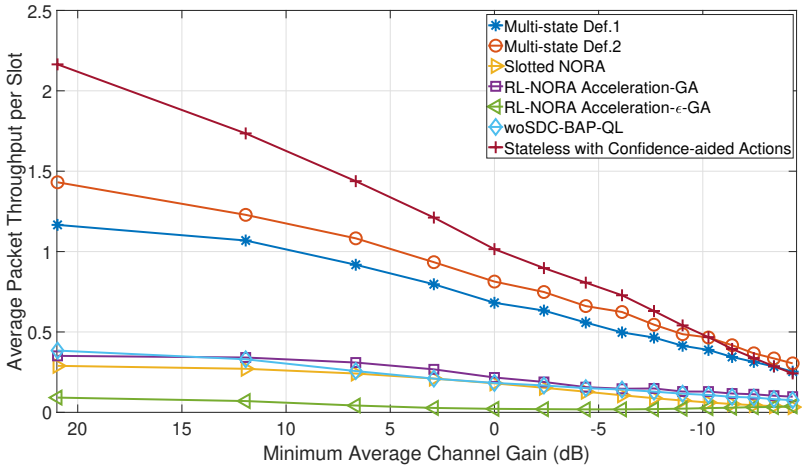
(b) Average number of users per slot with $ASR > 0.1$ versus K

Figure 4.8: Effect of K on average packet throughput (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $L = 3$, $N = 2LK$.

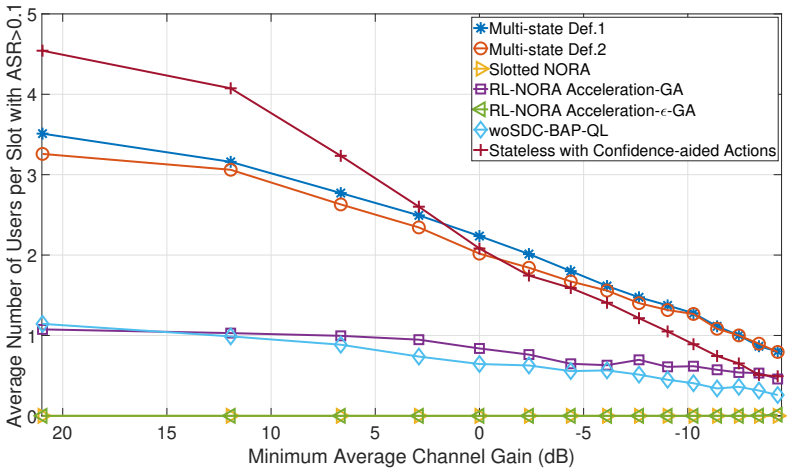
the schemes runs concave as N increases. This is because the number of users is smaller than the maximum capacity of the schemes at the beginning so that N is the restriction of the performance. The benchmarks are stateless algorithms, which make the users choose one of the actions under low

collisions. This is efficient for the users to ultimately find different actions with no collision when the number of users is smaller than the number of actions. However, when the number of users grows, there are more collisions. In this case, the proposed algorithms allow the users to better find dynamic strategies with the use of state space. Note that the benchmark schemes have a performance degradation when $N > LK$ while the two proposed multi-state algorithms can maintain their performance. It can be seen that the confidence-aided algorithm has the best performance compared to other algorithms in terms of average packet throughput for all the values of N. The Multi-state Def.1 is the best in terms of average number of users with desired ASR when $LK < N < 2LK$. It is worth noting that the Multi-state Def.2 performs in-between Multi-state Def.1 and the confidence-aided algorithm. Note that the average number of users with desired ASR of the confidence-aided algorithm keep increasing with the number of users even when $N > LK$, and outperforms the two multi-state algorithms when $N > 2LK$. Fig.4.5, Fig.4.6 and Fig.4.7 show the average number of users per slot with the ASR thresholds equal to 0.25, 0.5 and 0.75, respectively, which indicates the proposed algorithms outperforms the benchmarks in high congestion traffic even in higher ASR threshold settings.

The effect of number of slots on the performance of the proposed algorithms and the benchmark schemes are shown in Fig.4.8. The average packet throughput and average number of users with desired ASR of the proposed algorithms are concave with K. Particularly, the performance degrades seriously when K become relatively large, increasing the size of the Q-table. This leads to an insufficient exploration when using a lookup table method.



(a) Average packet throughput per slot versus minimum average channel gain



(b) Average number of users per slot with $ASR > 0.1$ versus minimum average channel gain

Figure 4.9: Effect of minimum average channel gain on average packet throughput (a) and average number of users with desired ASR (b). The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $N = 24$, $K = 4$, $L = 3$.

As shown in Table 4.1, the Q-table size of the two proposed multi-state algorithms increase much faster than the confidence-aided algorithm with K. Besides, the computation complexity of the Multi-state Def.2 is nearly one quarter of the complexity of Multi-state Def.1, which explained the reason

Table 4.1: Complexity Comparison

Algorithm	Complexity
Confidence-aided stateless	$\mathcal{O}(2(KL+1))$
Multi-state Def.1	$\mathcal{O}((KL+1)(4KL+2))$
Multi-state Def.2	$\mathcal{O}((KL+1)(KL+1))$
RL-NORA [15]	$\mathcal{O}(KL+1)$
woSDC-BAP-QL [86]	$\mathcal{O}(KL)$

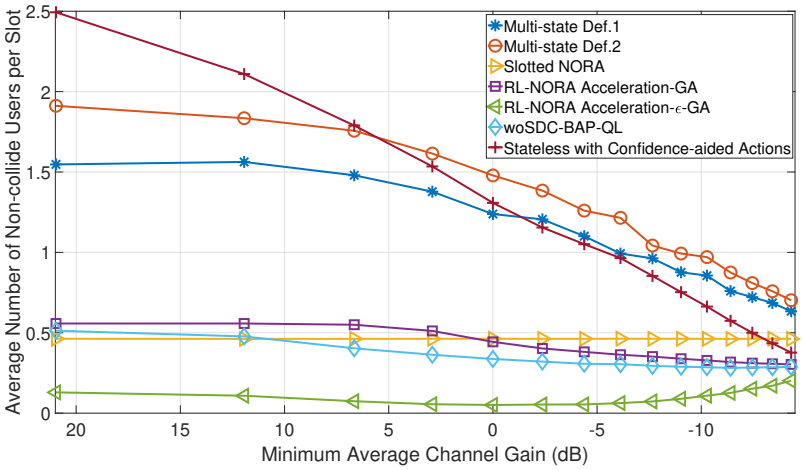


Figure 4.10: Number of non-collide users versus minimum average channel gain. The proposed methods are Multi-state Def.1, Multi-state Def.2, and Stateless with confidence-aided actions, and for comparison, the benchmarks, when $N = 24$, $K = 4$, $L = 3$.

that Multi-state Def.1 degrades earlier than Multi-state Def.2.

The effect of the minimum average channel gain on the performance of the proposed algorithms and the benchmark schemes are shown in Fig.4.9. The reduction of average channel gain leads to lower probabilities of successful decoding. Accordingly, with the increased difference between the average channel gains of the users, it can be seen that the two multi-state algorithms degrade less on both average packet throughput and average number of users with desired ASR than the confidence-aided algorithm. This means

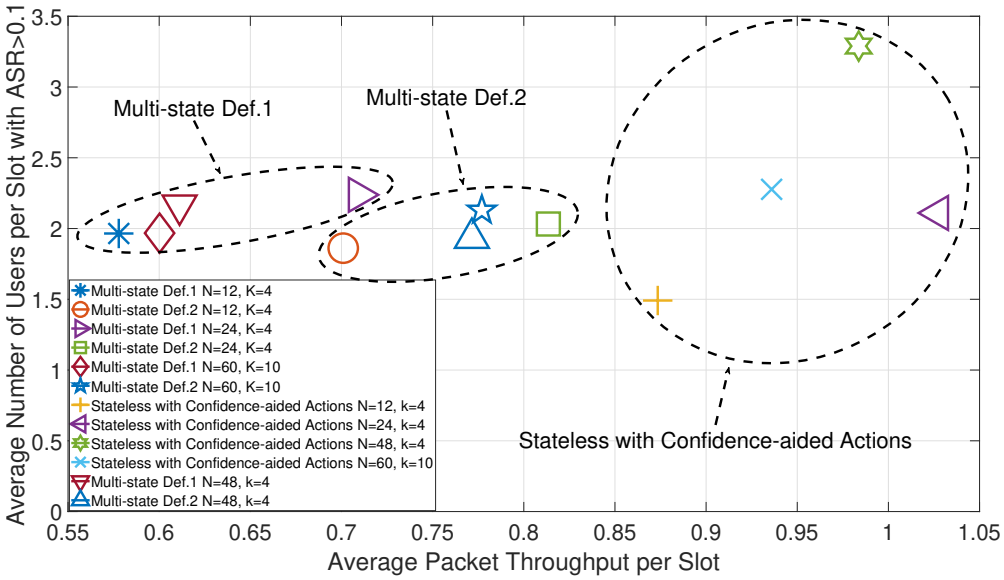


Figure 4.11: Trade-off between average packet throughput and average number of users with desired ASR, when (1) $N = 12$, $K = 4$, $L = 3$, (2) $N = 24$, $K = 4$, $L = 3$, (3) $N = 48$, $K = 4$, $L = 3$ and (4) $N = 60$, $K = 10$, $L = 3$.

the multi-state algorithms are more robust to the heterogeneity of the users' channel conditions. Fig.4.10 shows the reason for the above observation, that is, the two multi-state algorithms can maintain the number of non-collide users better than the confidence-aided algorithm. When the minimum average channel gain is lower than $-1dB$, the confidence-aided algorithm no longer has advantage over the multi-state algorithms.

The trade-offs of the proposed algorithms are compared through their average packet throughput and average number of users with desired ASR in Fig.4.11. The two proposed multi-state algorithms are preferable for the applications with medium amount of users ($LK < N < 2LK$). In particular, the Multi-state Def.2 suits to the applications in which both the system throughput and the number of users achieving target QoS are important with limited users' computation resources. Whereas Multi-state Def.1 is more

suitable when the number of users achieving target QoS is the main performance target. The confidence-aided algorithm is preferable for applications with massive number of users ($N > 2LK$), or when the users' computation resources are extremely limited.

4.5 Summary

In this chapter, a distributed reinforcement learning framework for joint slot and power level selecting problem in heterogeneous NOMA-ALOHA systems is proposed. Two multi-state Q-Learning algorithms and a confidence-aided algorithm are developed to find the action selection strategies in a distributed manner. Simulation results show that the proposed algorithms outperform the benchmarks in terms of system throughput and fairness in high congestion traffics, which is crucial for the massive connectivities in 6G. Additionally, the three algorithms have advantages in terms of fairness, system throughput and robustness to extreme congestion condition. Thanks for the model-free distributed learning framework and the NOMA-ALOHA procedure, the proposed schemes are capable of enabling efficient RA for resource limited MTC networks in heterogeneous environment.

The reinforcement learning (RL) algorithms proposed in this chapter effectively address the joint slot and power level selection in heterogeneous NOMA-ALOHA systems. However, as the complexity and scale of problems in practical 6G scenarios significantly increase, conventional RL approaches, such as Q-learning, face limitations due to their dependence on tabular representations of state-action spaces, making them infeasible for large-scale systems. To overcome these constraints, Deep Reinforcement Learning (DRL) is

introduced in Chapter 5, offering substantial benefits over traditional RL by utilizing neural networks to approximate the Q-value function. This transition allows DRL to efficiently handle larger, more complex state-action spaces and continuous input data, which is essential for practical deployments in multi-AP machine-type communication systems. Consequently, DRL provides the necessary scalability and flexibility to adapt to diverse, high-traffic network conditions, demonstrating its critical role in achieving intelligent and efficient random access strategies in future 6G ecosystems.

5 Distributed Multi-Agent Deep Reinforcement Learning for Multi-Point NORA Systems

5.1 Introduction

This chapter builds directly upon the single access point (AP) system model and reinforcement learning (RL)-based algorithms discussed in Chapter 4, by extending the analysis to a more complex multiple AP environment. Specifically, this chapter introduces distributed deep reinforcement learning (DRL) algorithms as opposed to traditional RL methods employed previously. This progression enhances flexibility by allowing users to autonomously select their preferred APs, thereby eliminating the need for explicit, controlled user clustering. Additionally, the shift from lookup-table-based RL to DRL methodologies substantially improves the scalability of the proposed framework, effectively handling larger-scale scenarios characteristic of practical beyond-6G networks.

This chapter proposes distributed deep Q-learning algorithms for multi-agent multi-point NORA systems to optimize the selection of APs, time slots, and power levels for each user without requiring any information sharing between users. Both action collision and fading are considered, and users' transmitters operate without CSI due to the limited spectrum and energy

resources of MTC devices. Notably, the presence of multiple APs within a network allows users to transmit data packets to any AP, which makes it challenging for each user to independently learn a strategy that maximizes throughput, particularly when APs are located in different positions, resulting in varying channel gains. Unlike many existing works that only consider the total throughput of the system, this study also measures fairness among users by considering the number of users under the time-averaged age of information (AAOI) constraint. Further details are in Section 5.2. Main contributions of this chapter are summarized as following:

- A multi-agent multi-point NORA system is proposed, where users dynamically transmit packets by exploiting one of APs, channel slots and power levels. To address and reduce both collisions and fading in the NOMA-ALOHA system, a multi-agent deep Q-learning framework is developed. This framework incorporates a novel state definition leveraging the age of information (AOI) and a reward function that dynamically adjusts the exploitation and exploration of action selections based on the user's AOI.

- Given the multi-agent deep Q-learning framework for multi-point NORA, two algorithms are developed for each user to find a strategy of selecting APs, channel slots and power levels, towards the enhanced throughput. They are double deep Q-network (DDQN) assisted multi-point NORA and dueling double deep Q-network (Dueling DDQN) assisted multi-point NORA. For this, insights into the benefits of AOI-assisted state and dueling network are discussed. Additionally, the choice of different type of output layers and action policies are illustrated.

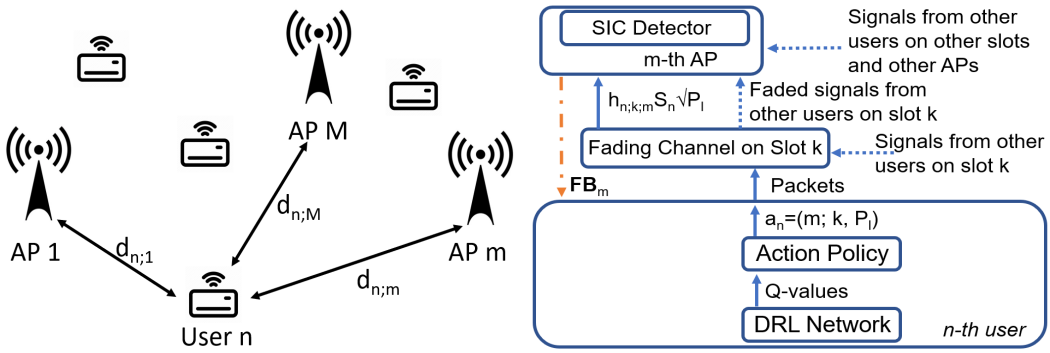
- Through simulative analysis, the impact of hyper parameters such as AOI truncation threshold and Fair Policy threshold, as well as the number of APs and users are investigated. In addition, the proposed algorithms are compared to the benchmarks in terms of both the system packet throughput and the number of users under the AAOI constraint, over several scenarios. In particular, as the number of APs increases, the proposed algorithms are shown to outperform the benchmarks that suffer from the performance degradation at large size of problems.

- Simulative and numerical results clearly show that under a medium congestion level, the Dueling DDQN algorithm performs the best in terms of the number of users under the AAOI constraint while the DDQN algorithm is the best candidate for the system packet throughput. When it comes to extreme congestion condition, the Dueling DDQN algorithm performs best on both system throughput and fairness.

The rest of the chapter is structured such that Section 5.2 introduces the system model of the proposed multi-agent multi-point NORA network. Section 5.3 provides the deep Q network learning model and algorithms. Section 5.4 presents the simulation results and discussions, followed by the conclusion in Section 5.5.

5.2 System Model

Consider a scenario in which N users are randomly dispersed across a region. The users become active with activation probability λ_a and transmit packets over MK slots to M access points (APs) connected to a core network, as



(a) Geographic layout of the NORA system (b) Uplink transmission process of the n-th user

Figure 5.1: System diagram of the DRL NORA system. (a) shows a example of the geographic layout of the users and the APs. $d_{n;m} \in (0, R]$ denotes the distance between the n-th user and the m-th AP. (b) shows the process of that the n-th user selects a AP-slot-power pair using the learned DRL network, and transmit the packet using the selected AP-slot-power pair. $\mathbf{FB}_m$ is the transmission feedback for $n \in \{1, \ldots, N\}$, $m \in \{1, \ldots, M\}$.

depicted in Fig. 5.1(a). Each AP has K slots, and slots used by different APs are orthogonal. The coverage area of the AP group has a radius R, and $d_{n;m} (\leq R)$ indicates the distance between the n-th user and the m-th AP, where $m \in \{1, \ldots, M\}$. For $i \neq j$, it is assumed that $d_{i;m} \neq d_{j;m}$. The channel coefficient from the n-th user to the m-th AP on slot k is denoted as $h_{n;k;m}$, where $k \in \{1, \ldots, K\}$. Under the assumption of a Rayleigh fading channel based on $d_{n;m}$, the coefficient $h_{n;k;m}$ follows a complex Gaussian distribution with a mean of zero and variance $\bar{g}_{n;k;m}$, such that $h_{n;k;m} \sim \mathcal{CN}(0, \bar{g}_{n;k;m})$. Here, $\mathcal{CN}(\cdot)$ represents a complex Gaussian distribution, and the variance is given by $\bar{g}_{n;k;m} = A_0 d_{n;m}^{-\kappa}$, where κ is the path-loss exponent and A_0 is the shadowing coefficient. The instantaneous channel gain of the n-th user, $g_{n;k;m} = |h_{n;k;m}|^2$, follows an exponential distribution $g_{n;k;m} \sim \mathbf{Exp}\left(\frac{1}{\bar{g}_{n;k;m}}\right)$.

Each user independently selects one of the K slots of one of the M APs

for packet transmission, without requiring coordination or interaction with other users. In traditional cellular networks, users are assigned to specific cells with a single AP based on their locations. However, in this study, each user not only chooses a time slot and transmission power but also selects an AP for transmission. Users dynamically cluster themselves into different APs. Within this framework, users must devise grant-free random access strategies to adapt to the heterogeneous conditions across varying users and APs. Inspired by the principles of GF-NORA [38] [94], the approach aims to minimize signaling overhead and improve spectral efficiency by leveraging power-domain differences to allow N users to transmit packets concurrently. Further details regarding the NORA system are discussed in the subsequent subsection.

5.2.1 NORA Process

At each transmission interval, users select their actions randomly without relying on any channel state information (CSI) at the transmitter. The action of the n-th user, $\mathbf{a}_n(t)$, is defined as a combination of selecting an access point m, a channel slot k, and a transmission power level P_l, as expressed by

$$\mathbf{a}_n(t) = \begin{cases} (0,0,0) & \text{no transmission at step } t \\ (m,k,P_l) & \end{cases} \tag{5.1}$$

where $m \in \{1,\ldots,M\}$, $k \in \{1,\ldots,K\}$, and $l \in \{1,\ldots,L\}$ represent the indices for the access point (AP), channel slot, and transmission power level, respectively. Here, L denotes the total number of power levels, and denote by $\mathcal{A}$ is the set of all possible actions. The power levels are given by $P_l \in (0,1)$,

with $P_1 < \ldots < P_L$ and $\sum_l P_l = 1$. The binary indicator function is denoted by $\mathbb{1}(\cdot)$. The selection status of actions is represented by:

$$Z_{n;m,k,l} = \mathbb{1}(\mathbf{a}_n = (m, k, P_l)). \tag{5.2}$$

The received signal at the m-th AP on slot k is given by

$$y_{m;k} = \sum_{l=1}^{L} \sum_{n=1}^{N} \sqrt{P_l} h_{n;m,k} S_n Z_{n;m,k,l} + w_{m,k} \tag{5.3}$$

where $w_{m,k} \in \mathcal{CN}(0, N_0)$ represents the additive white Gaussian noise (AWGN) on slot k at the receiver of the m-th AP, N_0 denotes the noise variance, and S_n represents the modulated symbol. As indicated in (5.3), it is possible that two or more users randomly contend for the same slot k at the same AP, where NOMA superposition decoding enables signal decoding through successive interference cancellation (SIC) steps.

Given $\mathbf{a}_n = (m, k, P_l)$ from the n-th user, the received signal-to-interference-plus-noise ratio (SINR) at the m-th AP on slot k with power level P_l based on the SIC procedure is expressed as:

$$SINR_{n;m,k,l} = \frac{P_l g_{n;m,k}}{\sum_{i=1}^{l-1} \sum_{n'=1, n' \neq n}^{N} P_i g_{n';m,k} Z_{n';m,k,i} + N_0}. \tag{5.4}$$

$SINR_{n;m,k,l}$ becomes $SNR_{n;m,k,l}$ if there is no interference ($\sum_{i=1}^{l-1} \sum_{n'=1}^{N} Z_{n';m,k,i} = 0$).

The criteria for successful decoding for action (m, k, P_l) are:

Con1) $\sum_{n=1}^{N} Z_{n;m,k,l'} \leq 1$, for $l' \geq l$ (no action collision)

Con2) $SINR_{n;m,k,l'} \geq \Gamma \sum_{n=1}^{N} Z_{n;m,k,l'}$, for $l' \geq l$ (SIC success)

where Γ denotes the SINR threshold. It is assumed that packets are successfully decoded only when both $Con1$ and $Con2$ are satisfied. $Con1$ represents a no-collision condition, ensuring that no more than two users select the same action. For instance, given an action (m, k, P_l), at most one user is permitted to choose this action for successful decoding. Specifically, if a user selects $P_{l'}$ with $l' \geq l$ for a given slot k at a given AP m, no other users may select the same action. This constraint is due to the power-domain NOMA technology, which allows multiple users to transmit packets simultaneously over the same resource block (RB) using distinct transmit power levels. Thus, an action (m, k, P_l) can be allocated to at most one user. Otherwise, packet decoding is assumed to fail due to random collisions (i.e., more than one user selects the same power level on the same RB), as the capture effect is not considered in this study.

$Con2$ pertains to channel fading, indicating that packet decoding is successful only if the SINR after SIC meets or exceeds the required threshold. Furthermore, packet transmission fails if the decoding of any higher power-level signal on the same RB is unsuccessful, as the SIC decoding order proceeds from higher to lower power levels.

Denote the broadcast feedback signal from the m-th AP as $\mathbf{FB}_m(t) = [FB_{m;1}(t), \ldots, FB_{m;KL}(t)]$, where $FB_{m;i}(t)$ is calculated as:

$$FB_{m;i}(t) = \begin{cases} 1 & \text{, for successful decoding at the } i\text{-th action} \\ 0 & \text{, otherwise.} \end{cases} \tag{5.5}$$

Based on the broadcast signals from all the APs, each user can obtain $\mathbf{Flag}(t) = [\mathbf{FB}_1(t), \cdots, \mathbf{FB}_M(t)]$ that indicates the decoding results on all the random access resources at each time step.

5.2.2 Problem formulation

Distributed grant-free self-clustering NORA algorithms are developed in order to enhance the system throughput while maintaining the fairness among users. For effectiveness, we consider the age of information (AOI) of user n at $t+1$ as

$$AOI_n(t+1) \triangleq \begin{cases} 1 & \text{, for successful decoding} \\ & \text{at step } t \\ AOI_n(t)+1 & \text{, otherwise.} \end{cases} \tag{5.6}$$

The time-averaged AOI (AAOI), proposed in [49], of user n is calculated by

$$AAOI_n = \frac{1}{2} + \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} AOI_n(t) \tag{5.7}$$

The algorithm design must ensure fairness among users by aiming to keep each user's $AAOI_n$ within specified threshold. The performance of the algorithms are evaluated using two metrics: the number of users meeting the AAOI constraint and the average system packet throughput.

For fairness-sensitive systems, fairness is assessed by counting the number of users whose $AAOI_n$ values are below the AAOI threshold. Conversely, for fairness-tolerant systems, the average system packet throughput serves as the main metric to evaluate system performance without considering fairness. These metrics are formally defined as:

- *Number of users under the AAOI constraint:* Given N users, the num-

ber of users with $AAOI < AAOI_0$ is calculated by

$$N_{users} = \sum_{n=1}^{N} \mathbb{1}(AAOI_n < AAOI_0). \tag{5.8}$$

- *Average system packet throughput:* Given N users, the average system packet throughput is

$$N_{packet} = \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{1}(AOI_n(t+1) = 1). \tag{5.9}$$

Denote by $\mathbf{x}_n \in \mathcal{X}$ the mixed strategy of the n-th user,

$$\mathbf{x}_n = [x_{n;m,k,l}, \cdots, x_{n;m,k,l}, \cdots, x_{n;M,K,L}]^T \tag{5.10}$$

where $x_{n;k,l}$ denotes the probability that the n-th user takes action (m, k, P_l), and

$$x_{n;0,0,0} = 1 - \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} x_{n;m,k,l}. \tag{5.11}$$

In deployed environment, each user intends to find its own mixed strategy, $\mathbf{x}_n$, for choosing actions.

Distributed deep reinforcement learning aided self-clustering NORA algorithms optimizing the random access strategy of individual user to enhance the performance are proposed in Section 5.3.

5.3 Proposed DRL Algorithms

In this section, a DRL model that can be leveraged to develop two DRL algorithms such as DDQN and Dueling DDQN is formulated. Each agent

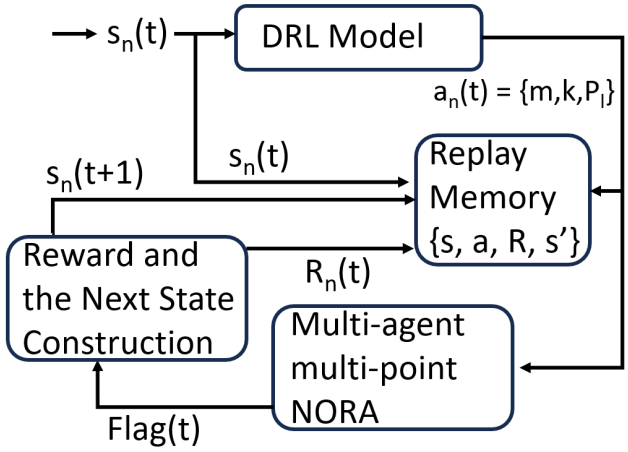


Figure 5.2: Markov decision process. $\mathbf{s}_n(t)$, $\mathbf{s}_n(t+1)$, $\mathbf{a}_n(t)$ and $R_n(t)$ denote the state at step t, the next state at step $t+1$, the action and the reward for step t, respectively. $\mathbf{Flag}(t)$ is the transmission feedback from the M APs. The RL information tuples in the replay memory are used in the training process further illustrated in Fig.3.

adopts the algorithm to independently conduct the RA resource selection in a multi-agent multi-point network.

5.3.1 Markov Decision Process Model

First, the state of each agent, which can be used for a DRL model to map the user's state from the state space to Q-value, is formulated by

$$\mathbf{s}_n(t) = [AOI_n^{\sigma}(t), \mathbf{Flag}(t), \mathbf{B}_n(t)] \tag{5.12}$$

where $\mathbf{B}_n(t)$ represents the behavior of user n at step t, and is given by

$$\mathbf{B}_n(t) \triangleq \begin{cases} \mathbf{a}_n(t) & \text{, if user } n \text{ takes an action} \\ (-1,-1,-1) & \text{, otherwise.} \end{cases} \tag{5.13}$$

and $AOI_n^{\sigma}(t)$ denotes a truncated AOI of user n:

$$AOI_n^{\sigma}(t) = \min\{AOI_n(t), \sigma\} \tag{5.14}$$

where σ denotes the AOI truncation threshold. The state definition should contain transmission status information of the user so that the DRL model can map the user's state to Q-value accurately. For this, an instantaneous AOI of each user is included to indicate the number of unsuccessful transmissions since the latest successful one. The AOI is used as a part of the state. The AOI grows if decoding continues to fail. Note that the corresponding state space faces a cursed dimensionality with a positive infinite range of AOI values. To avoid this, the AOI in the state definition is truncated by σ. Additionally, $\mathbf{Flag}(t)$, which indicates the status of all the RA resources, is designed as part of the state definition.

The reward of agent n at time step t is expressed with the $AOI_n^{\sigma}(t)$ as

$$R_n(t) \triangleq \begin{cases} AOI_n^{\sigma}(t) & \text{, if transmission success} \\ 0 & \text{, if } \mathbf{a}_n(t) = (0,0,0) \\ -AOI_n^{\sigma}(t) \cdot \mu & \text{, otherwise.} \end{cases} \tag{5.15}$$

where $\mu > 0$. By adopting this reward definition, users with relatively high AOI get a more dynamic range of reward than others. Thus, users may learn at varying speeds, leveraging the value for AOI, where users under high AOI learn faster than ones under low AOI.

Fig.5.2 illustrates the MDP model applicable to the proposed RA resource selection process. Each user takes an action to select one of the RA

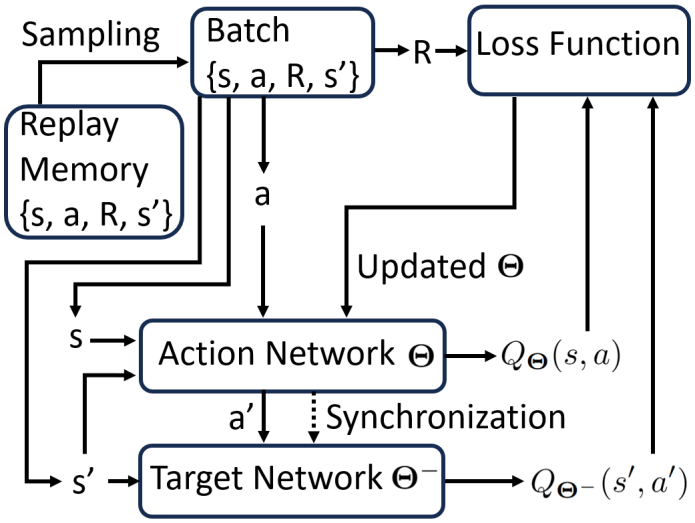


Figure 5.3: DRL model for each agent in the proposed network: both action and target networks and loss function settings rely on a choice of the proposed algorithms. $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^-$ denotes the parameters of action network and target network, respectively.

resources from $\mathbf{s}_n(t)$ at step t. After the multi-agent multi-point NORA process, the user observes the RL reward and formulates the next state, which are exploited to train DRL model.

5.3.2 DRL Model Updating

Fig.5.3 illustrates the DRL model applied for each agent to train the DRL network parameters. There are two DNNs adopted. One is action network to make action inference, while the other is target network used for training with a target Q-value calculation. The time step and user index are omitted for simplicity. With every step of action inference through the action network, a RL information tuple $\{\mathbf{s}, \mathbf{a}, R, \mathbf{s}'\}$, whose elements denote state, action, reward and next state of the MDP, respectively, is stored into the replay memory. When action network parameters are updated, N_s tuples are

randomly sampled from the replay memory to form a model training batch for loss calculation.

Given each tuple in the training batch at step t, current state $\mathbf{s}$ and action $\mathbf{a}$ are fed to the action network to output $Q_{\mathbf{\Theta}}(\mathbf{s}, \mathbf{a})$, where $\mathbf{\Theta}$ denotes the action network parameters. Based on the resulting **Flag**, AOI and $\mathbf{a}$, next state $\mathbf{s}'$ is computed and fed to the action network to get a prediction for action $\mathbf{a}'$ at the next step. Then both $\mathbf{s}'$ and $\mathbf{a}'$ are fed to the target network to calculate $Q_{\mathbf{\Theta}^-}(\mathbf{s}', \mathbf{a}')$. Based on these, the Double DQN loss function $\mathcal{L}(\mathbf{\Theta})$ is given by

$$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{N_s} \sum_{i=1}^{N_s} (Q_{Target}(R_i, \mathbf{s}'_i) - Q_{\mathbf{\Theta}}(\mathbf{s}_i, \mathbf{a}_i))^2 \tag{5.16}$$

where i is a tuple index, and

$$Q_{Target}(R_i, \mathbf{s}'_i) = R_i + \gamma \cdot Q_{\mathbf{\Theta}^-}(\mathbf{s}'_i, \arg\max_{\mathbf{a}} Q_{\mathbf{\Theta}}(\mathbf{s}'_i, \mathbf{a})) \tag{5.17}$$

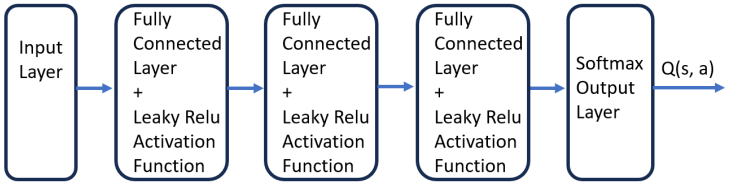
where $\mathbf{\Theta}^-$ denotes the target network parameters.

The $ADAM$ algorithm in [99] has been widely adopted in this field to reduce the loss function, updating the action network parameters $\mathbf{\Theta}$ due to its proven performance. To avoid the overshooting problem, the target network parameters $\mathbf{\Theta}^-$ is synchronized with $\mathbf{\Theta}$ every T_{target} updating iterations.

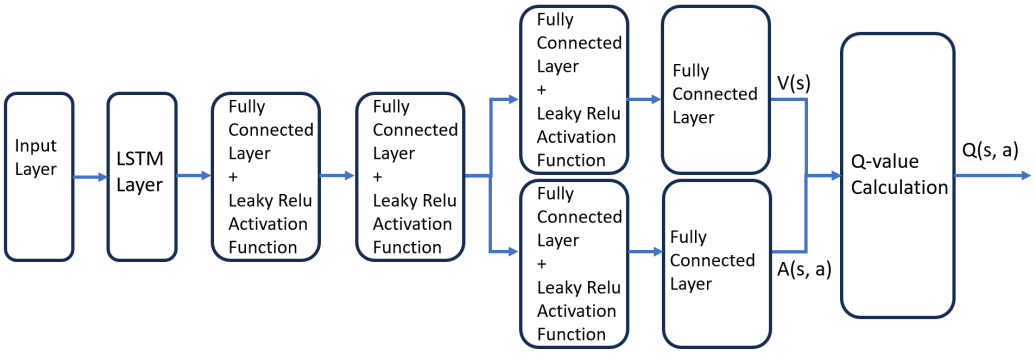
5.3.3 Deep Q Network Based Algorithms

Fig.5.4(a) shows the architecture of the proposed DDQN algorithm. The network consists of an input layer, 3 hidden FC layers with leaky Relu activation function, and 1 softmax output layer.

Given a set of Q-values at the output layer, exploration efficiency is im-



(a) DDQN network architecture



(b) Dueling DDQN network architecture

Figure 5.4: DRL Network architecture. Each FC layer is followed by a leacky Relu layer. (a) shows the DRL network used in the proposed DDQN algorithm. $Q(\mathbf{s}, \mathbf{a})$ denotes Q-values. (b) shows the DRL network used in the proposed Dueling DDQN algorithm. $V(\mathbf{s})$ and $A(\mathbf{s}, \mathbf{a})$ denote state value and advantages, respectively.

proved, adopting decaying ϵ-greedy as action policy, which is formulated by

$$\mathbf{a}_n(t) = \begin{cases} \text{random action from } \mathcal{A} & \text{, with probability} \\ & \epsilon(t) \\ \arg\max_{\mathbf{a}} Q_{\mathbf{\Theta}}(\mathbf{s}, \mathbf{a}) & \text{, with probability} \\ & 1 - \epsilon(t) \end{cases} \tag{5.18}$$

where $\epsilon(t)$ is given by

$$\epsilon(t) = \beta^{Upd} \cdot \epsilon_0 \tag{5.19}$$

where β, Upd and ϵ_0 denote the attenuation factor of the exploration prob-

ability, the counting index of action network parameter updates, and the initial exploration probability, respectively.

Fig.5.4(b) shows the architecture of the proposed Dueling DDQN algorithm. Since dueling networks have better ability of estimating state values, a LSTM layer is adopted to further leverage the history information hidden in the state. The network consists of 1 LSTM layer, 2 common FC layers, 2 FC layers for state value prediction, and 2 FC layers for advantage value prediction. Leaky Relu is adopted as activation function between the FC layers. Note that there is no activation function at the output layer so that the Dueling DDQN linearly outputs the state value and the advantages. The Q-values are calculated by

$$Q_{\mathbf{\Theta}}(\mathbf{s},\mathbf{a}) = V_{\mathbf{\Theta}}(\mathbf{s}) + (A_{\mathbf{\Theta}}(\mathbf{s},\mathbf{a}) - \frac{1}{|\mathcal{A}|}\sum_{\mathbf{a}^*\in\mathcal{A}} A_{\mathbf{\Theta}}(\mathbf{s},\mathbf{a}^*)) \tag{5.20}$$

where $V_{\mathbf{\Theta}}(\mathbf{s})$ and $A_{\mathbf{\Theta}}(\mathbf{s},\mathbf{a})$ denote state value and advantage value, respectively.

In terms of action selection, a precise prediction of Q-values can be enabled by the Dueling DDQN adopting softmax action policy [87]. This selects action according to a probability mass function (PMF) of the actions. Denoted by $\Pr(\cdot)$ the probability of an event and the probability of selecting action $\mathbf{a}$ at a given state $\mathbf{s}$ can be given by

$$\Pr(\mathbf{a}|\mathbf{s}) = \frac{e^{Q(\mathbf{s},\mathbf{a})}}{\sum_{\mathbf{a}^*\in\mathcal{A}} e^{Q(\mathbf{s},\mathbf{a}^*)}}. \tag{5.21}$$

Exploiting such softmax action policy, actions with higher Q-values have higher probabilities of being selected. Compared with greedy action policies

and its variations, this action policy responds to changes of all the Q-values rather than only focusing on which action has the highest Q-value.

In addition, a fair policy is proposed to assist the DRL model to improve the fairness between users. An active user transmits packet only if its instantaneous AOI is greater than a threshold. In other words, a user does not take any action if $AOI_n(t) \leq \theta_{Fair}$. The fair policy prevents users with relatively low AOI from continuing to occupy the RA resources, which helps to increase the success probabilities of other users.

Algorithm 7 represents the DQN based multi-point NORA process. It is worth noting that the process described in Algorithm 1 is applicable to both the DDQN algorithm and the Dueling DDQN algorithm, as different neural networks can be employed. In addition, this process is identical for all users. At each step, the user first identify whether it is active and whether its AOI is greater than the Fair threshold. If the user satisfies both the two conditions, it takes an action from its current state by feeding the state to its neural network and selecting an action according to its action policy. The user then transmits its packet using the selected slot and power level to the selected AP. After that, the user observes the feedback from the APs, and constructs the reward and the next state according to the feedback. Then the RL information tuple is stored in the replay memory for training usage. The network parameters will be updated every step once the replay memory is full by reducing the loss function in (5.17). The target network is synchronized with the action network every T_{target} iterations.

Algorithm 7 Deep Q Network based Multi-Point NORA

Initialization: Initialize the parameters of the action network to $\mathbf{\Theta}$, and the parameters of the target network $\mathbf{\Theta}^{-} = \mathbf{\Theta}$. The instantaneous age of information, AOI, is initialized to a random number in a range from 0 to 10, and state $\mathbf{s} = [AOI, zeros(1, MKL), -1, -1, -1]$. Initialize discount factor γ, target network update period T_{target}, batch size N_s, replay memory size N_m, number of updates $Upd = 0$.

while 1 **do**

\\take action according to the state

if Active && $AOI^{\sigma} > \theta_{Fair}$ **then**

Feed $\mathbf{s}$ to the action network to get the Q values.

Select $\mathbf{a}$ according to the Q values, and the action policy for the DDQN or the Dueling DDQN formulated in Chapter 5.3.3.

$\mathbf{B} \leftarrow \mathbf{a}$

Access the channel according to $\mathbf{a}$ and observe **Flag** through the broadcast signals from the APs.

$R \leftarrow AOI \cdot ((1+\mu) \cdot \mathbb{1}(success) - \mu \cdot \mathbb{1}(\mathbf{a} \neq (0,0,0)))$

else

$\mathbf{B} \leftarrow (-1,-1,-1)$

end if

\\update AOI and construct next state

$AOI = AOI \cdot (1 - \mathbb{1}(success)) + 1$

$AOI^{\sigma} = \min\{AOI, \sigma\}$

$\mathbf{s}' \leftarrow (AOI^{\sigma}, \mathbf{Flag}, \mathbf{B})$

\\update replay memory

if Active && $AOI^{\sigma} > \theta_{Fair}$ **then**

if Replay memory is full **then**

Delete the oldest tuple.

end if

Store the new tuple $\{\mathbf{s}, \mathbf{a}, R, \mathbf{s}'\}$ to the replay memory.

end if

\\update action network parameters

if Replay memory is full **then**

Randomly select N_s tuples from the replay memory to update $\mathbf{\Theta}$.

$\mathcal{L}(\mathbf{\Theta}) = \frac{1}{N_s} \sum_{i=1}^{N_s} (Q_{Target}(R_i, \mathbf{s}_i') - Q_{\mathbf{\Theta}}(\mathbf{s}_i, \mathbf{a}_i))^2$

where

$Q_{Target}(R_i, \mathbf{s}_i') = R_i + \gamma \cdot Q_{\mathbf{\Theta}^{-}}(\mathbf{s}_i', \arg\max_{\mathbf{a}} Q_{\mathbf{\Theta}}(\mathbf{s}_i', \mathbf{a}))$

Update $\mathbf{\Theta}$ using $ADAM$ algorithm.

$Upd \leftarrow Upd + 1$

end if

\\Synchronize target network with the action network

if Upd modulo $T_{target} = 0$ **then**

$\mathbf{\Theta}^{-} \leftarrow \mathbf{\Theta}$

end if

$\mathbf{s} \leftarrow \mathbf{s}'$

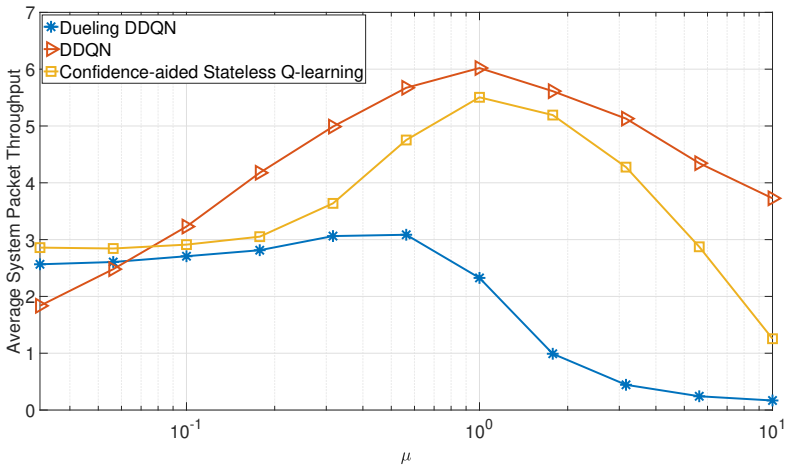
end while

Table 5.1: Learning Hyperparameters

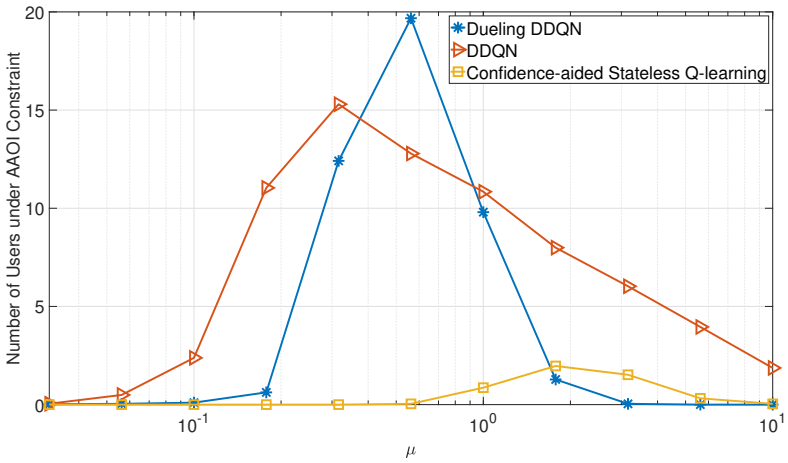
Hyperparameters	DDQN	Dueling-DDQN
Learning rate α	0.01	0.01
Discount factor γ	0.95	0.95
AOI truncation threshold θ_{Trunc}	2	20
μ	1	$10^{-0.25}$
Fair threshold θ_{Fair}	0	5
Width of LSTM layer	N/A	$MKL + 8$
Replay memory size	20	40
Batch size	5	2
Number of neurons in FC layers	$8MKL + 29$	$9MKL + 34$
Initial exploration probability ϵ_0	1	N/A
Exploration attenuation factor β	0.99	N/A

5.4 Simulations and Discussions

The effectiveness of the proposed DRL algorithms is examined and analyzed in Section 5.4. Channel inversion aided DRL GF-NOMA [87], confidence-aided stateless algorithm [90] and slotted NORA [38] are adopted as benchmarks for comparison. Note that [87] does not adopt the hard collision model that assumes packets can not be decoded successfully unless there is no collision. Instead, [87] only use SINR as the criteria for the SIC decoding process. Since the Channel inversion aided DRL GF-NOMA proposed in [87] is developed for single BS system where users are split into different clusters, the users are pre-assigned with different APs in this work. For all simulations, $\lambda_a = 0.5$, $A_0 = 1$, $\kappa = 3.5$, $\Gamma = -3dB$, $L = 3$ with $P_1 = 0.04$, $P_2 = 0.16$, $P_3 = 0.8$, $AAOI_0 = 10$. The learning hyperparameters are listed in Table 5.1. In order to emulate the inherent randomness of MTC networks, the instantaneous AOI of the users at the beginning of the simulations are modeled as uniformly distributed random variables in a range of 1 to 10.



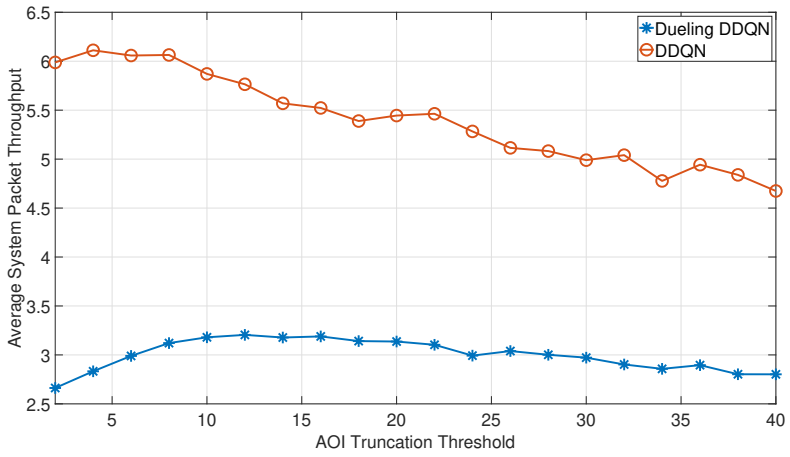
(a) Average system packet throughput versus μ



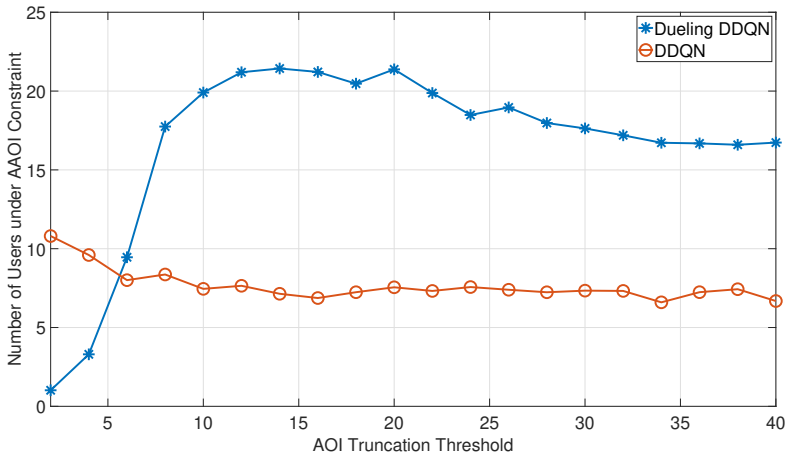
(b) Number of users under AAOI constraint versus μ

Figure 5.5: Effect of μ on average system packet throughput (a) and number of users under AAOI constraint (b). The proposed methods are Dueling DDQN and DDQN, and for comparison, the Confidence-aided Stateless Q-learning are depicted, when $M = 2$, $N = 48$, $K = 4$, $L = 3$.

Fig.5.5 shows the effect of reward magnitude for failure transmission, μ. DDQN performs best in terms of average system packet throughput, and it reaches its best performance when $\mu = 1$. Dueling DDQN achieves the highest number of users under AAOI constraint when $\mu = 10^{-0.25}$. Based on these observations, the value of μ is set to 1 and $10^{-0.25}$ for DDQN and



(a) Average system packet throughput versus σ



(b) Number of users under AAOI constraint versus σ

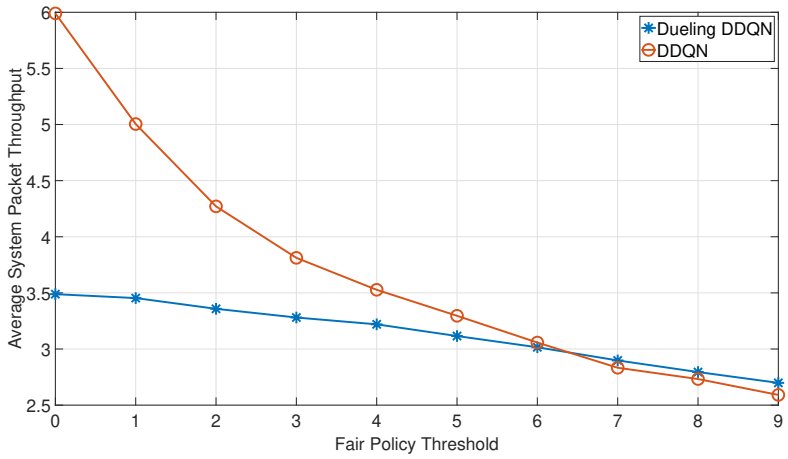
Figure 5.6: Effect of σ on average system packet throughput (a) and number of users under AAOI constraint (b). The proposed methods are Dueling DDQN and DDQN, when $M = 2$, $N = 48$, $K = 4$, $L = 3$.

Dueling DDQN, respectively, in other simulations.

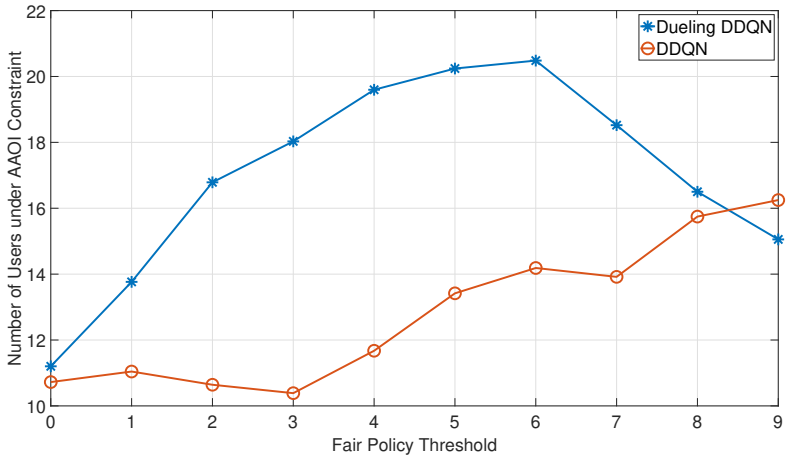
Fig.5.6 shows the effect of the AOI truncation threshold, σ, on system throughput and number of users under AAOI constraint. The performance of the DDQN algorithm degrades in terms of system throughput as σ increases, and it exhibits insensitivity to changes in σ when considering the number of users under the AAOI constraint. In contrast, the Dueling DDQN

algorithm achieves peak system throughput at $\sigma = 12$ and maximizes the number of users under the AAOI constraint at $\sigma = 14$. This is due to the inherent characteristics of the dueling network architecture, which split the estimation of Q-values into state values and advantages, and thus enabling the algorithm to leverage the AOI information through the truncated AOI in the state. Since the size of the state space increases with σ, the Dueling DDQN algorithm has a significant performance gain when σ increases from 2 to 14. However, the algorithm suffers updating efficiency degradation when the state space gets large, which is the reason why the performance degrades gradually when σ is greater than 20.

Fig.5.7 shows the effect of the Fair policy threshold, θ_{Fair}, on the system throughput and the number of users under AAOI constraint. It can be seen that the system throughput of the Dueling DDQN algorithm drops more slower than the DDQN algorithm when θ_{Fair} increases. Moreover, the Dueling DDQN algorithm gains more users under AAOI constraint than the DDQN algorithm as θ_{Fair} increases, and it reaches its highest average number of users under AAOI constraint when θ_{Fair} equals to 6. In other words, the Dueling DDQN algorithm can obtain a relative improvement on number of users under AAOI constraint with a minor cost of system throughput. The above observations shows that the Fair policy fits the Dueling DDQN algorithm very well while it is not worth adopting it for the DDQN algorithm. The reason behind this is that the Dueling DDQN algorithm achieves more users with medium user throughput so that there will be more RA resources enabled when applying the Fair policy to force the users with medium to high throughput to wait. Additionally, the system throughput cost of the



(a) Average system packet throughput versus θ_{Fair}

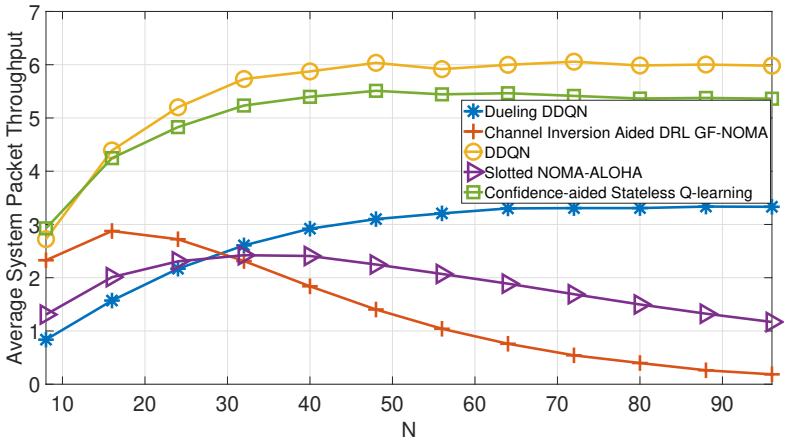


(b) Number of users under AAOI constraint versus θ_{Fair}

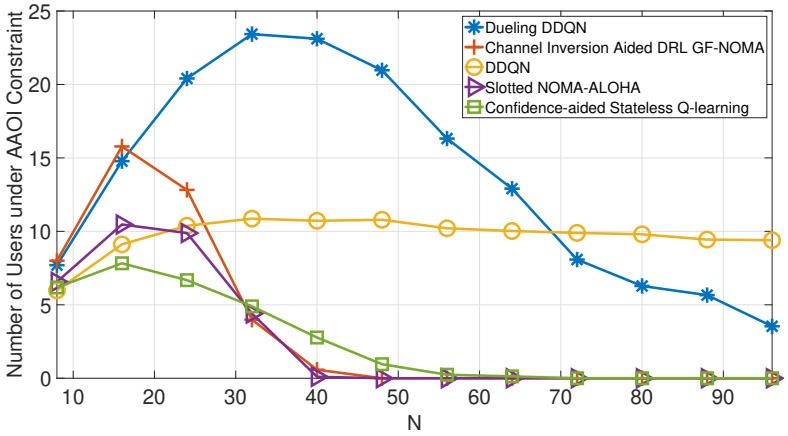
Figure 5.7: Effect of θ_{Fair} on average system packet throughput (a) and number of users under AAOI constraint (b). The proposed methods are Dueling DDQN and DDQN, when $M = 2$, $N = 48$, $K = 4$, $L = 3$.

Fair policy is relatively high for DDQN algorithm due to the high-throughput users.

Fig.5.8 shows the effect of number of users on the average packet throughput and the number of users under AAOI constraint. It can be seen that the DDQN algorithm is the best in terms of system throughput when $N \geq 16$. However, the Dueling DDQN algorithm achieves 200% the number of users



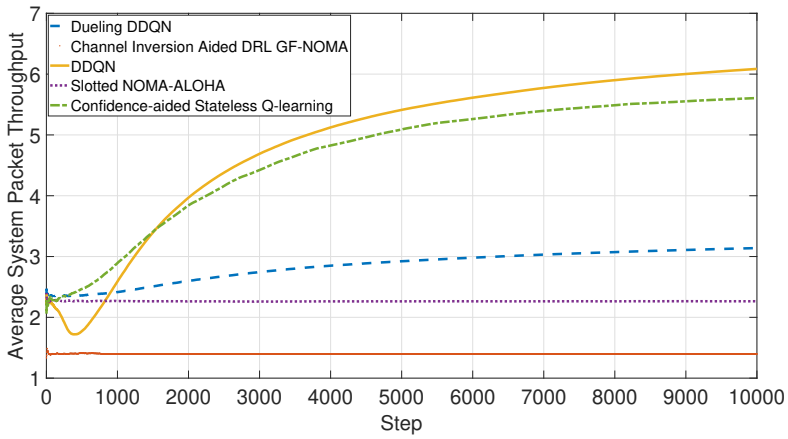
(a) Average system packet throughput versus N



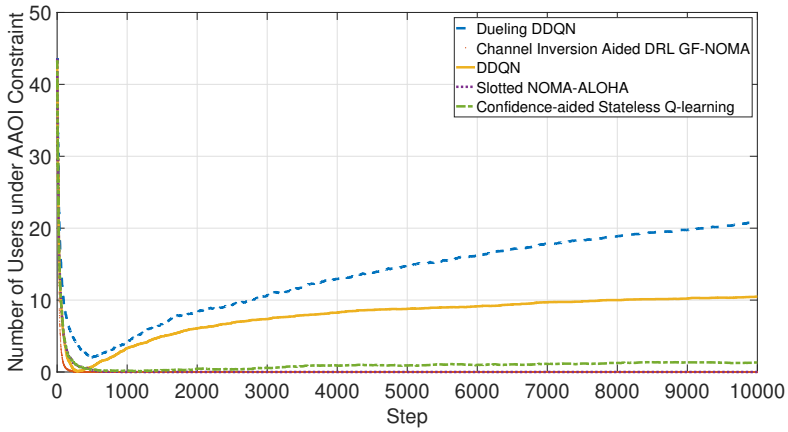
(b) Number of users under AAOI constraint versus N

Figure 5.8: Effect of N on average system packet throughput (a) and number of users under AAOI constraint (b). The proposed methods are Dueling DDQN and DDQN, and for comparison, the benchmark schemes are depicted, when $M = 2$, $K = 4$, $L = 3$.

under AAOI constraint when $N = 1.33MLK$, compared with the DDQN algorithm. This is because the Dueling DDQN algorithm results in a more dynamic action strategy than the DDQN algorithm due to its large state space and softmax action policy. The consequential result is that the RA resources can be interchangeably leveraged by the users, which alleviates collisions. However, the DDQN algorithm is still consistent when $N \geq 3MLK$



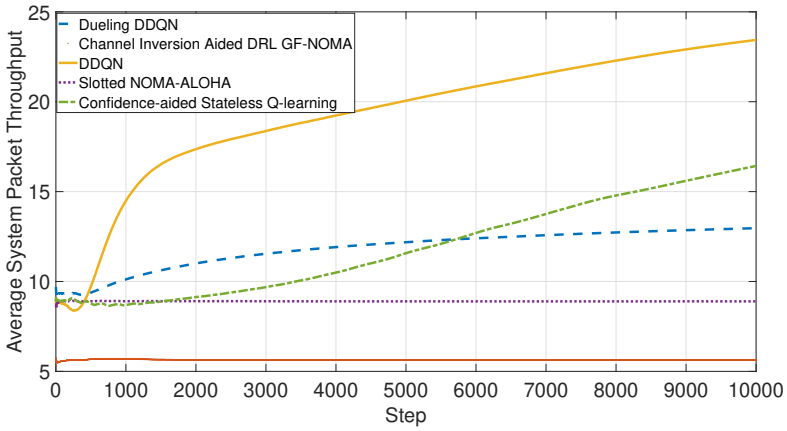
(a) Average system packet throughput versus step



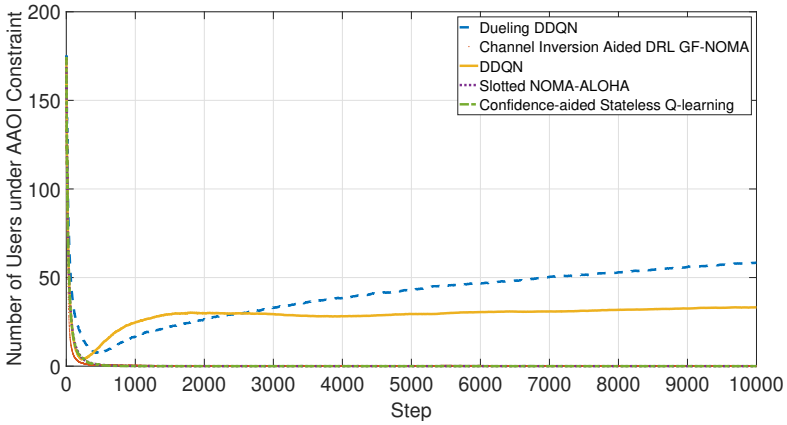
(b) Number of users under AAOI constraint versus step

Figure 5.9: Performances related to average system packet throughput (a) and number of users under AAOI contraints (b) against steps. The proposed methods are Dueling DDQN and DDQN, and for comparison, the benchmark schemes are depicted, when $M = 2$, $N = 48$, $K = 4$, $L = 3$.

while the number of users under AAOI constraint of the Dueling DDQN algorithm significantly decreases when $N > 40$. The reason is that the collision is not avoidable by interchangably choosing different RA resources when the number of users is much higher than the number of RA resources. In this situation, it is more effective to make part of the users always success



(a) Average system packet throughput versus step



(b) Number of users under AAOI constraint versus step

Figure 5.10: Performances related to average system packet throughput (a) and number of users under AAOI contraints (b) against steps. The proposed methods are Dueling DDQN and DDQN, and for comparison, the benchmark schemes are depicted, when $M = 8$, $N = 192$, $K = 4$, $L = 3$.

with fixed RA resources. The above observations indicate that the Dueling DDQN algorithm is more suitable for application scenarios where users are latency-critical, and the number of users need to be limited to match the amount of available RA resources. However, the DDQN algorithm is good at application scenarios where the number of users is higher than the number of RA resources, and system throughput is the main concern.

Fig.5.9 and Fig.5.10 show the convergence properties of the proposed algorithms and the benchmarks for $M = 2$ and $M = 8$, respectively. It can be seen that the DDQN algorithm achieves the highest average packet throughput while the Dueling DDQN algorithm is the best in terms of number of users under AAOI constraint. The system throughput gap between the Confidence-aided Stateless Q-learning and the DDQN algorithm is enlarged when the problem size increases from $M = 2$ to $M = 8$. Furthermore, the crossover point of the system throughput of the DDQN algorithm and the Confidence-aided Stateless Q-learning shift from around 1400 steps to around 400 steps as the number of APs increases from 2 to 8. This observation shows that the proposed DDQN algorithm outperforms the lookup-table method in terms of both performance and scalability. This is due to the different function approximators used. The Q-table only updates one of its Q-values at each step while most of the parameters of DNN-based function approximator are updated at every steps, which means that DRL methods are more robust than lookup-table methods as the problem size increases.

Due to the multiple APs and the heterogeneous average channel gains between users in the system model adopted in this chapter, there is a gap between the system throughput of the proposed algorithms and the maximum possible throughput of the system. Nevertheless, the proposed system are capable of achieving better system level performance than single AP systems when the geographic distribution of users is uneven between APs. Additionally, it also offers the users more flexibilities when part of the APs provide low QoS due to shadowing effect.

5.5 Summary

This chapter presents a distributed deep reinforcement learning framework designed to address the joint selection of AP, transmission slots, and power levels in multi-agent multi-point NORA systems. Two deep Q network based algorithms were developed to enable distributed action selection strategies. The simulation results demonstrate that the proposed algorithms exhibit good scalability and surpass existing benchmarks in terms of system throughput and fairness under high-traffic congestion scenarios, supporting massive connectivity in 6G networks. Moreover, the two algorithms exhibit complementary strengths, with distinct advantages in fairness, system throughput, and resilience to extreme congestion random access states. By leveraging the model-free distributed learning framework and the multi-point NORA methodology, the proposed schemes enable efficient random access for resource-constrained MTC networks operating in asymmetric environments. Additionally, the proposed scheme produces extra flexibility compared with conventional single AP system, which can be a potential candidate for addressing the challenge of high mobility in vehicular communications. Joint optimization for vehicular network related parameters will be investigated in future works.

6 Conclusions and Future Work

6.1 Conclusions

This thesis has presented a comprehensive investigation into edge intelligent multi-user uplink radio access methods, tailored for emerging 6G wireless networks. Addressing the critical challenges of adaptive modulation, massive connectivity, and distributed learning within dynamic environments, the research herein contributes significantly to the development of intelligent, decentralized, and efficient communication strategies.

Firstly, a novel federated learning-based method for adaptive orthogonal frequency division multiplexing with index modulation (OFDM-IM) was developed, leveraging distributed k-means clustering. The proposed Fed-k-means approach effectively aggregated learning outcomes across distributed user devices into a cohesive global model. This method demonstrated substantial improvements in throughput and bit-error-rate performance while minimizing the federation overhead, making it highly suitable for resource-constrained machine-type communication devices.

Secondly, addressing the critical issue of massive connectivity, this thesis introduced a learning-based non-orthogonal random access (NORA) strategy optimized through tailored multi-state and confidence-aided Q-learning

algorithms. By modeling joint slot and power-level selection as a Markov decision process (MDP), the proposed methods significantly enhanced system throughput and user fairness, particularly under conditions of high congestion and without requiring channel state information (CSI). Crucially, the federated k-means clustering adaptive OFDM-IM in Chapter 3 highlighted the necessity for efficient random access technologies, directly motivating the developments detailed in Chapter 4.

Lastly, extending the principles of distributed reinforcement learning, a decentralized deep Q-network-based NORA framework was presented for multi-AP scenarios. The developed algorithms allowed users to autonomously and intelligently select access points, transmission slots, and power levels without CSI. Extensive simulations demonstrated that this approach significantly improved system throughput, fairness, and scalability, underscoring its robust applicability in heterogeneous high-traffic scenarios common to future 6G networks.

Collectively, these contributions demonstrate the potential of integrating edge intelligence with adaptive and learning-driven methods to overcome current limitations in uplink radio access, supporting the ambitious performance targets of future 6G ecosystems.

6.2 Future work

While this thesis advances the state-of-the-art significantly, several promising directions for future research have emerged:

- **Hybrid Learning Models**: Exploring the integration of federated learning and deep reinforcement learning in a hybrid framework could

offer enhanced adaptive capabilities and further reduce convergence time and computational load.

- **Energy-Efficient Learning Algorithms**: Given the stringent power constraints in many edge devices, future work should aim to develop even more energy-efficient learning algorithms that minimize power consumption without compromising communication performance.

- **Security and Privacy**: Incorporating advanced cryptographic and secure aggregation techniques into the federated learning framework to enhance data privacy and resist potential adversarial attacks warrants detailed exploration.

- **Real-Time Deployment**: Practical implementation and real-time validation of the proposed methods in actual wireless testbeds would offer valuable insights, enabling refinements to adapt theoretical models to realistic operational conditions.

- **Cross-layer Optimization**: Future studies should examine the benefits of cross-layer optimization, considering not only the physical and MAC layers but also higher network protocol layers, to create comprehensive and efficient wireless communication solutions.

Addressing these future research avenues will not only enhance the robustness and effectiveness of intelligent uplink access methods but also significantly contribute to realizing the vision of fully autonomous, intelligent, and resilient 6G wireless networks.

Appendices

A

Derivation of $ASR_{n;k,l}$

In this appendix, the $ASR_{n;k,l}$ in Chapter 4 is derived. Consider $L = 3$ power levels, ASRs of action (k, P_l) for $l \in \{1, 2, 3\}$ are discussed in the following. To simplify the equations in the later derivations, denote $\beta_{n;k,l} = \frac{\Gamma_l}{\bar{g}_{n;k} P_l}$ where $n \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$, and $l \in \{1, \ldots, L\}$.

A.1 When the transmit power level P_3 is chosen

The expectation of the ASR is given by

$$\begin{aligned} \mathbb{E}[ASR_{n;k,3}] &= \Pr(Con1, Con2) \\ &= \Pr(Con2|Con1) \Pr(Con1) \end{aligned} \tag{A.1}$$

where

$$\Pr(Con1) = \prod_{n' \neq n} (1 - x_{n';k,3}) \tag{A.2}$$

and $\Pr(Con2|Con1)$ is given by

$$\Pr\left(g_{n;k} \geq \frac{\Gamma_3}{P_3}[N_0 + \sum_{l=1}^{2}\sum_{n'\neq n} g_{n';k}P_l Z_{n';k,l}]|Con1\right)$$
$$= e^{-\beta_{n;k,3}N_0}\prod_{n'\neq n}\phi_{n';k,3} \tag{A.3}$$

where $\phi_{n';k,3}$ is averaged over $g_{n';k}$, and is given by

$$\phi_{n';k,3} = \mathbb{E}[\frac{1}{1+\bar{g}_{n';k}\beta_{n;k,3}(P_2 Z_{n';k,2} + P_1 Z_{n';k,1})}|Con1]. \tag{A.4}$$

When *Con1* is satisfied, for $l \in \{1,2\}$

$$Z_{n';k,l} = \begin{cases} 1 & \text{w.p. } \frac{x_{n';k,l}}{1-x_{n';k,3}}, \\ 0 & \text{w.p. } 1-\frac{x_{n';k,l}}{1-x_{n';k,3}}. \end{cases} \tag{A.5}$$

By substituting (A.5) into (A.4), $\phi_{n';k,3}$ is averaged over $Z_{n';k,l}$,

$$\begin{aligned}\phi_{n';k,3} = {} & (1-\frac{x_{n';k,2}}{1-x_{n';k,3}})(1-\frac{x_{n';k,1}}{1-x_{n';k,3}}) \\ & + \frac{1}{1+\bar{g}_{n';k}\beta_{n;k,3}P_1}(1-\frac{x_{n';k,2}}{1-x_{n';k,3}})(\frac{x_{n';k,1}}{1-x_{n';k,3}}) \\ & + \frac{1}{1+\bar{g}_{n';k}\beta_{n;k,3}P_2}(\frac{x_{n';k,2}}{1-x_{n';k,3}})(1-\frac{x_{n';k,1}}{1-x_{n';k,3}}) \\ & + \frac{1}{1+\bar{g}_{n';k}\beta_{n;k,3}(P_2+P_1)}(\frac{x_{n';k,2}}{1-x_{n';k,3}})(\frac{x_{n';k,1}}{1-x_{n';k,3}}).\end{aligned} \tag{A.6}$$

A.2 When the transmit power level P_2 is chosen

For $ASR_{n;k,2}$, the $Con1$ can be decomposed into two situations: $Con1')$ $\forall n' \neq n, Z_{n';k,2} = Z_{n';k,3} = 0$; $Con1'')$ $\forall n' \neq n, Z_{n';k,2} = 0, \sum_{n' \neq n} Z_{n';k,3} = 1$. The $ASR_{n;k,2}$ under the two situations are derived respectively so that

$$\begin{aligned} \mathbb{E}[ASR_{n;k,2}] &= \Pr(Con1, Con2) \\ &= \Pr(Con1', Con2) + \Pr(Con1'', Con2). \end{aligned} \tag{A.7}$$

A.2.1 For $Con1'$

The probability of successful transmission when there is NOT any user choosing $l = 3$ is given by

$$\begin{aligned} \Pr(Con1', Con2) &= \Pr(Con2|Con1') \Pr(Con1') \\ &= \Pr(SINR_{n;k,2} \geq \Gamma_2 | Con1') \prod_{n' \neq n} (1 - x_{n';k,3} - x_{n';k,2}) \end{aligned} \tag{A.8}$$

where

$$\begin{aligned} &\Pr(SINR_{n;k,2} \geq \Gamma_2 | Con1') \\ &= \Pr\left(g_{n;k} \geq \frac{\Gamma_2}{P_2}(N_0 + \sum_{n' \neq n} g_{n';k} P_1 Z_{n';k,1}) | Con1' \right) \\ &= e^{-\beta_{n;k,2} N_0} \prod_{n' \neq n} \phi_{n';k,2'} \end{aligned} \tag{A.9}$$

where $\phi_{n';k,2}$ is averaged over $g_{n';k}$, and is given by

$$\begin{aligned}\phi_{n';k,2'} &= \mathbb{E}[e^{-\beta_{n;k,2}g_{n';k}P_1Z_{n';k,1}}|Con1'] \\ &= \mathbb{E}[\frac{1}{1+\bar{g}_{n';k}\beta_{n;k,2}P_1Z_{n';k,1}}|Con1'].\end{aligned} \tag{A.10}$$

When $Con1$ is satisfied

$$Z_{n';k,1} = \begin{cases} 1 & \text{w.p. } \frac{x_{n';k,1}}{1-x_{n';k,3}-x_{n';k,2}}, \\ 0 & \text{w.p. } 1-\frac{x_{n';k,1}}{1-x_{n';k,3}-x_{n';k,2}}. \end{cases} \tag{A.11}$$

By substituting (A.11) into (A.10), $\phi_{n';k,2'}$ is averaged over $Z_{n';k,1}$,

$$\phi_{n';k,2'} = 1 - \frac{\bar{g}_{n';k}\beta_{n;k,2}P_1}{1+\bar{g}_{n';k}\beta_{n;k,2}P_1}\frac{x_{n';k,1}}{1-x_{n';k,3}-x_{n';k,2}}. \tag{A.12}$$

A.2.2 For $Con1''$

The probability of successful transmission when there is only one user choosing $l = 3$ is given by

$$\begin{aligned}&\Pr(Con1'', Con2) \\ &= \sum_{m\neq n} x_{m;k,3} \prod_{n'\neq m,n} (1-x_{n';k,3}-x_{n';k,2})\xi_{m,n;k}\end{aligned} \tag{A.13}$$

where

$$\begin{aligned}\xi_{m,n;k} &= \Pr(SINR_{m;k,3} \geq \Gamma_3, SINR_{n;k,2} \geq \Gamma_2 | Con1'') \\ &= \Pr\Bigg(g_{m;k} \geq \frac{\Gamma_3}{P_3}(N_0 + g_{n;k}P_2 + \sum_{n' \neq n} g_{n';k}P_1 Z_{n';k,1}), \\ &\qquad g_{n;k} \geq \frac{\Gamma_2}{P_2}(N_0 + \sum_{n' \neq n} g_{n';k}P_1 Z_{n';k,1}) | Con1'' \Bigg). \end{aligned} \tag{A.14}$$

Because $g_{m;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\begin{aligned}\xi_{m,n;k} &= \mathbb{E}[\int_{\beta_{n;k,2}b}^{+\infty} \int_{ay+\beta_{m;k,3}b}^{+\infty} e^{-x}dx e^{-y}dy | Con1''] \\ &= \mathbb{E}[\frac{e^{-[(1+a)\beta_{n;k,2}+1]b}}{1+a} | Con1''] \\ &= \frac{e^{-[(1+a)\beta_{n;k,2}+1]N_0}}{1+a} \prod_{n' \neq m,n} \phi_{n';k,2''} \end{aligned} \tag{A.15}$$

where $b = N_0 + \sum_{n' \neq m,n} g_{n';k}P_1 Z_{n';k,1}$, $a = \beta_{m;k,3}\bar{g}_{n;k}P_2$, $c = (1+a)\beta_{n;k,2} + 1$, and $\phi_{n';k,2''}$ is averaged over $g_{n';k}$ by

$$\begin{aligned}\phi_{n';k,2''} &= \mathbb{E}[e^{-cg_{n';k}P_1 Z_{n';k,1}} | Con1''] \\ &= \mathbb{E}[\frac{1}{1 + \bar{g}_{n';k}cP_1 Z_{n';k,1}} | Con1''].\end{aligned} \tag{A.16}$$

By substituting (A.11) into (A.16), $\phi_{n';k,2''}$ is averaged over $Z_{n';k,1}$, and is given by

$$\phi_{n';k,2''} = 1 - \frac{\bar{g}_{n';k}cP_1}{1 + \bar{g}_{n';k}cP_1} \frac{x_{n';k,1}}{1 - x_{n';k,3} - x_{n';k,2}}. \tag{A.17}$$

A.3 When the transmit power level P_1 is chosen

For $ASR_{n;k,1}$, the $Con1$ can be decomposed into three situations: $Con1')$ $\forall n' \neq n, Z_{n';k,1} = Z_{n';k,2} = Z_{n';k,3} = 0$; $Con1'')$ $\forall n' \neq n, Z_{n';k,1} = 0, \sum_{n' \neq n} Z_{n';k,3} = 1, \sum_{n' \neq n} Z_{n';k,2} = 1$; $Con1''')$ $\forall n' \neq n, Z_{n';k,1} = 0, \sum_{n' \neq n}(Z_{n';k,2} + Z_{n';k,3}) = 1$. The $ASR_{n;k,1}$ under the three situations are derived respectively so that

$$\begin{aligned} \mathbb{E}[ASR_{n;k,1}] &= \Pr(Con1, Con2) \\ &= \Pr(Con1', Con2) + \Pr(Con1'', Con2) + \Pr(Con1''', Con2). \end{aligned} \tag{A.18}$$

A.3.1 For $Con1'$

The probability of successful transmission when there is no user choosing $l = 3$ or $l = 2$ is

$$\begin{aligned} &\Pr(Con1', Con2) = \Pr(Con2|Con1') \Pr(Con1') \\ &= \Pr(SINR_{n;k,1} \geq \Gamma_1) \prod_{n' \neq n} (1 - x_{n';k}) \end{aligned} \tag{A.19}$$

and

$$\begin{aligned} &\Pr(SINR_{n;k,1} \geq \Gamma_1) = \Pr(g_{n;k} \geq \frac{\Gamma_1}{P_1} N_0) \\ &= e^{-\beta_{n;k,1} N_0}. \end{aligned} \tag{A.20}$$

A.3.2 For $Con1''$

The probability of successful transmission with only two users, choosing $l = 3$ and $l = 2$, is given by

$$\begin{aligned}&\Pr(Con1'', Con2)\\&= \sum_{v\neq n} x_{v;k,3} \sum_{m\neq v,n} x_{m;k,2} \prod_{n'\neq v,m,n} (1 - x_{n';k})\xi_{v,m,n;k}\end{aligned} \tag{A.21}$$

where

$$\begin{aligned}&\xi_{v,m,n;k}\\&= \Pr(SINR_{v;k,3} \geq \Gamma_3, SINR_{m;k,2} \geq \Gamma_2, SINR_{n;k,1} \geq \Gamma_1)\end{aligned} \tag{A.22}$$

Because $g_{v;k}$, $g_{m;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\xi_{v,m,n;k} = \frac{1}{(q\beta_{v;k,3} + 1)ur} e^{-u(r\beta_{n;k,1}+1)N_0} \tag{A.23}$$

where $q = \bar{g}_{m;k}P_2$, $r = \bar{g}_{n;k}P_1$, $u = (q\beta_{v;k,3} + 1)\beta_{m;k,2} + \beta_{v;k,3}$.

A.3.3 For $Con1'''$

The probability of successful transmission when there is only one user choosing $l = 3$ or $l = 2$ is

$$\Pr(Con1''', Con2) = \sum_{i=2}^{3} \sum_{j\neq n} x_{j;k,i} \prod_{n'\neq j,n} (1 - x_{n';k})\xi_{j,n;k} \tag{A.24}$$

where

$$\xi_{j,n;k} = \Pr\left(g_{j;k} \geq \frac{\Gamma_i}{P_i}(N_0 + g_{n;k}P_1), g_{n;k} \geq \frac{\Gamma_1}{P_1}N_0 \right). \tag{A.25}$$

Because $g_{j;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\xi_{j,n;k} = \frac{e^{-(1+f)o-p}}{1+f} \tag{A.26}$$

where $f = \Gamma_i \frac{P_1 \bar{g}_{n;k}}{P_i \bar{g}_{j;k}}$, $p = \beta_{j;k,i}N_0$, and $o = \beta_{n;k,1}N_0$.

B

Matlab Implementation of Deep Reinforcement Learning

In this appendix, the Matlab implementation details of the DRL assisted NORA are presented. MTDs become active randomly to save energy due to inherent sporadic traffic. To simulate this, the DRL agents need to be active or inactive based on the status of the MTDs. Matlab does have a reinforcement learning toolbox, while it provides limited flexibility, which makes it impossible to have full control of the agents. Therefore, the DRL algorithms are implemented from scratch using the deep learning toolbox.

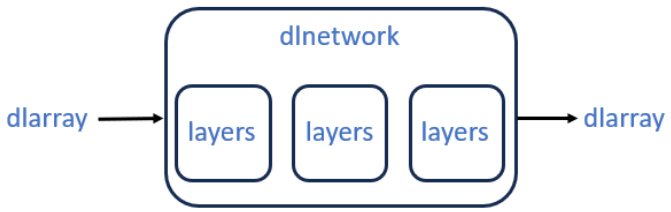


Figure B.1: Data Structure of Matlab Deep Learning Toolbox

Fig.B.1 shows the data structure of the deep learning toolbox. There is a class called layers in Matlab, which is for the various types of network layer provided by the deep learning toolbox, and a class called dlnetwork is for DNNs constructed from given layers. Note that all the input and output data of a dlnetwork are dlarray.

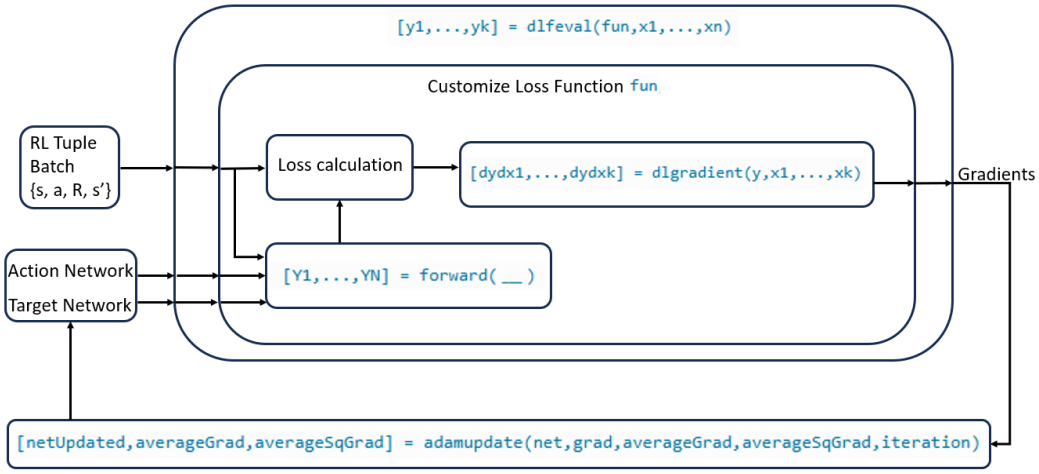


Figure B.2: Matlab Implementation Diagram

The implementation diagram is shown in Fig.B.2. To train a DNN using a customized loss function, a Matlab function called dlfeval must be used as an interface for the automatic differentiation process to track the gradient of the customized loss function. dlfeval must be called with the same input and output arguments as the user-defined function fun that calculates the loss and the gradients. In each iteration of the DRL, the action network and the target network, and a batch of the RL tuples that contains the current state $\mathbf{s}$, the action $\mathbf{a}$, the reward R, and the next state $\mathbf{s}'$, are fed to fun. fun calls forward to compute the Q-values used for calculating the loss, and then calls dlgradient to compute the gradients with respect to the learnable parameters of the action network. Note that every operation in the loss calculation must be trackable by dlgradient to ensure a successful gradients calculation. Once the gradients are acquired, a Matlab function that updates the network parameters (e.g. adamupdate if $ADAM$ algorithm is adopted) can be called to update the action network parameters using the gradients. The target network parameters are synchronized to the action

network parameters periodically or gradually.

References

[1] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular iot networks: Current issues and machine learning-assisted solutions," *IEEE Communications Surveys Tutorials*, vol. 22, no. 1, pp. 426–471, 2020.

[2] E. Peltonen, M. Bennis, and et al., "6G white paper on edge intelligence," *arXiv preprint arXiv:2004.14850*, 2020.

[3] T. Gong, L. Zhu, F. R. Yu, and T. Tang, "Edge intelligence in intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 9, pp. 8919–8944, 2023.

[4] F. Foukalas and A. Tziouvaras, "Edge artificial intelligence for industrial internet of things applications: an industrial edge intelligence solution," *IEEE Industrial Electronics Magazine*, vol. 15, no. 2, pp. 28–36, 2021.

[5] V. Hayyolalam, M. Aloqaily, Ö. Özkasap, and M. Guizani, "Edge intelligence for empowering iot-based healthcare systems," *IEEE Wireless Communications*, vol. 28, no. 3, pp. 6–14, 2021.

[6] W. Y. B. Lim, Z. Xiong, D. Niyato, X. Cao, C. Miao, S. Sun, and Q. Yang, "Realizing the metaverse with edge intelligence: A match made in heaven," *IEEE Wireless Communications*, vol. 30, no. 4, pp. 64–71, 2022.

[7] G. White and S. Clarke, "Urban intelligence with deep edges," *IEEE Access*, vol. 8, pp. 7518–7530, 2020.

[8] R. Zhang, Y.-C. Liang, and S. Cui, "Dynamic resource allocation in cognitive radio networks," *IEEE signal processing magazine*, vol. 27, no. 3, pp. 102–114, 2010.

[9] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (urllc)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, 2019.

[10] M. U. A. Siddiqui, H. Abumarshoud, L. Bariah, S. Muhaidat, M. A. Imran, and L. Mohjazi, "Urllc in beyond 5g and 6g networks: An interference management perspective," *IEEe Access*, vol. 11, pp. 54639–54663, 2023.

[11] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3347–3366, 2021.

[12] W. Mao, Z. Zhao, Z. Chang, G. Min, and W. Gao, "Energy-efficient industrial internet of things: Overview and open issues," *IEEE transactions on industrial informatics*, vol. 17, no. 11, pp. 7225–7237, 2021.

[13] M. A. Qureshi and C. Tekin, "Fast learning for dynamic resource allocation in ai-enabled radio networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 95–110, 2019.

[14] O. Habachi, M.-A. Adjif, and J.-P. Cances, "Fast uplink grant for noma: A federated learning based approach," in *International symposium on ubiquitous networking*, pp. 96–109, Springer, 2019.

[15] Y. Ko and J. Choi, "Reinforcement learning for NOMA-ALOHA under fading," *IEEE Transactions on Communications*, vol. 70, no. 10, pp. 6861–6873, 2022.

[16] A.-T. H. Bui and A. T. Pham, "Deep reinforcement learning-based access class barring for energy-efficient mmtc random access in lte networks," *IEEE Access*, vol. 8, pp. 227657–227666, 2020.

[17] M. Al-Quraan, L. Mohjazi, and et al., "Edge-native intelligence for 6G communications driven by federated learning: A survey of trends and challenges," *arXiv preprint arXiv:2111.07392*, 2021.

[18] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[19] B. McMahan, E. Moore, and et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, pp. 1273–1282, PMLR, 2017.

[20] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[21] H. H. Kumar, V. Karthik, and M. K. Nair, "Federated k-means clustering: A novel edge AI based approach for privacy preservation," in *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 52–56, IEEE, 2020.

[22] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *In Proceedings of the Seventeenth International Conference on Machine Learning*, Citeseer, 2000.

[23] H. Sun, X. Ma, and R. Q. Hu, "Adaptive federated learning with gradient compression in uplink NOMA," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16325–16329, 2020.

[24] R. S. Sutton and et al., *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[25] M. S. Sarwar, I. N. A. Ramatryana, G. Budiman, and S. Y. Shin, "An uplink frequency-time index modulation multiple access for 6g networks," *IEEE Transactions on Vehicular Technology*, vol. 74, no. 5, pp. 7866–7880, 2025.

[26] J. Li, S. Dang, M. Wen, Q. Li, Y. Chen, Y. Huang, and W. Shang, "Index modulation multiple access for 6g communications: Principles, applications, and challenges," *IEEE Network*, vol. 37, no. 1, pp. 52–60, 2023.

[27] T. Van Luong and Y. Ko, "Spread OFDM-IM with precoding matrix and low-complexity detection designs," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11619–11626, 2018.

[28] E. Başar, "OFDM with index modulation using coordinate interleaving," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 381–384, 2015.

[29] A. T. Dogukan and E. Basar, "Super-mode OFDM with index modulation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7353–7362, 2020.

[30] E. Başar, Aygölü, E. Panayırcı, and H. V. Poor, "Orthogonal frequency division multiplexing with index modulation," *IEEE Transactions on Signal Processing*, vol. 61, no. 22, pp. 5536–5549, 2013.

[31] P. Yang, Y. Xiao, M. Xiao, Y. L. Guan, S. Li, and W. Xiang, "Adaptive spatial modulation MIMO based on machine learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 9, pp. 2117–2131, 2019.

[32] Y. Ko and J. Choi, "Unsupervised machine intelligence for automation of multi-dimensional modulation," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1783–1786, 2019.

[33] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6G communications: Challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, pp. 105–118, 2020.

[34] Y. Wang, G. Gui, H. Gacanin, B. Adebisi, H. Sari, and F. Adachi, "Federated learning for automatic modulation classification under class imbalance and varying noise condition," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 1, pp. 86–96, 2022.

[35] Y. Liu, Z. Ma, Z. Yan, Z. Wang, X. Liu, and J. Ma, "Privacy-preserving federated k-means for proactive caching in next generation cellular networks," *Information Sciences*, vol. 521, pp. 14–31, 2020.

[36] C. G. Kang, A. T. Abebe, and J. Choi, "Noma-based grant-free massive access for latency-critical internet of things: A scalable and reliable framework," *IEEE Internet of Things Magazine*, vol. 6, no. 3, pp. 12–18, 2023.

[37] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[38] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, 2017.

[39] J. G. Proakis and M. Salehi, *Digital communications*, vol. 4. McGraw-hill New York, 2001.

[40] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *2013 IEEE 77th vehicular technology conference (VTC Spring)*, pp. 1–5, IEEE, 2013.

[41] M. S. Ali, H. Tabassum, and E. Hossain, "Dynamic user clustering and power allocation for uplink and downlink non-orthogonal multiple access (NOMA) systems," *IEEE access*, vol. 4, pp. 6325–6343, 2016.

[42] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[43] T. P. C. D. Andrade and et al., "The random access procedure in long term evolution networks for the internet of things," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 124–131, 2017.

[44] J. Kim, G. Lee, S. Kim, T. Taleb, S. Choi, and S. Bahk, "Two-step random access for 5g system: Latest trends and challenges," *IEEE Network*, vol. 35, no. 1, pp. 273–279, 2021.

[45] J. Choi, "On fast retrial for two-step random access in mtc," *IEEE Internet of Things Journal*, vol. 8, no. 3, pp. 1428–1436, 2021.

[46] N. Abramson, "The aloha system: Another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*, pp. 281–285, 1970.

[47] R. T. Ma, V. Misra, and D. Rubenstein, "An analysis of generalized slotted-aloha protocols," *IEEE/ACM Transactions on networking*, vol. 17, no. 3, pp. 936–949, 2008.

[48] C. Namislo and et al., "Analysis of mobile radio slotted ALOHA networks," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 4, pp. 583–588, 1984.

[49] D. C. Atabay and et al., "Improving age of information in random access channels," in *IEEE INFOCOM 2020 - IEEE Conf. on Computer Commun. Workshops (INFOCOM WKSHPS)*, pp. 912–917, 2020.

[50] How LTE Stuff Works, "5g nr: 2-step random access procedure." `https://howltestuffworks.blogspot.com/2020/04/5g-nr-2-step-random-access-procedure.html`, 2020. Accessed: 2025-03-19.

[51] S. Khairy, P. Balaprakash, L. X. Cai, and Y. Cheng, "Constrained deep reinforcement learning for energy sustainable multi-uav based random access iot networks with noma," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1101–1115, 2020.

[52] H. S. Jang, H. Lee, and T. Q. Quek, "Deep learning approach for outage-constrained non-orthogonal random access," *IEEE Wireless Communications Letters*, vol. 11, no. 3, pp. 645–649, 2021.

[53] C. Zhang, X. Sun, W. Xia, J. Zhang, H. Zhu, and X. Wang, "Deep learning based double-contention random access for massive machine-type communication," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1794–1807, 2022.

[54] H. V. Hasselt and et al., "Deep reinforcement learning with double q-learning," in *AAAI Conf. on Artificial Intelligence*, vol. 30, 2016.

[55] Y. Liu and et al., "Deep reinforcement learning-based grant-free NOMA optimization for murllc," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1475–1490, 2023.

[56] J. Choi, "Multichannel noma-aloha game with fading," *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4997–5007, 2018.

[57] S. Khairy, P. Balaprakash, L. X. Cai, and H. V. Poor, "Data-driven random access optimization in multi-cell iot networks using noma," *IEEE Transactions on Wireless Communications*, vol. 21, no. 7, pp. 4938–4953, 2021.

[58] 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*, 2024.

[59] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1028–1042, 2020.

[60] U. K. Ganesan, E. Björnson, and E. G. Larsson, "Clustering-based activity detection algorithms for grant-free random access in cell-free massive MIMO," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7520–7530, 2021.

[61] Q. Zhang, J. Zhang, and S. Jin, "Grant-free random access in cell-free massive MIMO systems with UE detection thresholds: A stochastic geometry approach," *IEEE transactions on vehicular technology*, vol. 72, no. 6, pp. 8117–8121, 2023.

[62] R. B. Di Renna and R. C. de Lamare, "Adaptive LLR-based APs selection for grant-free random access in cell-free massive MIMO," in *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 196–201, IEEE, 2022.

[63] E. De Carvalho, E. Björnson, J. H. Sørensen, E. G. Larsson, and P. Popovski, "Random pilot and data access in massive mimo for machine-type communications," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7703–7717, 2017.

[64] J. Ding, D. Qu, H. Jiang, and T. Jiang, "Success probability of grant-free random access with massive mimo," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 506–516, 2018.

[65] R. Kassab, O. Simeone, A. Munari, and F. Clazzer, "Space diversity-based grant-free random access for critical and non-critical iot services," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, IEEE, 2020.

[66] Z. Xu, Z. Zhang, S. Wang, X. Hu, Y. Jia, and B. Ren, "Energy-Constrained Distributed MAC in CR-IoT Networks: A Budgeted Multi-Player Multi-Armed Bandit Approach," *IEEE Transactions on Cognitive Communications and Networking*, 2024.

[67] I. Oueslati, O. Habachi, J.-P. Cances, and V. Meghdadi, "Grant-free access for massive mtc: a low-complexity noma-based framework (loconoma)," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, IEEE, 2024.

[68] A. T. Abebe, J. Lee, M. Rim, and C. G. Kang, "Multi-cell performance of grant-free and non-orthogonal multiple access," in *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pp. 1–6, IEEE, 2017.

[69] A. T. Abusabah, N. M. Balasubramanya, and R. Oliveira, "Performance evaluation of uplink grant-free access networks based on spreading-based

noma," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12953–12965, 2023.

[70] J. Youn, J. Park, S. Kim, S. Ahn, Y. Kim, D. Kim, and S. Cho, "Marl-based access control for grant-free non-orthogonal random access in udn," *IEEE Internet of Things Journal*, 2024.

[71] Z. Shi and J. Liu, "Massive access in 5g and beyond ultra-dense networks: An marl-based nora scheme," *IEEE Transactions on Communications*, vol. 71, no. 4, pp. 2170–2183, 2023.

[72] X. Tang, S. Liu, X. Du, and M. Guizani, "Sparsity-aware intelligent massive random access control for massive mimo networks: A reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, 2024.

[73] W. Wang, W. Yu, C. H. Foh, D. Gao, and Q. Ni, "User scheduling in noma random access using contextual multi-armed bandits," in *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 112–117, IEEE, 2022.

[74] B. Zhao, G. Ren, X. Dong, and H. Zhang, "Distributed q-learning based joint relay selection and access control scheme for iot-oriented satellite terrestrial relay networks," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1901–1905, 2021.

[75] P. Agostini, J.-F. Chamberland, F. Clazzer, J. Dommel, G. Liva, A. Munari, K. Narayanan, Y. Polyanskiy, S. Stanczak, and Z. Utkovski, "Evolution of the 5g new radio two-step random access towards 6g unsourced mac," *arXiv preprint arXiv:2405.03348*, 2024.

[76] L. Zhang, H. Yin, Z. Zhou, S. Roy, and Y. Sun, "Enhancing wifi multiple access performance with federated deep reinforcement learning," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pp. 1–6, IEEE, 2020.

[77] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[78] M. Elsayem, H. Abou-Zeid, A. Afana, and S. Givigi, "Intelligent Resource Allocation for Grant-Free Access: A Reinforcement Learning Approach," *IEEE Networking Letters*, vol. 5, no. 3, pp. 154–158, 2023.

[79] C. Zhang, X. Sun, W. Xia, J. Zhang, H. Zhu, and X. Wang, "Deep learning based double-contention random access for massive machine-type communication," *IEEE Transactions on Wireless Communications*, vol. 22, no. 3, pp. 1794–1807, 2022.

[80] M. V. da Silva, S. Montejo-Sánchez, R. D. Souza, H. Alves, and T. Abrão, "D2d assisted q-learning random access for noma-based mtc networks," *IEEE Access*, vol. 10, pp. 30694–30706, 2022.

[81] R. Huang, V. W. Wong, and R. Schober, "Throughput optimization for grant-free multiple access with multiagent deep reinforcement learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 228–242, 2020.

[82] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE transactions on wireless communications*, vol. 18, no. 1, pp. 310–323, 2018.

[83] E. Nisioti and N. Thomos, "Fast q-learning for improved finite length performance of irregular repetition slotted aloha," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 844–857, 2019.

[84] Z. Shi and et al., "Distributed q-learning-assisted grant-free NORA for massive machine-type communications," in *GLOBECOM 2020 - IEEE Global Commun. Conf.*, pp. 1–5, 2020.

[85] S. K. Sharma and et al., "Collaborative distributed q-learning for rach congestion minimization in cellular iot networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, 2019.

[86] D.-D. Tran, V. N. Ha, and S. Chatzinotas, "Novel reinforcement learning based power control and subchannel selection mechanism for grant-free noma urllc-enabled systems," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*, pp. 1–5, IEEE, 2022.

[87] J. Zhang, X. Tao, H. Wu, N. Zhang, and X. Zhang, "Deep reinforcement learning for throughput improvement of the uplink grant-free noma system," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6369–6379, 2020.

[88] M. Sohaib, J. Jeong, and S.-W. Jeon, "Dynamic multichannel access via multi-agent reinforcement learning: Throughput and fairness guarantees," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 3994–4008, 2021.

[89] X. Wu, Y. Ko, and A. M. Tyrrell, "Decentralized multi-state q-learning for NOMA-ALOHA systems," in *2024 IEEE 35th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1–6, 2024.

[90] X. Wu, Y. Ko, and A. M. Tyrrell, "Distributed multi-agent reinforcement learning for heterogeneous noma-aloha systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 3, pp. 1902–1912, 2025.

[91] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE transactions on cognitive communications and networking*, vol. 4, no. 2, pp. 257–265, 2018.

[92] C. Zhong, Z. Lu, M. C. Gursoy, and S. Velipasalar, "A deep actor-critic reinforcement learning framework for dynamic multichannel access," *IEEE Transactions on Cognitive Communications and Networking*, vol. 5, no. 4, pp. 1125–1139, 2019.

[93] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, "A noma-based q-learning random access method for machine type communications," *IEEE Wireless Communications Letters*, vol. 9, no. 10, pp. 1720–1724, 2020.

[94] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink grant-free access solutions for urllc services in 5g new radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, pp. 607–612, IEEE, 2019.

[95] C. M. Bishop and et al., *Pattern Recognition and Machine Learning*. Springer, 2006.

[96] S. Fujimoto and et al., "Addressing function approximation error in actor-critic methods," in *Int. Conf. on Machine Learning*, pp. 1587–1596, PMLR, 2018.

[97] W. Xu, J. An, C. Huang, L. Gan, and C. Yuen, "Deep reinforcement learning based on location-aware imitation environment for ris-aided mmwave mimo systems," *IEEE Wireless Communications Letters*, vol. 11, no. 7, pp. 1493–1497, 2022.

[98] P. Auer and et al., "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.

[99] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.