IEEE ComSoc IEEE Communications Society | IEEE COMPUTER SOCIETY | IEEE Signal Processing Society | VTS | IEEE Transactions on Machine Learning in Communications and Networking

Received 30 December, 2026; revised 8 May, 2026; accepted 16 June, 2026.

QoS-aware State-Augmented Learning for Wireless Coexistence Parameter Management

Mohammad Reza Fasihi[1], Student Member, IEEE, Brian L. Mark[1], and Omar A. Alotaibi[2], Senior Member, IEEE

[1]Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030, USA
[2]Department of Computer Engineering, King Saud University, Riyadh, 11543, Saudi Arabia

Corresponding author: Brian L. Mark (email: bmark@gmu.edu).

"This research was partially funded by NSF Award No. 2034616 and the NSF IUCRC WISPER, under NSF Award No. 2413168."

ABSTRACT Efficient and fair coexistence in unlicensed spectrum is critical for next-generation wireless networks such as 5G New Radio in Unlicensed Spectrum (NR-U) and Wi-Fi, which contend for shared resources under diverse Quality of Service (QoS) requirements. Existing reinforcement learning (RL) approaches either rely on multi-objective formulations with heuristic weight tuning or primal-dual constrained methods that struggle with stability in dynamic environments. To address these limitations, we propose **QaSAL-CPM**, a novel QoS-aware State-Augmented Learning framework for Coexistence Parameter Management (CPM). Building on the recently introduced concept of state augmentation, QaSAL-CPM embeds dual variables directly into the agent's observation space, enabling real-time responsiveness to constraint violations without retraining or complex penalty tuning. This design achieves strict QoS guarantees for high-priority traffic while maintaining fairness across traffic classes, even under heavy contention. Unlike prior methods, QaSAL-CPM separates offline learning from lightweight online execution, making it practical for real-world deployments such as Ultra-Reliable Low-Latency Communications (URLLC) and the Internet of Things (IoT). Through extensive simulations of 5G NR-U/Wi-Fi coexistence scenarios, we show that QaSAL-CPM enforces 95th-percentile delay compliance and improves policy robustness. These results demonstrate that state-augmented constrained RL offers a scalable and adaptive solution for real-time coexistence optimization in dynamic wireless networks.

INDEX TERMS 5G NR-U, Wi-Fi, coexistence, unlicensed spectrum, reinforcement learning, quality of service (QoS), medium access delay, fairness.

I. Introduction

The rapid growth of wireless connectivity driven by applications such as mobile broadband, industrial IoT, and ultra-reliable low-latency communications (URLLC) has intensified reliance on unlicensed spectrum. These bands are shared by heterogeneous Radio Access Technologies (RATs), notably 5G New Radio in Unlicensed Spectrum (NR-U) and Wi-Fi, which contend for channel access using distinct medium access protocols: Listen-Before-Talk (LBT) for NR-U and CSMA/CA for Wi-Fi. Without effective coordination, this coexistence can lead to severe performance degradation, including increased latency, collisions, and unfair spectrum utilization [1]–[4].

Managing coexistence under these conditions requires adaptive strategies that satisfy diverse Quality of Service (QoS) requirements while maintaining fairness across competing technologies. Prior approaches based on multi-objective reinforcement learning (RL) optimize weighted combinations of performance metrics [5], [6], but lack guarantees for strict QoS compliance and incur significant overhead in weight tuning. Constrained RL methods address this limitation by incorporating QoS constraints into the learning process via Lagrangian relaxation [7]. However, conventional primal-dual formulations struggle in dynamic environments due to unstable convergence and the need for frequent online updates [8], [9].

State augmentation was recently introduced in [8] as a method to embed dual variables directly into the agent's observation space, enabling policies to adapt dynamically to constraint violations. Building on this idea, we apply state-augmented constrained RL to wireless coexistence for the first time. Our proposed framework, referred to as *QoS-aware State-Augmented Learning for Coexistence Parameter Management* (QaSAL-CPM), leverages this approach to enforce QoS constraints tightly while maintaining fairness across traffic classes, even under heavy contention and dynamic network conditions [1]. This capability is critical for real-world deployments such as URLLC and industrial IoT, where strict delay guarantees and fairness are essential for reliable operation. Furthermore, QaSAL-CPM is scalable to emerging coexistence scenarios in the 6 GHz band [12], which will host next-generation Wi-Fi and 5G NR-U systems.

Our main contributions are summarized as follows:

- We analyze the training dynamics, convergence behavior, and deployment complexity of QaSAL-CPM, highlighting the benefits of offline learning and lightweight online execution for practical NR-U and Wi-Fi coexistence control.
- We introduce a bounded and asymmetric violation-scaling mechanism that stabilizes learning under bursty delay dynamics and ensures consistent constraint handling during both training and execution.
- We study the impact of controlling different medium access control (MAC) parameters, including Contention Window (CW), Arbitration Inter-Frame Space Number (AIFSN), and Maximum Channel Occupancy Time (MCOT), both individually and jointly, and identify CW as the dominant lever for delay-sensitive coexistence under the considered traffic conditions.
- We study the impact of enhanced Listen-Before-Talk mechanisms and show that contention-aware access significantly reduces delay spikes and improves tail-delay behavior when combined with QaSAL-CPM.

By explicitly modeling constraint dynamics and leveraging state augmentation, QaSAL-CPM provides a practical and generalizable solution for next-generation wireless networks operating in shared spectrum environments, including future deployments in the 6 GHz band.

The rest of the paper is organized as follows. Section II reviews related work on coexistence strategies and RL-based approaches. Section III presents the system model and problem formulation. Section IV discusses two baseline RL approaches for CPM: multi-objective RL and constrained RL based on Lagrangian relaxation (primal-dual method). Section V develops the proposed QaSAL-CPM framework. Section VI reports extensive simulation results demonstrating the performance of QaSAL-CPM alongside the baseline approaches in a variety of coexistence scenarios. Finally, Section VII concludes the paper.

[1]Our preliminary work on QoS-aware reinforcement learning for the wireless coexistence problem appeared in [10], [11].

II. Related Work

Research on coexistence in unlicensed spectrum spans multiple dimensions, including protocol design, resource allocation, and learning-based optimization. We summarize the most relevant directions below.

A. General Coexistence Strategies

Coexistence of 5G NR-U/LAA and Wi-Fi has been extensively studied in recent literature. Surveys such as [1], [2] provide comprehensive overviews of coexistence challenges in sub-7 GHz unlicensed bands. Detailed evaluations of NR-U specifications and coexistence mechanisms with Wi-Fi are presented in [3]. [13] presents an analytical framework to evaluate the coexistence of LTE and Wi-Fi in unlicensed spectrum, focusing on the impact of priority classes on channel access fairness and performance metrics.

In [14], Oh et al. discuss key enhancements in channel access mechanisms which are designed to ensure fair coexistence with incumbent technologies, as well as the regulatory considerations and ongoing 3GPP standardization efforts that shape NR-U deployment. Emerging opportunities and challenges in the 6 GHz band are discussed in [12], while [15] investigates interference risks between NR-U, Wi-Fi 6E, and incumbent services. In [15], the coexistence of licensed Fixed Service (FS) systems with unlicensed Radio Local Area Networks (RLANs) and 5G NR-U in the 6 GHz band is studied, focusing on potential 5G NR-U and Wi-Fi interference risks on FS systems. A comprehensive analysis of NR-U by examining its transmission protocols, PHY layer design, and coexistence mechanisms in comparison to LTE-LAA is provided in [16]. These studies primarily focus on protocol-level enhancements and spectrum-sharing policies rather than adaptive learning-based control.

B. Bandit-Based Reinforcement Learning

In [17], an MAB-based online learning distributed channel selection algorithm is proposed to enable NR-U users to make optimal channel selections without requiring complete environmental knowledge. An AI-driven framework is introduced in [18] for adaptive sensing threshold selection in shared spectrum environments by leveraging a clustering-based multi-armed bandit algorithm, ensuring efficient coexistence among Wi-Fi, LTE-LAA, and 5G NR-U systems. A novel framework for fair and efficient NR-U coexistence in shared spectrum environments is presented in [19], which formulates the channel access problem as a multi-objective MAB that jointly optimizes spectrum efficiency and fairness. A probabilistic MAB algorithm in introduced in [20] to ensure fair coexistence in heterogeneous networks by optimizing probabilistic

transmission strategies. While MAB-based approaches offer lightweight online adaptation, they operate under a context-free paradigm and cannot capture long-term dependencies or QoS constraints, limiting their applicability in dynamic scenarios.

C. State-Driven and Constrained RL

Deep RL methods have been explored for resource allocation and coexistence optimization in heterogeneous networks. In [21], an RL-based approach is proposed for joint allocation of transmission opportunities and spectral resources in 5G NR-U and Wi-Fi coexistence while ensuring fair spectrum sharing. In [22], a joint base station and resource allocation framework based on deep RL is presented for a heterogeneous network with vehicle-to-everything users. Deep RL-based approaches for energy detection threshold optimization in NR-U and Wi-Fi coexistence systems focusing on downlink URLLC are proposed in [23]. In [24], the challenge of fair coexistence between 5G NR and Wi-Fi while accommodating URLLC traffic is addressed through a mixed priority scheduling mechanism that aims to balance latency constraints and data size. Constrained RL approaches based on Lagrangian relaxation have been proposed for network slicing and resource orchestration in [25]. State-augmented constrained RL was introduced in [8] to improve constraint handling by embedding dual variables into the agent's state. This concept has been applied to Wi-Fi slicing [26], opportunistic routing in wireless networks [27], and wireless resource allocation [9], [28].

Despite the extensive body of work on wireless coexistence and reinforcement learning-based optimization, several important limitations remain. Existing multi-objective RL approaches typically rely on scalarization techniques, which require careful weight tuning and do not guarantee strict satisfaction of QoS requirements. On the other hand, constrained RL methods based on primal-dual formulations often suffer from instability, slow convergence, and limited responsiveness in highly dynamic environments. Moreover, prior works generally do not explicitly incorporate constraint dynamics into the learning process, which is critical for real-time QoS-aware decision making in heterogeneous coexistence scenarios. To address these challenges, this work proposes a state-augmented constrained reinforcement learning framework, where dual variables are embedded directly into the state space. This enables the agent to dynamically adapt to constraint violations while maintaining robust performance, thereby achieving reliable QoS compliance and improved coexistence efficiency.

III. System Model and Problem Formulation

In this section, we describe the system model for wireless coexistence, in particular, 5G NR-U/Wi-Fi coexistence. We discuss the MAC control parameters and the performance metrics of interest for CPM. We then formulate the CPM problem in terms of a *Markov Decision Process* (MDP).

A. System Model

We consider a coexistence scenario where a 5G NR-U network operates alongside a Wi-Fi network in an unlicensed frequency band, as illustrated in Fig. 1. Transmitters from both technologies contend for channel access to perform downlink transmissions. To evaluate worst-case coexistence behavior, we consider a saturated traffic condition in which all transmitters are continuously backlogged and transmit for the maximum allowed transmission duration, as specified by the corresponding standards [29].

The Wi-Fi network employs the IEEE 802.11 Enhanced Distributed Channel Access (EDCA) protocol, which uses binary exponential backoff with a random initial backoff value. EDCA defines four access categories, each with distinct initial contention window (CW) sizes, maximum backoff stages, and transmission opportunity (TXOP) durations. NR-U gNBs adopt the Listen-Before-Talk (LBT) mechanism, which aligns with EDCA principles and supports four priority classes corresponding to the Wi-Fi access categories [30]. For the PHY layer, NR-U gNBs use flexible numerologies and slot-based scheduling, where each transmission can only start at the next slot boundary. If the channel is idle when finished the backoff procedure, the gNB transmits a reservation signal (RS) until the next slot boundary to prevent other nodes from accessing the channel. In the considered framework, a centralized CPM agent communicates coexistence parameter updates to participating NR-U and Wi-Fi transmitters.

To capture heterogeneous traffic requirements, we consider two priority classes:

- PC1: High-priority traffic with strict delay constraints.
- PC3: Low-priority traffic with relaxed delay requirements.

The two traffic classes considered in this work can be interpreted as representative abstractions of 5G service categories. Specifically, PC1 models latency-critical traffic with stringent delay requirements, consistent with URLLC-type services, while PC3 represents throughput-oriented or best-effort traffic, similar to eMBB traffic. This abstraction enables a focused investigation of QoS-aware coexistence while maintaining a tractable learning formulation. While mMTC traffic is not explicitly modeled in this work, it can be incorporated as an additional low-rate traffic class with relaxed latency requirements, which we leave for future investigation.

The coexistence environment is highly dynamic due to asynchronous channel access, collisions, and varying traffic loads. These factors significantly impact latency performance, particularly for PC1 traffic, which may experience unpredictable delays under heavy contention. Our model assumes that NR-U and Wi-Fi transmitters

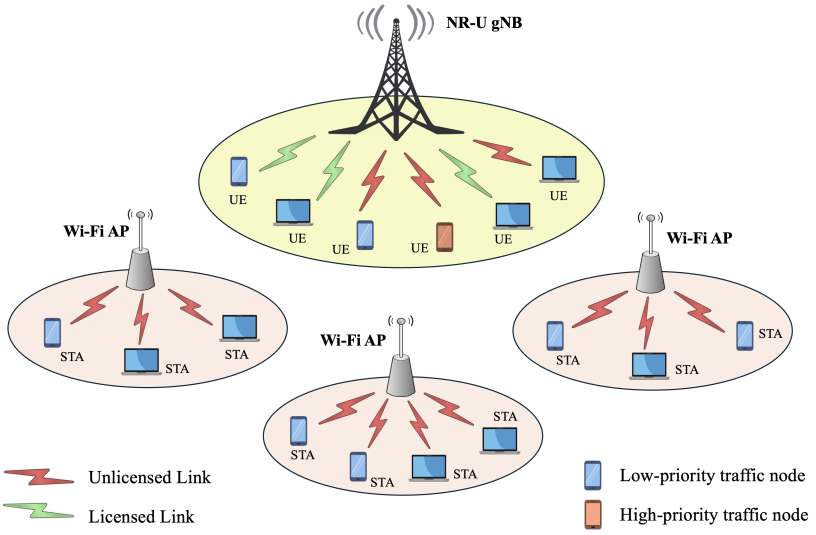


FIGURE 1: 5G NR-U/Wi-Fi coexistence in an unlicensed spectrum with multi-priority traffic transmitters.

have similar channel occupancy durations for successful transmissions and collisions, ensuring fairness in airtime accounting.

Key MAC parameters influencing coexistence include:

- **Contention Window (CW):** Determines the backoff range before channel access attempts;
- **Arbitration Inter-Frame Space Number (AIFSN):** Controls deferment time before backoff starts;
- **Maximum Channel Occupancy Time (MCOT):** Specifies the maximum duration a transmitter can hold the channel after gaining access.

These parameters are selected as the primary control variables because they directly govern channel access priority, contention dynamics, and channel occupation behavior, and thus have the most significant impact on medium access delay and fairness. Other MAC-layer parameters, such as slot time, inter-frame spacing, retransmission rules, and sensing thresholds, are typically fixed by protocol standards or operate at different timescales, and are therefore not included in the control space to maintain a tractable and interpretable learning problem. Dynamic adjustment of the selected parameters is essential for meeting QoS requirements while maintaining fairness across traffic classes.

B. Performance Metrics

We focus on two key performance metrics to evaluate the effectiveness of CPM: *medium access delay* and *airtime fairness*. Medium access delay is a critical factor for ensuring low-latency communication, particularly for high-priority traffic, while Jain's fairness index provides a quantitative measure of how equitably resources are distributed among competing nodes. The two metrics are elaborated below.

1) Medium Access Delay

In coexistence scenarios where 5G NR-U and Wi-Fi share unlicensed spectrum, medium access delay plays a crucial role in ensuring QoS, particularly for high-priority traffic such as URLLC, which demands stringent latency guarantees, often requiring end-to-end delays as low as 1 ms with high reliability. Any excessive delay in accessing the channel or transmitting packets can lead to QoS degradation, packet drops, or violations of reliability constraints, making medium access delay a critical metric in CPM. By dynamically adjusting the contention parameters, the medium access delay can be controlled to ensure that high-priority URLLC traffic meets its latency requirements.

In this paper, we define medium access delay as the time interval between the completion of a node's successful transmission and the start of its next successful transmission. This delay includes backoff delay, which accounts for the time spent in the exponential backoff process before attempting transmission, and contention delay, which encompasses the total time a node waits due to channel occupancy by other devices. Additionally, when a collision occurs, the node must restart the contention process, further increasing the delay before the next successful transmission. Fig. 2 illustrates the calculation of medium access delay across the learning time steps. The cumulative medium access delay, D_t, is experienced by a node up to the end of time step t, accounting for the delays incurred in successful transmissions and collisions.

In an NR-U/Wi-Fi coexistence scenario the delay experienced by a high-priority PC1 transmitter is strongly influenced by the activity of other coexisting nodes, particularly lower-priority transmitters like those in PC3. These nodes often operate with fixed MCOT, such as 8 ms

for PC3, which means that when a PC1 transmitter has to wait for the channel, its delay tends to increase in discrete steps. For instance, the delay could jump to 8 ms, 16 ms, or more, depending on how many consecutive MCOT periods it is blocked. As a result, the delay values do not vary smoothly but appear in distinct levels. This makes it difficult to access performance based on a single delay value, since it might be unusually high due to transient interference or contention. To address this, we use a *smoothed* medium access delay, $\overline{D}_t$, which averages the last few delays (e.g., the last five) observed for the transmitter. This helps reduce the sensitivity to transient spikes caused by contention or collision, and provides a more consistent reflection of ongoing latency performance. As a result, the learning algorithm can make more informed and stable decisions when optimizing coexistence under QoS constraints.

2) Airtime Fairness

While reducing medium access delay is critical for high-priority traffic, it must be balanced with fair resource allocation between PC1 and PC3 transmitters from both the NR-U and Wi-Fi networks. An overly aggressive optimization for PC1 traffic could starve lower-priority traffic, leading to unfair spectrum access and potential performance degradation. To prevent this, fairness should be incorporated as a key metric in CPM optimization. *Jain's Fairness Index* (JFI) is a widely used metric to assess the fairness of resource allocation among competing entities. Considering the PC1 and PC3 transmitters as these entities, we define JFI as follows:

$$\text{JFI} = \frac{(\text{Airtime}_{\text{PC1}} + \text{Airtime}_{\text{PC3}})^2}{2\left(\text{Airtime}^2_{\text{PC1}} + \text{Airtime}^2_{\text{PC3}}\right)}, \tag{1}$$

where $\text{Airtime}_{\text{PC1}}$ and $\text{Airtime}_{\text{PC3}}$ denote, respectively, the total airtime received by PC1 and PC3 traffic over a given total time period. The JFI value ranges from 0.5 to 1, with values closer to 1 indicating a more equitable distribution of airtime. This ensures that neither PC1 nor PC3 traffic classes disproportionately dominates spectrum access.

C. Problem Formulation

The coexistence of 5G NR-U and Wi-Fi in unlicensed spectrum introduces multiple conflicting objectives, such as minimizing delay for high-priority traffic while maintaining fairness across all nodes. These objectives can be modeled within a MDP framework.

Let $\mathcal{S} \subset \mathbb{R}^n$ denote the set of environment states capturing network conditions, including traffic load, collision rates, and delay statistics, where n is the dimension of the state space. At each discrete time step $t \in \mathbb{Z}_{\geq 0}$, the agent observes state $\mathbf{S}_t \in \mathcal{S}$ and selects a coexistence action $\mathbf{a}_t \in \mathcal{A}$, where $\mathcal{A} = \{\mathbf{a} : \mathcal{S} \to \mathbb{R}^a\}$ represents the set of possible CPM decisions (e.g., CW, AIFSN, MCOT adjustments) and $\mathbb{R}^a$ is the a-dimensional action space. The system evolves

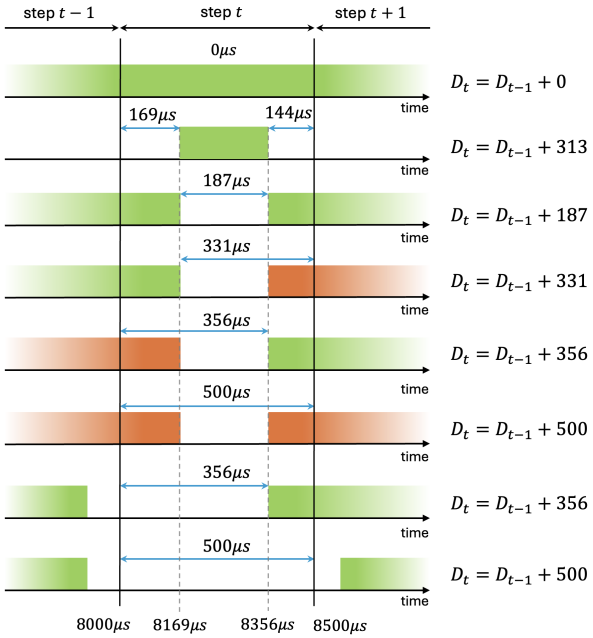


FIGURE 2: Illustration of medium access delay calculation across time steps. Each row represents a different scenario where transmissions (green) and collisions (orange) occur within a given time step.

according to transition probability $p(\mathbf{S}_{t+1}|\mathbf{S}_t, \mathbf{a}_t(\mathbf{S}_t))$. The agent receives a performance vector:

$$\mathbf{f}(\mathbf{S}_t, \mathbf{a}_t(\mathbf{S}_t)) = \left[f_0(\mathbf{S}_t, \mathbf{a}_t(\mathbf{S}_t)), \ldots, f_{m-1}((\mathbf{S}_t, \mathbf{a}_t(\mathbf{S}_t))\right], \tag{2}$$

where f_0 denotes the primary objective (e.g., JFI) and f_i for $i \geq 1$ represent QoS-related metrics (e.g., medium access delay).

The long-term performance under CPM policy π is defined by the ergodic average:

$$\tilde{V}_i(\pi) = \frac{1}{T}\sum_{t=0}^{T-1} f_i(\mathbf{S}_t, \mathbf{a}_t(\mathbf{S}_t)), \quad i = 0, \ldots, m-1. \tag{3}$$

The CPM problem can be expressed in a multi-objective formulation as

$$\max_{\pi} \left[\tilde{V}_0(\pi), \tilde{V}_1(\pi), \ldots, \tilde{V}_{m-1}(\pi)\right]. \tag{4}$$

In the context of 5G NR-U and Wi-Fi coexistence, the network state $\mathbf{S}_t$ is designed to comprehensively capture the dynamic environment on an unlicensed spectrum. It encodes critical metrics that influence decision-making, including network performance and resource utilization such as the average and smoothed medium access delay of PC1 transmitter, collision rates, channel utilization ratio, rate of QoS violations, and JFI. Additionally, it tracks trends in delay variation and short-term collision statistics to provide insights into ongoing network conditions.

The CPM actions $\mathbf{a}_t = \mathbf{a}(\mathbf{S}_t)$ depends on the MAC parameter that is controlled. When controlling the CW, $\mathbf{a}_t = \{a_{i,t}\}, i \in \{\text{PC1}, \text{PC3}\}$ and $a_{i,t} \in \{0, 1, \ldots, 6\}$. The maxi-

mum contention window for priority class i, this, is

$$\mathrm{CW}_{\max,i} = 2^{a_{i,t}+b_i} - 1, \quad \text{with } b_{\mathrm{PC1}} = 0,\ b_{\mathrm{PC3}} = 4,$$

which directly affects backoff duration and channel access aggressiveness. For AIFSN control, the action selects from predefined sets $\mathrm{AIFS}_{\mathrm{PC1}} \in [1,2,3]$ and $\mathrm{AIFS}_{\mathrm{PC3}} \in [2,\ldots,8]$, tuning the deferment period before backoff and thus the relative priority among transmitters. When MCOT is the control parameter, the agent selects $\mathrm{MCOT}_{\mathrm{PC1}} \in [1, 1.5, \ldots, 3]$ ms and $\mathrm{MCOT}_{\mathrm{PC3}} \in [2,3,\ldots,8]$ ms, which governs the maximum duration a transmitter may occupy the channel after winning access. These actions collectively determine how aggressively each node contends for the medium and how long it holds the channel, jointly shaping delay, collision probability, and fairness. Moreover, the objectives are designed as

$$f_0(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t)) = \mathrm{JFI}_t, \qquad f_1(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t)) = \overline{D}_{\mathrm{PC1},t},$$

where JFI_t measures airtime fairness (PC1/PC3) and $\overline{D}_{\mathrm{PC1},t}$ is the smoothed medium-access delay for PC1. We define a constraint for medium access delay of PC1 in terms of a predefined threshold D_{th}.

IV. Reinforcement Learning for CPM

In this section, we discuss two baseline reinforcement learning approaches for CPM, which we will compare with our proposed QaSAL-CPM.

A. Baseline Approach 1: Multi-Objective RL

Prior works have addressed CPM using multi-objective RL (MORL) by scalarizing performance metrics into a weighted sum [6]:

$$f_{\mathrm{scalar}}(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t)) = \sum_{i=0}^{m-1} w_i f_i(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t)), \quad \sum_{i=0}^{m-1} w_i = 1. \tag{5}$$

We apply this approach to the CPM problem . Optimization of $f_{\mathrm{scalar}}(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t))$ in (5) involves an infinite-dimensional search over the space of CPM decisions $\mathbf{a}(\mathbf{S})$ for any given network state $\mathbf{S}$, making direct optimization impractical. To address this, a parameterized approach for the CPM policy can be adopted by replacing $\mathbf{a}(\mathbf{S})$ with $\mathbf{a}(\mathbf{S};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ denotes a finite-dimensional set of learning parameters, and maximization is performed iteratively over the set $\boldsymbol{\theta}$. This leads to the *parameterized multi-objective CPM* problem

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{T} \sum_{t=0}^{T-1} f_{\mathrm{scalar}}(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})), \tag{6}$$

which can be solved through multi-objective RL algorithms. We call this approach MORL-CPM.

In the 5G NR-U/Wi-Fi coexistence scenario, the agent jointly optimizes medium access delay and airtime fairness through a weighted objective function. We normalize $\overline{D}_{\mathrm{PC1},t}$ by $\mathrm{D}_{\max}$ to normalize the delay term to $[0,1]$. This is critical for combining it with JFI, which is inherently bounded in the same range. Utilizing the linear scalarization approach, we define the weight vector as $\boldsymbol{w} = \{1-\alpha, \alpha\}$ according to (5). Therefore, the *multi-objective CPM* problem in (6) can be formulated as

$$\max_{\{\mathbf{a}_t\}_{t=0}^{T-1}} \frac{1}{T} \sum_{t=0}^{T-1} \left\{ (1-\alpha)\mathrm{JFI}_t + \alpha \left(1 - \frac{\overline{D}_{\mathrm{PC1},t}}{\mathrm{D}_{\max}} \right) \right\}. \tag{7}$$

The trade-off parameter $\alpha \in [0,1]$ balances the relative importance of the objectives. Higher values of α put more emphasize on reducing $\overline{D}_{\mathrm{PC1},t}$, whereas lower values put more weight on JFI.

Although linear scalarization is computationally efficient and easy to implement, it assumes that the trade-offs between objectives can be accurately captured by the predefined weights in (5). In practice, selecting appropriate weight values can be challenging, especially when objectives exhibit nonlinear dependencies. These limitations motivate the constrained RL formulation adopted in this work.

B. Baseline Approach 2: Constrained RL with Primal-Dual

As an alternative to multi-objective RL, the QoS requirements can be explicitly incorporated into a constrained RL problem. To formulate the CPM problem within constrained RL, we designate the first objective function in (2), $f_0(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t))$, as the *primary* objective to maximize, while treating the remaining objectives as *QoS constraints* that must be satisfied. The goal of the CPM problem is to determine the optimal CPM decision $\mathbf{a}(\mathbf{S}_t)$ vector for any given network state $\mathbf{S}_t \in \mathscr{S}$, such that the primary objective is optimized while ensuring compliance with QoS constraints. Accordingly, the generic *parameterized constrained CPM* problem is defined as

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \frac{1}{T} \sum_{t=0}^{T-1} f_0(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})), \tag{8a}$$

$$\text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} f_i(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})) \geq \mathrm{c}_i, \quad i = 1,\ldots,m-1, \tag{8b}$$

where the constant $\mathrm{c}_i \in \mathbb{R}$ represents the threshold value for the i-th objective function. In this paper, we develop a learning algorithm to solve (8) for any given coexistence environment state $\mathbf{S}_t \in \mathscr{S}$. In the 5G NR-U/Wi-Fi coexistence scenario, the *parameterized constrained CPM* problem is as follows:

$$\max_{\{\mathbf{a}_t\}_{t=0}^{T-1}} \quad \frac{1}{T} \sum_{t=0}^{T-1} \mathrm{JFI}_t, \tag{9a}$$

$$\text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} \overline{D}_{\mathrm{PC1},t} \leq D_{\mathrm{th}}. \tag{9b}$$

A customary approach to solve (8) is to consider a penalized version in the *Lagrangian dual* domain and solving it through the *primal-dual* approach. We apply this approach introduced in [25] to the CPM problem. Formally, we introduce dual variables $\boldsymbol{\lambda} \in \mathbb{R}_+^{m-1}$ associated

with the constraints in (8b) and define the Lagrangian

$$\mathscr{L}(\boldsymbol{\lambda};\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=0}^{T-1} f_0(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})) + \sum_{i=1}^{m-1} \lambda_i \left(\left(\frac{1}{T}\sum_{t=0}^{T-1} f_i(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})) \right) - \mathrm{c}_i \right). \quad (10)$$

The Lagrangian in (10) should be maximized over $\boldsymbol{\theta}$, while subsequently minimizing over the dual variables $\boldsymbol{\lambda}$, i.e.,

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^{m-1}} \quad \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \mathscr{L}(\boldsymbol{\lambda};\boldsymbol{\theta}). \quad (11)$$

The advantage of replacing the objective in (8) with the Lagrangian in (10) is that the latter can be optimized using any parameterized learning framework, such as standard RL algorithms. One limitation of (11) is the ambiguity in determining suitable values for the dual variables. The optimal choice for $\boldsymbol{\lambda}$ depends on the transition probability $p(\mathbf{S}_{t+1}|\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta}))$, which is typically unknown. This challenge can be circumvented by dynamically adjusting the dual variables $\boldsymbol{\lambda}$. To accomplish this, we define an iteration index $k \in \{0, 1, \ldots, \lfloor T/T_0 \rfloor - 1\}$, where T_0 is the duration of one epoch, which is the number of time steps between successive updates of the model parameters. Also, we introduce the learning rates $\eta_{\lambda_i} \in \mathbb{R}_+$ corresponding to each λ_i. The model parameters $\boldsymbol{\theta}_k$ and dual variables $\lambda_{i,k}$ are updated iteratively according to equations (12) and (13).

During the k-th epoch, the agent collects samples over a window of length T_0 and uses them to compute the constraint violation. The primal update in (12) seeks to maximize the Lagrangian, which includes the main objective f_0 and a penalty term weighted by the current dual variable $\lambda_{i,k}$ and the degree of constraint violations. The dual update in (13) adjusts each $\lambda_{i,k}$ based on the extent to which its corresponding constraint c_i is violated. If the objective f_i falls below the c_i, the dual variable increases proportionally to the violation, encouraging the agent to reduce the violation in future updates. The $[\cdot]^+$ operator ensures that the dual variables remain non-negative. The update rate η_{λ_i} can be tuned individually for each constraint to control how aggressively violations are penalized.

One of the important limitations of the primal-dual approach is that it only guarantees convergence to a feasible and near-optimal solution in the long run, i.e., as the total operation time T becomes very large, theoretically approaching infinity. In practice, this means we cannot simply stop the algorithm after a finite number of iterations and claim that the solution it has reached is close to optimal [8]. The performance and feasibility guarantees only hold in the limit, which may not be practical for systems that need to make decisions in real-time or within limited time frames.

Another challenge with this method is that optimizing the Lagrangian function at each iteration requires information about future network states. For example, when the algorithm reaches the beginning of the k-th epoch (at time step $t = kT_0$), the optimization process needs access to all network states from $t = kT_0$ to $t = (k+1)T_0 - 1$. However, in an online or real-time setting, this future information is not available, making it impossible to perform the optimization exactly as required [8]. While this may be manageable in offline training where future states can be simulated or assumed, it poses a serious limitation during actual deployment.

V. Proposed QaSAL-CPM Framework

In this section, we describe our proposed QaSAL-CPM framework, which is based on state-augmented constrained RL. We develop deep RL algorithms for training and execution of QaSAL-CPM based on a *Double Deep Q-Network* (DDQN) architecture. We further develop a procedure for bounded violation scaling to improve learning stability. We also discuss the computational complexity of QaSAL-CPM at deployment.

A. State-Augmented Constrained RL

The recent *state augmentation* approach provides a practical and scalable alternative to conventional primal-dual methods [8], [9]. Its key idea is to treat constraint satisfaction as part of the system state by embedding the dual variables directly into the agent's observation space. This transforms the constrained optimization problem into a state-augmented MDP, enabling the agent to learn a single policy that explicitly accounts for constraint dynamics and adapts to changes in constraint pressure over time.

From a theoretical perspective, classical constrained MDPs may require randomized mixtures of multiple deterministic policies to satisfy constraints. In contrast, state-augmented approaches avoid explicit policy randomization by learning a unified policy conditioned on the dual variables. As these dual variables evolve based on observed constraint violations, the learned policy exhibits an *implicit policy switching* behavior, where the agent

$$\boldsymbol{\theta}_k = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left[\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1} f_0(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})) + \sum_{i=1}^{m-1} \lambda_{i,k} \left(\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1} \left(f_i(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta})) - \mathrm{c}_i \right) \right) \right], \quad (12)$$

$$\lambda_{i,k+1} = \left[\lambda_{i,k} - \frac{\eta_{\lambda_i}}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1} \left(f_i(\mathbf{S}_t, \mathbf{a}(\mathbf{S}_t;\boldsymbol{\theta}_k)) - \mathrm{c}_i \right) \right]^+, \quad i = 1, \ldots, m-1. \quad (13)$$

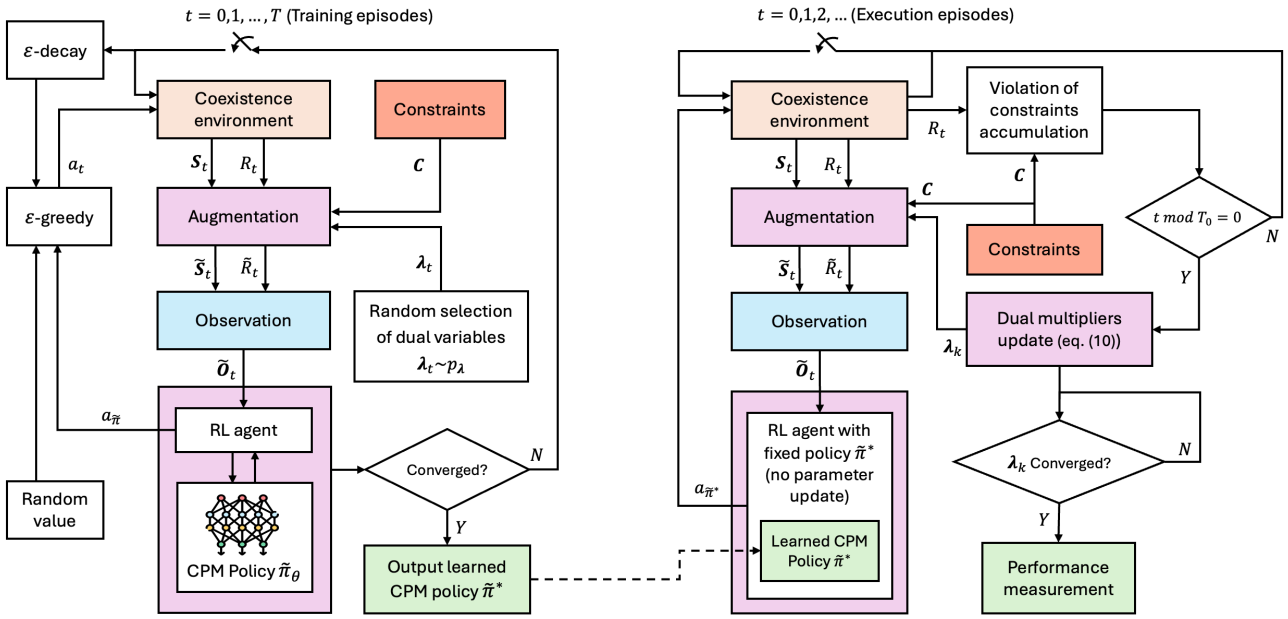


FIGURE 3: Flowchart of the proposed QaSAL-CPM framework. The offline training phase (left) employs state-augmented reinforcement learning with randomized dual variables to improve robustness across different constraint regimes. The online execution phase (right) deploys a fixed policy and performs periodic monitoring and updates of the dual variable. The index k denotes the update interval for the dual variable during execution.

naturally transitions between different operating regimes depending on the current level of constraint pressure.

In the proposed framework, dual variables are updated online and embedded in the state, forming a closed feedback loop between constraint violations and action selection. When violations persist, the corresponding dual variable increases, steering the policy toward more conservative actions that prioritize constraint satisfaction. Conversely, when the constraint is satisfied, the dual variable decreases, allowing the policy to relax and focus on optimizing the primary objective. This interaction induces a *state-driven adaptive behavior*, where policy adjustments emerge continuously rather than through discrete switching. As a result, the learned policy can be interpreted as a continuous manifold indexed by the dual state, rather than a fixed set of candidate policies.

Unlike classical primal-dual methods, where constraints affect learning only through Lagrangian penalties in the reward, state augmentation allows the policy to directly observe the constraint state. This explicit awareness improves responsiveness and stability in time-varying environments, as the agent can proactively adjust its actions rather than reacting indirectly to delayed reward signals. When combined with gradient-based dual updates, the resulting policy remains feasible and near-optimal even in non-convex settings [8].

These properties are especially important for CPM, where coexistence dynamics are driven by heterogeneous transmitters and rapidly changing QoS requirements. By aligning coexistence parameter decisions with the real-time status of QoS constraints, state augmentation helps maintain critical guarantees, such as bounded medium access delay for high-priority traffic, in highly dynamic environments.

B. QaSAL-CPM

In light of the state augmentation approach, we propose the QaSAL-CPM framework. Let us consider state $\mathbf{S}_t$ at time step t of the k-th epoch. Augmentation of the dual variables $\boldsymbol{\lambda}_k$ into the state space results in a augmented state $\tilde{\mathbf{S}}_t = (\mathbf{S}_t, \boldsymbol{\lambda}_k)$. CPM action is $\mathbf{a}(\tilde{\mathbf{S}}; \tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}} \in \tilde{\boldsymbol{\Theta}}$ denotes the set of parameters of the state-augmented CPM policy. We define the augmented version of the Lagrangian in (10) as

$$\mathscr{L}(\boldsymbol{\lambda}; \tilde{\boldsymbol{\theta}}) = \frac{1}{T}\sum_{t=0}^{T-1} f_0(\tilde{\mathbf{S}}_t, \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}})) + \sum_{i=1}^{m-1} \lambda_i \left(\left(\frac{1}{T}\sum_{t=0}^{T-1} f_i(\tilde{\mathbf{S}}_t, \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}})) \right) - \mathrm{c}_i \right). \quad (14)$$

Considering a probability distribution $p_{\boldsymbol{\lambda}}$ for the dual variables $\boldsymbol{\lambda}$, we define the optimal *state-augmented CPM* policy which maximizes the expected augmented Lagrangian over this distribution as

$$\tilde{\boldsymbol{\theta}}^* = \arg\max_{\tilde{\boldsymbol{\theta}} \in \tilde{\boldsymbol{\Theta}}} \mathbb{E}_{\boldsymbol{\lambda} \sim p_{\boldsymbol{\lambda}}} \{ \mathscr{L}(\boldsymbol{\lambda}; \tilde{\boldsymbol{\theta}}) \}. \quad (15)$$

Using the augmented policy parameterized by (15), we can obtain the CPM decisions $\mathbf{a}(\tilde{\mathbf{S}}; \tilde{\boldsymbol{\theta}})$ which maximize the

Lagrangian corresponding to the dual variables $\boldsymbol{\lambda}$ at each iteration k. The dual variable update equation in (13) is replaced with its augmented version:

$$\lambda_{i,k+1} = \left[\lambda_{i,k} - \frac{\eta_{\lambda_i}}{T_0} \sum_{t=kT_0}^{(k+1)T_0-1} \left(f_i(\tilde{\mathbf{S}}_t, \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}}_k^*)) - \mathrm{c}_i\right)\right]^+, \quad (16)$$

where $i = 1, \ldots, m-1$. Note that the optimal model parameters $\tilde{\boldsymbol{\theta}}^*$ are directly utilized in (16), which mitigates the challenge posed in (12), where computing and storing the optimal set of parameters for every possible configuration of dual variables would be required. This was, in fact, the primary motivation for adopting the state-augmented parameterization approach. A flowchart depicting the QaSAL-CPM framework is given in Fig. 3.

C. Deep Reinforcement Learning for QaSAL-CPM

We apply a Double Deep Q-Network (DDQN) [31] to solve (15), where the maximization of the augmented Lagrangian is approximated through parameterized value learning. DDQN decouples action selection and action evaluation during target computation by using the online Q-network to select the action and the target network to evaluate its value. This design mitigates overestimation bias and improves learning stability, particularly in complex and dynamic environments.

The training phase begins by initializing the online and target network parameters, along with an experience replay buffer. At the start of each episode, the environment is reset and a set of dual variables associated with the QoS constraints is sampled from a predefined distribution. In this work, we adopt a uniform distribution the interval $(0, \lambda_{\max})$:

$$\lambda_i \sim \mathscr{U}(0, \lambda_{\max}), \quad \forall i \in \{1, \ldots, m\}. \quad (17)$$

This sampling strategy exposes the agent to a diverse range of constraint conditions, enabling it to learn a policy that generalizes across different levels of constraint pressure. The parameter $\lambda_{\max}$ plays a critical role in shaping the learned behavior. A small value may limit the agent's ability to enforce constraints effectively, while excessively large values may bias the policy toward overly conservative actions, potentially degrading the primary objective (e.g., airtime fairness). In our implementation, the parameter $\lambda_{\max}$ is selected empirically to achieve a balance between effective constraint enforcement and maintaining performance of the primary objective.

The sampled dual variables are concatenated with the environment state to form an augmented state representation. Based on this augmented state, the agent selects a coexistence action that balances the primary performance objective and constraint satisfaction. After executing the action, the environment returns an augmented reward derived from the Lagrangian formulation, and the agent transitions to a new augmented state. Each transition is stored in the replay buffer for subsequent training.

During learning, the agent samples mini-batches from the replay memory and applies the DDQN update rule to compute stable target Q-values. The action is selected using the online network, while its value is estimated using the target network. The resulting temporal-difference error defines the augmented Lagrangian loss, which is minimized via gradient descent to update the network parameters. The target network is periodically synchronized with the online network to maintain training stability. This process continues over multiple episodes until convergence, yielding a policy capable of making real-time CPM decisions under QoS constraints.

In the execution phase, the agent observes the current network state and augments it with the current dual variables before selecting an action using the trained policy. The policy parameters remain fixed during deployment. Every T_0 steps, the dual variables are updated based on the average constraint violations observed over the preceding interval according to (16). This mechanism enables the deployed policy to adapt its behavior to evolving constraint pressure while preserving lightweight online operation. The overall training and execution procedures of QaSAL-CPM are summarized in Algorithms 1 and 2, respectively.

When jointly optimizing multiple MAC parameters, we employ a *multi-head* DDQN architecture. In this design, a shared feature extractor first maps the augmented state $\mathbf{S}_t$ into a latent representation, which is then processed by multiple output heads. Each head corresponds to a specific MAC parameter and outputs a vector of Q-values over the discrete action set associated with that parameter. The joint CPM action at time t is given by

$$\mathbf{a}_t = \left(a_t^{\mathrm{CW}}, a_t^{\mathrm{AIFSN}}, a_t^{\mathrm{MCOT}}\right)$$

where each component is selected independently using an ϵ-greedy policy based on the Q-values of its corresponding head.

The experience reply buffer stores transitions of the form $(\mathbf{S}_t, \mathbf{a}_t, r_t, \mathbf{S}_{t+1})$. During training, temporal-difference targets for all heads are computed using the standard DDQN mechanism, and the same augmented reward is used to update each head simultaneously. This factorized multi-head representation avoids enumerating the full Cartesian product of MAC parameter actions, which would otherwise result in a prohibitively large discrete action space. At the same time, it enables coordinated learning across parameters through shared state representations, allowing the agent to learn coherent joint CPM policies that balance access aggressiveness and channel occupancy, thereby improving both delay performance and airtime fairness.

Algorithm 1 Training phase of QaSAL-CPM with DDQN.

Input: Number of training episodes N, number of time steps T, QoS constraints $\mathbf{c}$, target network update frequency τ, batch size B, replay memory capacity M, primal learning rate $\eta_{\tilde{\boldsymbol{\theta}}} \in (0,1]$, discount factor $\gamma \in (0,1]$.
Output: Optimal model parameters $\tilde{\boldsymbol{\theta}}^*$.

1: Initialize model parameters $\tilde{\boldsymbol{\theta}}_0$, target parameters $\tilde{\boldsymbol{\theta}}^{\text{target}} \leftarrow \tilde{\boldsymbol{\theta}}_0$, and experience replay memory $\mathscr{D} \leftarrow \varnothing$.
2: **for** $n = 0, 1, \ldots, N-1$ **do**
3: Observe the initial network state $\mathbf{S}_0$.
4: Randomly sample $\boldsymbol{\lambda}_n = \{\lambda_{i,n} \sim p_\lambda\}_{i=1}^{m-1}$.
5: **for** $t = 0, 1, \ldots, T-1$ **do**
a: Augment network state $\tilde{\mathbf{S}}_t = (\mathbf{S}_t, \boldsymbol{\lambda}_n)$.
b: Generate CPM decision $\mathbf{a}_t = \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}}_n)$.
c: Calculate main reward $r_t = f_0(\tilde{\mathbf{S}}_t, \mathbf{a}_t)$.
d: Calculate violations $v_t = \sum_{i=1}^{m-1} \lambda_{i,n} \left(f_i(\tilde{\mathbf{S}}_t, \mathbf{a}_t) - \mathrm{c}_i\right)$.
e: Observe and augment next state $\tilde{\mathbf{S}}_{\mathbf{t+1}} = (\mathbf{S}_{t+1}, \boldsymbol{\lambda}_n)$.
f: Store transition $\left(\tilde{\mathbf{S}}_t, \mathbf{a}_t, r_t, v_t, \tilde{\mathbf{S}}_{t+1}\right)$ in $\mathscr{D}$.
g: Sample minibatch $\left\{\left(\tilde{\mathbf{S}}_j, \mathbf{a}_j, r_j, v_j, \tilde{\mathbf{S}}_{j+1}\right)\right\}_{j=1}^{B}$ from $\mathscr{D}$.
h: Calculate $\left\{y_j = r_j + \gamma \max_{\mathbf{a}'} Q(\tilde{\mathbf{S}}_{j+1}, \mathbf{a}' \mid \tilde{\boldsymbol{\theta}}_n^{\text{target}})\right\}_{j=1}^{B}$.
i: Compute the augmented Lagrangian loss:
$$L(\tilde{\boldsymbol{\theta}}_n) = \tfrac{1}{B} \textstyle\sum_{j=1}^{B} \left(y_j + v_j - Q(\tilde{\mathbf{S}}_j, \mathbf{a}_j \mid \tilde{\boldsymbol{\theta}}_n)\right)^2.$$
j: Update model parameters via gradient descent:
$$\tilde{\boldsymbol{\theta}}_{n+1} \leftarrow \tilde{\boldsymbol{\theta}}_n - \eta_{\tilde{\boldsymbol{\theta}}} \nabla_{\tilde{\boldsymbol{\theta}}_n} L(\tilde{\boldsymbol{\theta}}_n).$$
k: Update $\tilde{\boldsymbol{\theta}}^{\text{target}} \leftarrow \tilde{\boldsymbol{\theta}}_{n+1}$ every τ steps.
6: **end for**
7: **end for**
8: $\tilde{\boldsymbol{\theta}}^* \leftarrow \tilde{\boldsymbol{\theta}}_N$

Algorithm 2 Execution phase of QaSAL-CPM.

Input: Optimal model parameters $\tilde{\boldsymbol{\theta}}^*$, dual variables update rates $\boldsymbol{\eta}_\lambda$, QoS constraints $\mathbf{c}$, epoch duration T_0.
Output: Sequence of CPM decisions $\{\mathbf{a}_t; t = 0, 1, \ldots\}$.

1: Initialize: $\boldsymbol{\lambda}_0 \leftarrow \mathbf{0}$, $k \leftarrow 0$.
2: **for** $t = 0, 1, \ldots$ **do**
a: Observe and augment the state $\tilde{\mathbf{S}}_{\mathbf{t}} = (\mathbf{S}_t, \boldsymbol{\lambda}_k)$.
b: Generate CPM decision $\mathbf{a}_t = \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}}^*)$.
c: **if** $t+1 \mod T_0 = 0$ **then**
Update the dual variables $\{\lambda_i\}_{i=1}^{m-1}$ as:
$$\lambda_{i,k+1} = \left[\lambda_{i,k} - \tfrac{\eta_{\lambda_i}}{T_0} \textstyle\sum_{t=kT_0}^{(k+1)T_0-1} \left(f_i(\tilde{\mathbf{S}}_t, \mathbf{a}_t) - \mathrm{c}_i\right)\right]^+.$$
$k \leftarrow k+1$.
d: **end if**
3: **end for**

While we adopt a fixed sampling distribution for dual variables during training, more advanced strategies could further improve learning efficiency. In particular, sampling dual variables from trajectories collected during training may better capture the distribution of relevant constraint dynamics (see [28]). Exploring such adaptive sampling schemes in the QaSAL framework is an interesting direction for future work.

D. Bounded Asymmetric Violation Scaling

In constrained reinforcement learning, a key challenge is transforming noisy and time-varying constraint signals into stable learning targets while optimizing the primary objective. This challenge is especially acute in NR-U and Wi-Fi coexistence, where decentralized access, collisions, and reservation signaling produce bursty delay behavior and a narrow feasible operating region. As a result, even small fluctuations in constraint violations can destabilize learning or lead to oscillatory policies. State augmentation mitigates this issue by explicitly exposing constraint dynamics to the policy through the dual variables, enabling the agent to reason directly about constraint pressure when making decisions.

1) Bounded Asymmetric Constraint Violation Scaling

Learning stability depends strongly on how constraint violations are represented. Instead of using smooth nonlinear functions, we employ a bounded, signed, and asymmetric scaling that preserves linear sensitivity near the constraint boundary while preventing extreme values from dominating learning. Let

$$e_{\text{raw}} \triangleq f_i(\tilde{\mathbf{S}}_t, \mathbf{a}(\tilde{\mathbf{S}}_t; \tilde{\boldsymbol{\theta}})) - c_i,$$

denote the raw constraint deviation, where positive values indicate constraint satisfaction and negative values indicate violations. We first clip this signal to a bounded range to limit the influence of rare but severe outliers:

$$e_{\text{clip}} = \text{clip}(e_{\text{raw}}, -c_{\max}, c_{\max}),$$

where

$$\text{clip}(x, a, b) \triangleq \begin{cases} a, & \text{if } x \le a \\ b, & \text{if } x \ge b \\ x, & \text{otherwise.} \end{cases}$$

We then apply asymmetric linear scaling:

$$e_{\text{scaled}} = \text{scale}(e_{\text{clip}}),$$

where

$$\text{scale}(x) \triangleq \begin{cases} \kappa x, & \text{if } x \ge 0 \\ x, & \text{otherwise,} \end{cases}$$

and $0 < \kappa < 1$ decreases sensitivity to constraint satisfaction while penalizing the violations. This asymmetric treatment reflects the fact that violations should be corrected aggressively, whereas satisfying the constraint should not introduce unnecessary learning pressure. The RL agent receives only the negative component of the

scaled signal as a cost, ensuring that learning focuses on avoiding violations rather than exploiting slack above the constraint. At the same time, the full signed scaled signal is retained in the augmented state through the dual variable. This allows the policy to anticipate whether the system is approaching a violation boundary and adjust its actions proactively.

2) Consistent Dual Update with Scaled Violations

To avoid inconsistencies between training and execution, the dual variable is updated using the same scaled signal. Every T_0 steps, the dual variable is updated as

$$\lambda \leftarrow \left[\lambda - \eta_\lambda \, \overline{e_{\text{scaled}}}\right]^+, \qquad \lambda \in [0, \lambda_{\max}],$$

where $\overline{e_{\text{scaled}}}$ denotes a moving average of the scaled deviation. Using identical scaling and smoothing for both learning and dual updates prevents feedback mismatch that can lead to oscillations or slow convergence. The update parameters (η_λ, T_0) are selected based on contention dynamics, with heavier traffic loads favoring more frequent or stronger updates. If the dual variable saturates at $\lambda_{\max}$ without eliminating violations, $\lambda_{\max}$ can be increased. The bounded and asymmetric scaling provides a well-conditioned constraint signal that preserves sensitivity near the violation boundary while maintaining stability under bursty traffic. Combined with state augmentation, it enables QaSAL-CPM to enforce QoS constraints reliably without complex reward shaping or fragile penalty tuning.

E. Computational Complexity at Deployment

QaSAL-CPM enhances the agent's awareness of system dynamics by embedding the dual variables associated with QoS constraints directly into the state space. This allows the agent to observe the current level of constraint pressure explicitly and account for it when selecting coexistence parameters. As a result, policy learning and deployment are naturally separated. All policy optimization and parameter learning are carried out offline during training, after which the learned CPM policy is fixed and deployed for online operation. During execution, the agent does not perform gradient-based optimization or update model parameters; it simply evaluates the trained policy through a forward pass based on the current observed state.

By observing the dual variables as part of the state, the agent can also anticipate the evolution of the system. The dual variable provides a direct indication of whether the system is approaching a constraint violation, enabling the agent to proactively adjust its actions to prevent future violations. This proactive behavior improves QoS compliance and reduces oscillatory effects that commonly arise when constraint handling affects decisions only indirectly through reward penalties.

TABLE 1: Simulation Setup

Parameter	Value
Interaction time	10,000 episodes
Episode duration	100 steps
Step duration	2.5 ms
Discount factor	0.99
Replay buffer size	100,000
Range of ϵ	1.0 to 0.01
DDQN learning rate	10^{-5}
Batch size	256
DDQN Hidden layers	3×256
$\lambda_{\max}, T_0, \eta_\lambda, \kappa$	5.0, 5, 0.05, 0.1
D_{th}	2 ms

From a computational standpoint, the online complexity of QaSAL-CPM is therefore limited to lightweight policy inference and simple constraint monitoring. In contrast, classical primal-dual constrained reinforcement learning methods interleave dual updates with ongoing policy learning, requiring the policy to continuously adapt through online gradient updates as the optimization objective changes. This coupling increases runtime overhead and demands careful time-scale tuning to maintain stability, which can be challenging in fast, real-time environments such as MAC-layer control. By decoupling learning from execution and confining optimization to the offline phase, QaSAL-CPM achieves robust QoS-aware control with low deployment complexity.

VI. Simulation Results and Analysis

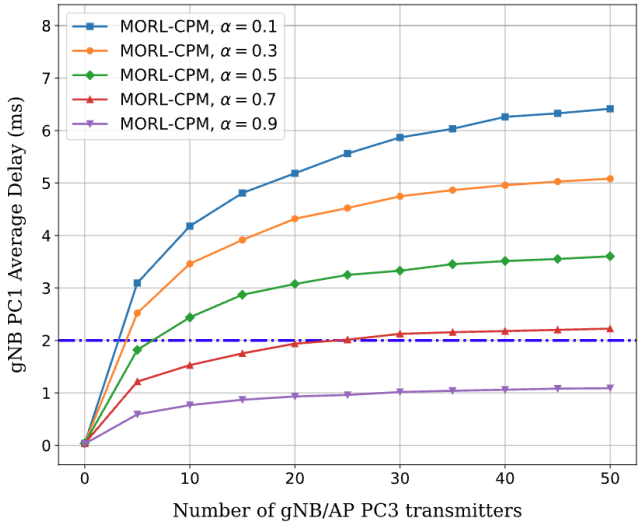
A. MORL-CPM (Scenario 1)

In this section, we evaluate the performance of QaSAL-CPM and the baseline methods MORL-CPM and Primal-Dual (i.e., constrained RL without state augmentation) across a sequence of increasingly complex coexistence scenarios. We implemented a Python-based simulation environment that models the MAC-layer behavior of 5G NR-U and Wi-Fi under saturated traffic conditions[2]. Two representative scenarios are considered:

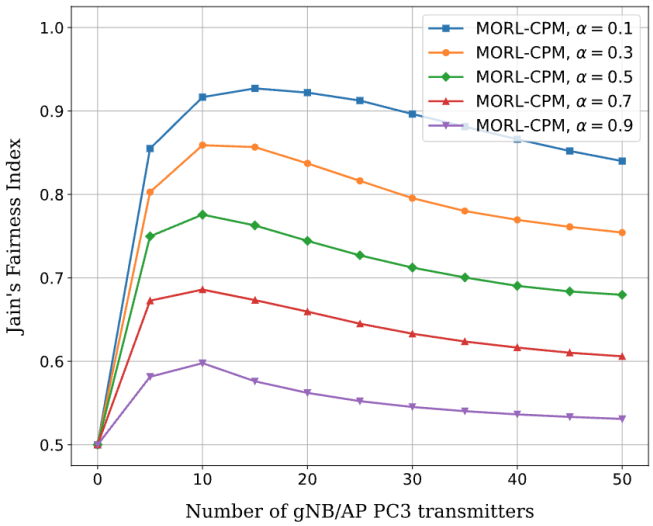
- *Scenario 1:* A gNB PC1 transmitter coexists with a fixed set of 25 PC3 transmitters drawn from both NR-U and Wi-Fi networks.
- *Scenario 2:* The gNB PC1 transmitter coexists with a symmetric and varying number of PC3 transmitters from both gNB and AP sides, ranging from 1 to 50.

We first compare MORL-CPM and QaSAL-CPM in Scenario 1 with the action space restricted to the contention window (CW). This setting provides a controlled baseline to isolate the effect of state augmentation when both approaches have identical control capabilities. We

[2]The code is available on Github: https://github.com/mfasihi/QaSAL_Coexistence_MAC.

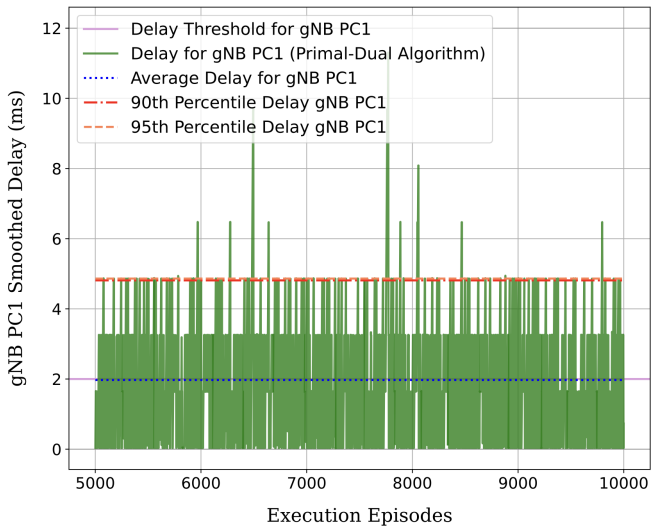


(a) Average medium access delay of gNB PC1

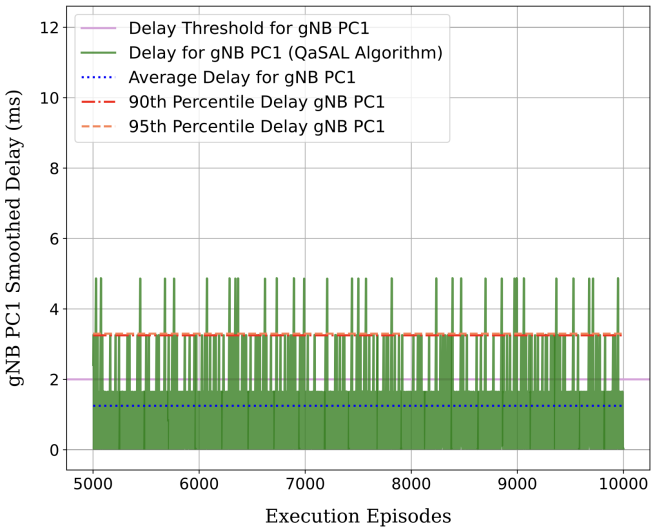


(b) Fairness between PC1 and PC3 transmitters

FIGURE 4: Performance of the MORL-CPM algorithm in Scenario 2 for different weight combinations when the action space includes only the contention window (CW). The results show (a) medium access delay of gNB PC1 and (b) Jain's fairness index between PC1 and PC3 transmitters, averaged over a large number of realizations (confidence intervals are omitted for readability), as the number of contending PC3 nodes increases. The dashed line indicates the QoS delay threshold for the PC1 transmitter ($D_{\text{th}} = 2$ ms).



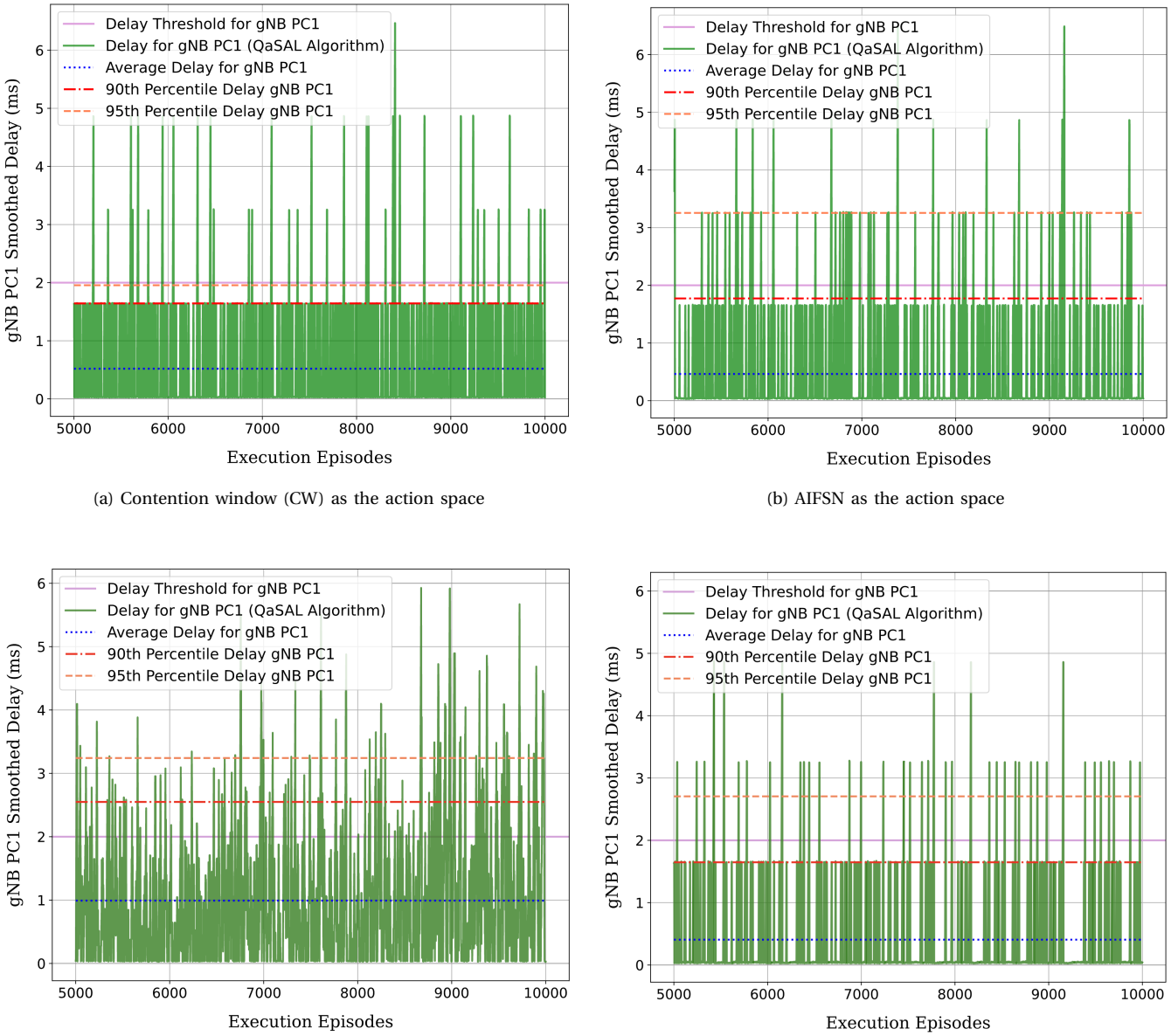
(a) Primal-Dual



(b) QaSAL (without violation scaling)

FIGURE 5: Comparison between the primal-dual constrained RL approach and QaSAL-CPM without violation scaling in Scenario 1, where the action space includes only the contention window (CW). The plots show the smoothed medium access delay of gNB PC1 during execution, including the average delay and the 90th- and 95th-percentile delays. The horizontal line denotes the QoS delay threshold for PC1.

then expand the CPM action space to include additional MAC parameters and study each parameter independently to quantify its individual impact on coexistence performance. Next, we study the joint control of CW and AIFSN using a multi-head DDQN extension of QaSAL-CPM, motivated by the observation that MCOT has limited influence during the contention phase and therefore a weaker effect on delay-sensitive coexistence behavior.



(a) Contention window (CW) as the action space

(b) AIFSN as the action space

(c) MCOT as the action space

(d) Joint control of CW and AIFSN

FIGURE 6: Smoothed medium access delay of the gNB PC1 transmitter in Scenario 1 under the QaSAL-CPM algorithm with violation scaling. Results are shown for different MAC parameters used as the action space: (a) CW, (b) AIFSN, (c) MCOT, and (d) joint control of CW and AIFSN. The plots include the average delay as well as the 90th- and 95th-percentile delays, with the horizontal line indicating the QoS delay threshold for PC1.

Finally, we assess the robustness of QaSAL-CPM under different delay thresholds and increasing levels of network congestion. The hyperparameters used in all experiments are summarized in Table 1. The step duration is selected to span multiple transmission opportunities, enabling accurate estimation of medium access delay.

Figure 4 illustrates how the trade-off coefficient α (see Eq. (7)) in the MORL-CPM framework affects the medium access delay experienced by a PC1 transmitter and JFI between PC1 and PC3 transmitters from both networks, respectively. As the number of PC3 transmitters increases, the delay performance degrades across all configurations; however, the degree of degradation depends significantly on the choice of α. Notably, lower values of α, which place greater emphasis on the fairness (as the secondary objective), result in higher delays for the PC1 transmitter, clearly exceeding the delay threshold of 2 ms. Even moderate scalarization settings (e.g., $\alpha = 0.5$) fail to consistently

enforce the QoS constraint as the network becomes more congested. Only high scalarization values (e.g., $\alpha = 0.7$ and $\alpha = 0.9$), which strongly prioritize delay minimization for PC1, maintain the delay below or near the required threshold.

These results highlight a key limitation of scalarized multi-objective based optimization in constrained environments: there is no guarantee of strict QoS satisfaction for high-priority traffic, especially under varying network conditions. This limitation motivates the need for more explicit constraint-handling strategies.

B. Primal-Dual vs. QaSAL-CPM (Scenario 1)

We compare the performance of the Primal-Dual approach and the proposed QaSAL method under Scenario 1. Figure 5 illustrates the evolution of the smoothed medium access delay for gNB PC1 for Primal-Dual and QaSAL, respectively, over 5000 execution episodes, each of duration 1.25 second. While Primal-Dual satisfies the delay constraint on average, it frequently experiences constraint violations, indicating weaker adherence to real-time QoS requirements. The horizontal, dashed orange and red lines indicate the 90th- and 95th-percentile smoothed delays, respectively, which are commonly used performance metrics for URLLC traffic. The percentile delays well exceed the 2 ms delay threshold for Primal-Dual. For QaSAL without violation scaling, even though the average delay stays well below the threshold, the delay signal shows frequent spikes above the delay threshold, and the percentile delays remain above the threshold. This behavior indicates that the dual update reacts too sharply to raw violation signals, causing oscillations and weak adherence to the QoS constraint.

C. Impact of Violation Scaling (Scenario 1)

Figure 6a shows the gNB PC1 delay using QaSAL with violation scaling and controlling the CW parameter as the action. Compared to QaSAL without violation scaling in Fig. 5b, the smoothed delay stays consistently below the threshold with fewer and shorter violations, showing that the agent learns to balance fairness and constraint satisfaction more effectively. Signed violation scaling combined with temporal smoothing produces a well-conditioned feedback signal which improves convergence stability and long-term compliance while preserving delay performance and significantly better compliance with QoS constraints. The delay dynamics in QaSAL are directly influenced by the evolution of the dual variable associated with the delay constraint, as defined in equation (9b). Compared to Primal-Dual, QaSAL demonstrates a more adaptive and responsive adjustment of the dual variable by spiking in response to constraint violations and decaying as the constraint becomes satisfied and enabling more reliable and timely QoS management.

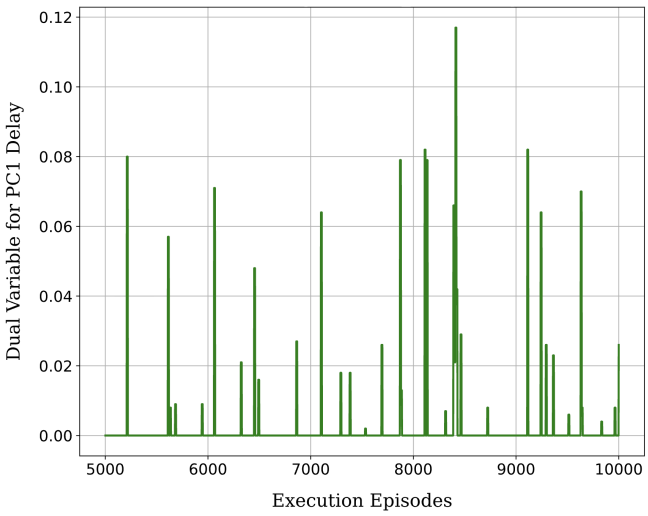


FIGURE 7: Evolution of the dual variable associated with the PC1 delay constraint during the execution phase of QaSAL-CPM.

D. Impact of Coexistence Parameter Selection (Scenario 1)

Figures 6b and 6c depict QaSAL performance when the action controls AIFSN and MCOT, respectively. Compared with the CW-controlled case, AIFSN yields more frequent delay spikes, and the 95th-percentile delay sits above the threshold. Two issues likely cause this result: (i) AIFSN changes the fixed deferment period before backoff begins, causing class-wide timing alignment shifts that can suddenly increase contention in the early slots; and (ii) the available AIFSN levels are coarse, so single-step policy changes cause large priority shifts and not just statistical spacing. In contrast, CW offers a finer gain on contention aggressiveness, smoothing access without amplifying burstiness. Therefore, under assumed traffic load and priority mix, CW is the more effective primary knob for delay constraint adherence; AIFSN should be used cautiously to avoid tail inflation. Figure 6c clearly shows that MCOT is a weaker lever on medium access delay, especially under saturation conditions. This is reasonable, as MCOT only impacts how long the transmitter holds the channel after it has already won a contention round.

The performance of QaSAL under joint control of the CW and AIFSN parameters is shown in Fig. 6d. Compared to using AIFSN alone, the delays becomes more regular and the occurrence of isolated spikes is reduced, which indicates better short-term stabilization of medium access. At the same time, the 95th-percentile delay remains above the QoS threshold, so joint control does not eliminate worst-case delay events in this setting. This suggests that while combining CW and AIFSN helps smooth fluctuations, it does not fundamentally change the tail behavior driven by contention under high load. In

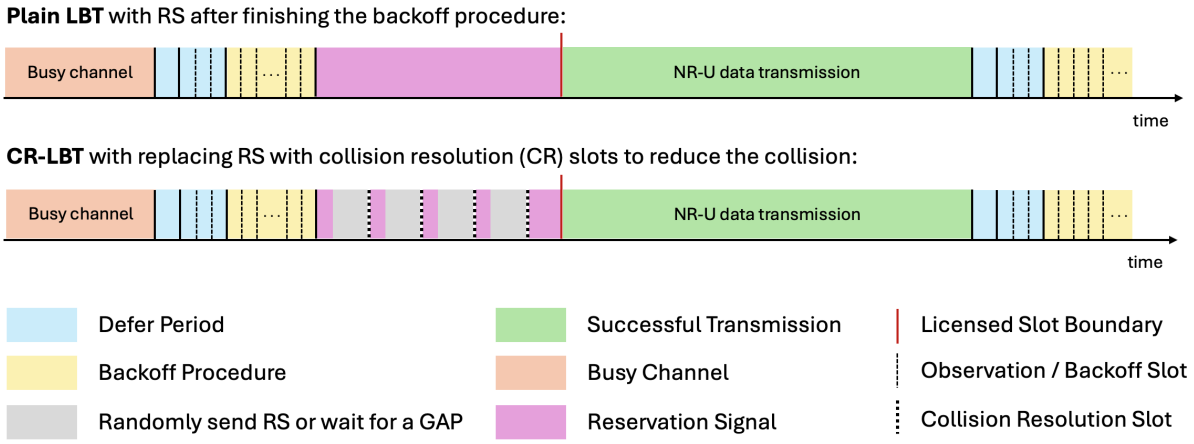


FIGURE 8: Enhanced LBT in NR-U gNBs. Plain LBT transmits a reservation signal (RS) after backoff until the next slot, whereas CR-LBT [29] replaces RS with short collision-resolution (CR) slots that reduce collisions while keeping slot alignment.

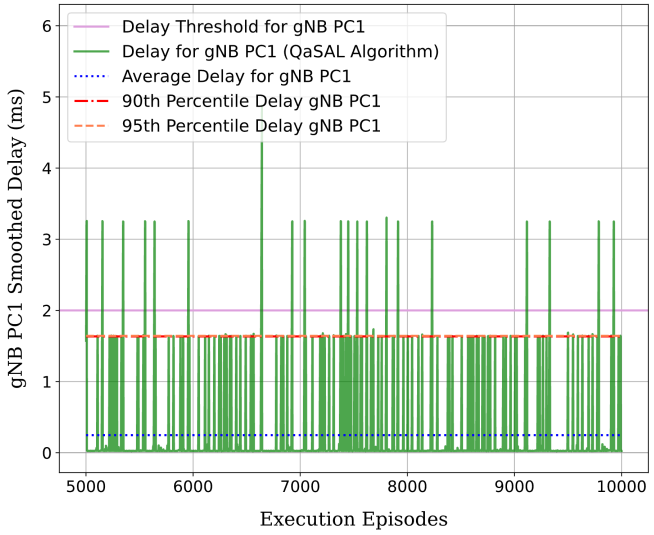


FIGURE 9: Smoothed medium access delay of the gNB PC1 transmitter in Scenario 1 under the QaSAL-CPM algorithm with violation scaling and CR-LBT; action is CW.

TABLE 2: Performance comparison of QaSAL-CPM with Plain LBT and enhanced LBT (CR-LBT).

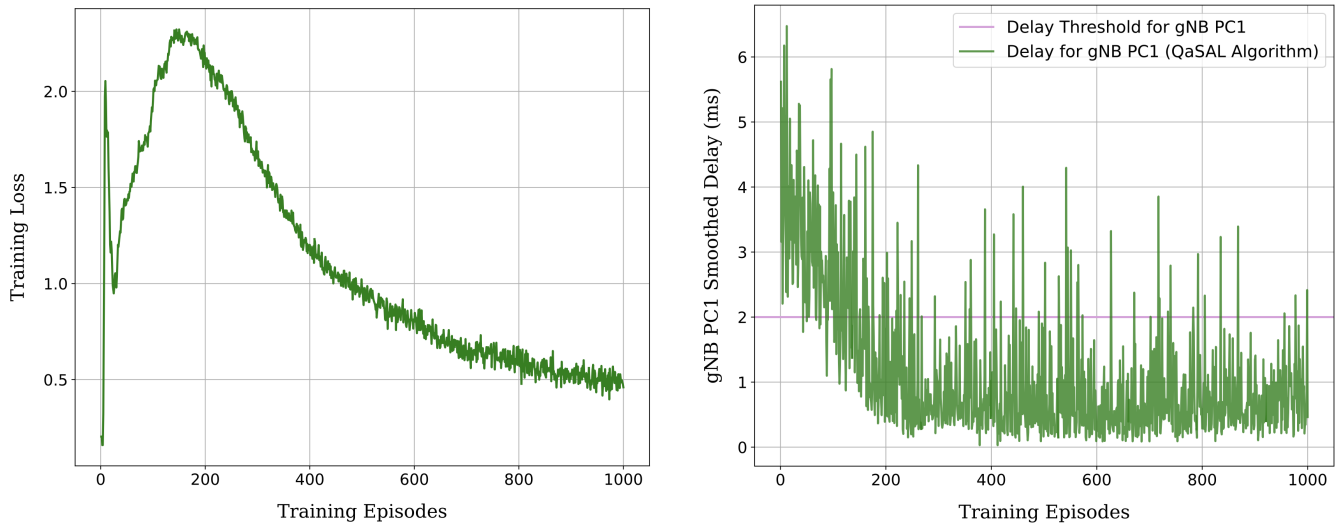
Average Collision Probability			
Node	Plain LBT	CR-LBT	Δ
gNB PC1	10.7%	0.25%	−10.45%
gNB PC3	97.4%	45.7%	−51.7%
AP PC3	98.6%	61.6%	−37.0%
Average Airtime Efficiency			
Node	Plain LBT	CR-LBT	Δ
gNB PC1	89.0%	100.0%	+11.0%
gNB PC3	9.0%	56.0%	+47.0%
AP PC3	8.0%	32.0%	+24.0%
Average Delay			
Node	Plain LBT	CR-LBT	Δ
gNB PC1	0.52 ms	0.25 ms	−0.27 ms

practice, the learned policy appears to use AIFSN mainly to moderate variability, while CW continues to determine whether strict delay bounds can be met. As a result, joint control improves stability but does not outperform CW-only control in terms of tail-delay guarantees. These observations motivate the next experiment, where CW control is combined with a enhanced LBT (CR-LBT), showing that explicitly avoiding synchronized contention events can effectively suppress delay spikes and maintain the 95th-percentile delay below the QoS threshold.

E. Evolution of Dual Variables

To further illustrate the behavior of the proposed QaSAL-CPM algorithm, we examine the evolution of the dual variable associated with the delay constraint during the execution phase. Fig. 7 shows the trajectory of the dual variable λ over time in Scenario 1 using QaSAL with violation scaling and controlling the CW parameter as the action (Fig. 6a).

The dual variable dynamically adjusts in response to constraint violations, increasing when the medium access delay exceeds the prescribed threshold and decreasing



(a) Training loss evolution of the QaSAL-CPM algorithm

(b) Smoothed medium access delay of gNB PC1 during training

FIGURE 10: Training behavior of the QaSAL-CPM algorithm in Scenario 1: (a) the evolution of the smoothed medium access delay of the gNB PC1 transmitter, illustrating progressive constraint satisfaction during learning, and (b) the corresponding training loss, exhibiting an initial transient followed by gradual stabilization.

when the constraint is satisfied. This behavior reflects the role of λ as a measure of constraint pressure, guiding the agent toward actions that reduce violations while maintaining the primary objective. Compared to conventional primal-dual approaches, the state-augmented formulation enables more responsive and stable dual dynamics. Since the dual variable is embedded directly in the agent's state, the learned policy can immediately adapt its actions based on the current level of constraint pressure, without requiring repeated re-optimization. This results in smoother transitions and improved real-time adherence to QoS requirements. These observations highlight a key advantage of QaSAL-CPM: the ability to tightly couple constraint enforcement with policy adaptation, leading to more reliable and efficient coexistence control in dynamic wireless environments.

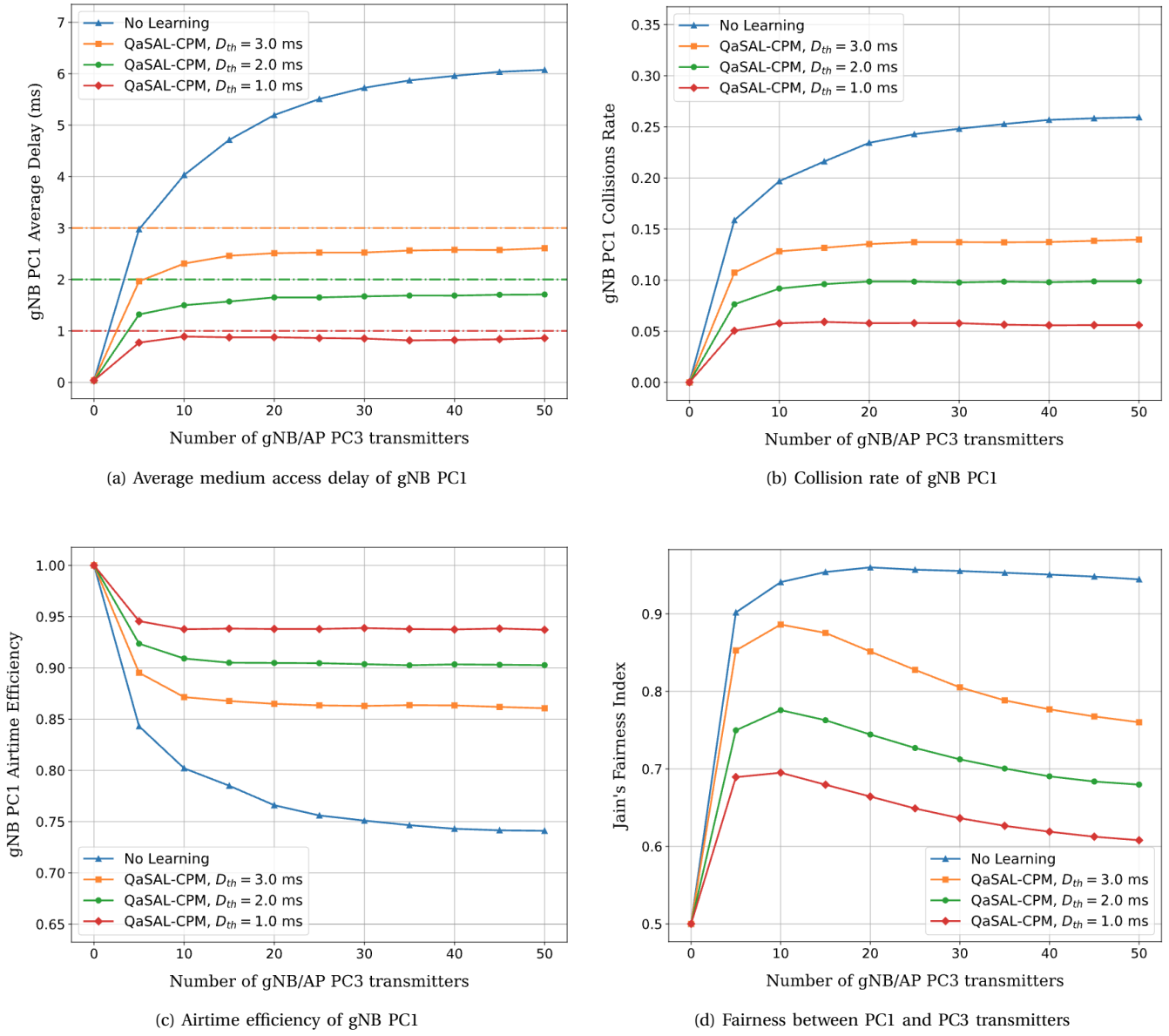
F. Impact of Enhanced LBT Mechanism (Scenario 1)

Collisions in unlicensed bands arise from heterogeneous carrier-sense thresholds, asynchronous timing, and partial visibility between NR-U and Wi-Fi devices. Misaligned slot boundaries and long reservation intervals can cause overlapping transmissions and extra delay, especially for high-priority traffic, while also compromising overall fairness. Enhanced Listen-Before-Talk (LBT) mechanisms address these issues by turning short pre-transmission gaps into decision points for collision avoidance. In particular, Collision-Resolution LBT (CR-LBT) [29] interleaves brief sensing opportunities before slot boundaries, allowing gNBs to detect active transmissions and defer access when necessary. As illustrated in Fig. 8, CR-LBT replaces the reservation signal with short collision-resolution slots, reducing collisions and improving airtime efficiency under heavy contention.

Figure 9 illustrates the delay performance of the QaSAL controller when CR-LBT is enabled alongside violation scaling. Compared with plain LBT (Fig. 6a), CR-LBT significantly reduces both the frequency and magnitude of delay spikes, keeping the average and 95th-percentile delays consistently below the threshold. This improvement stems from the additional collision-resolution slots, which mitigate cross-technology collisions and contention near slot boundaries. Moreover, CR-LBT enhances channel utilization for all nodes by reducing the overall collision probability, particularly for the nodes carrying low-priority traffic. Table 2 summarizes the performance across different network components under QaSAL with plain LBT and CR-LBT. The results show that enhanced collision resolution effectively complements QoS-aware learning by lowering retransmissions and improving overall airtime efficiency.

G. Training Dynamics and Convergence Behavior (Scenario 1)

Figure 10a shows the corresponding training loss. After an initial transient period, the loss increases as the Q-function adapts to the evolving policy and constraint feedback, followed by a gradual and sustained decrease. The loss stabilizes toward the later stages of training, suggesting convergence of value estimation despite the non-stationary environment. These results confirm that QaSAL-CPM achieves stable learning while progressively



(a) Average medium access delay of gNB PC1

(b) Collision rate of gNB PC1

(c) Airtime efficiency of gNB PC1

(d) Fairness between PC1 and PC3 transmitters

FIGURE 11: Performance of the QaSAL-CPM algorithm in Scenario 2 under different PC1 delay thresholds $D_{\text{th}} = 1.0$, 2.0, and 3.0 ms. The results show (a) average medium access delay of gNB PC1, (b) collision rate of gNB PC1, (c) airtime efficiency of gNB PC1, and (d) Jain's fairness index between PC1 and PC3 transmitters, as the number of contending PC3 nodes increases. The dashed lines represent the corresponding QoS delay thresholds.

improving QoS compliance during training. Figure 10b depicts the evolution of the smoothed medium access delay of the gNB PC1 transmitter during training under the QaSAL-CPM algorithm. In the early episodes, the delay exhibits large fluctuations and frequent violations of the QoS threshold, reflecting the exploratory phase of learning. As training progresses, the delay steadily decreases and becomes mostly confined below the threshold, with only occasional spikes caused by stochastic contention effects. This trend indicates that the agent gradually learns to enforce the delay constraint through the augmented state representation.

H. Varying Number of Contending Nodes (Scenario 2)

Figure 11 illustrates the performance of the QaSAL-CPM algorithm in Scenario 2 under different PC1 delay threshold configurations, averaged over a sufficiently large number of realizations to achieve small 95% confidence intervals (omitted from the curves for readability). Specifically, Fig. 11a shows the medium access delay of the gNB PC1 transmitter, Fig. 11b reports the corresponding

collision rate, Fig. 11c presents the airtime efficiency of gNB PC1, and Fig. 11d depicts JFI between PC1 and PC3 transmitters as the number of contending PC3 nodes increases. Across all threshold settings, QaSAL-CPM effectively regulates the average delay of PC1 and keeps it close to the desired QoS target, while more relaxed thresholds naturally permit higher delays. In contrast, the "No Learning" baseline [29] exhibits a steadily increasing delay and collision rate as contention grows, highlighting the need for learning-based adaptation in dense coexistence scenarios. Notably, QaSAL-CPM stabilizes delay performance even under high congestion, while simultaneously reducing collisions relative to the baseline. In these experiments, a single QaSAL-CPM policy is trained over multiple delay-threshold configurations rather than training separate policies for each threshold. The state-augmented formulation enables the learned policy to adapt to varying constraint levels through the embedded dual variables. By training over a sufficiently broad range of delay thresholds and dual variable values, the learned policy generalizes across multiple QoS requirements without requiring retraining for each threshold configuration.

The results also reveal a clear trade-off between delay enforcement and fairness. As shown in Fig. 11d, tighter delay thresholds improve delay performance but lead to lower fairness, since the algorithm prioritizes protecting PC1 latency over airtime distribution. Conversely, more relaxed delay thresholds, such as $D_{\text{th}} = 2.0$, and 3.0 ms, allow greater flexibility in channel access, resulting in improved fairness and higher airtime efficiency, as seen in Fig. 11c. In addition, when the delay threshold is relaxed, the agent tends to maintain the average delay well below the threshold to reduce the risk of violations under increasing congestion. Overall, these results highlight the inherent trade-off between delay and fairness in QoS-aware coexistence control and demonstrate that QaSAL-CPM can flexibly navigate this trade-off.

VII. Conclusion

We presented QaSAL-CPM, a QoS-aware state-augmented learning framework for coexistence parameter management in unlicensed spectrum. By embedding dual variables directly into the agent's observation space, QaSAL-CPM achieves real-time responsiveness to constraint violations and tighter QoS enforcement compared to multi-objective reinforcement learning and conventional primal-dual constrained reinforcement learning. The QaSAL-CPM framework separates offline learning from lightweight online execution, making this approach practical for latency-sensitive applications such as Ultra-Reliable Low-Latency Communications (URLLC) and Internet of Things (IoT) deployments. Extensive simulations under diverse 5G NR-U and Wi-Fi coexistence scenarios demonstrate that QaSAL-CPM enforces 95th-percentile delay compliance, improves fairness, and maintains robust performance under heavy contention. The introduction of bounded violation scaling further stabilizes learning and enhances QoS adherence, while integration with enhanced Listen-Before-Talk mechanisms significantly reduces delay spikes and collision rates.

Overall, QaSAL-CPM provides a scalable and adaptive foundation for real-time coexistence performance optimization in dynamic wireless environments. Future work will extend this framework to multi-channel coexistence, incorporate physical-layer parameters, and explore distributed implementations for edge and device-level deployment. Furthermore, the proposed framework can be extended to support a broader mix of heterogeneous traffic types, including combinations of URLLC, eMBB, and mMTC services with diverse traffic arrival patterns and QoS requirements. Incorporating such traffic diversity is an important direction for enhancing the realism of coexistence modeling.

REFERENCES

[1] V. Sathya *et al.*, "Standardization advances for cellular and Wi-Fi coexistence in the unlicensed 5 and 6 GHz bands," *GetMobile, Mobile Comput. Commun.*, vol. 24, no. 1, pp. 5–15, 2020.

[2] R. K. Saha, "Coexistence of Cellular and IEEE 802.11 Technologies in Unlicensed Spectrum Bands - A Survey," *IEEE Open Journal of the Commun. Soc.*, vol. 2, pp. 1996–2028, 2021.

[3] M. Hirzallah, M. Krunz, B. Kecicioglu, and B. Hamzeh, "5G New Radio Unlicensed: Challenges and Evaluation," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 7, no. 3, pp. 689–701, 2021.

[4] S. Muhammad, H. H. Refai, and M. O. A. Kalaa, "5G NR-U: Homogeneous Coexistence Analysis," in *IEEE Globecom*, Taipei, Taiwan, 2020.

[5] S. Mannor and N. Shimkin, "A geometric approach to multi-criterion reinforcement learning," *J. Mach. Learn. Res.*, vol. 5, p. 325–360, 2004.

[6] K. V. Moffaert, M. M. Drugan, and A. Nowé, "Scalarized multi-objective reinforcement learning: Novel design techniques," in *IEEE Symp. Adapt. Dyn. Prog. and Reinforcement Learning (ADPRL)*, Singapore, 2013, pp. 191–199.

[7] S. Bhatnagar and K. Lakshmanan, "An Online Actor–Critic Algorithm with Function Approximation for Constrained Markov Decision Processes," *J. Optim. Theory Appl.*, vol. 153, p. 688–708, 2012.

[8] M. Calvo-Fullana, S. Paternain, L. F. O. Chamon, and A. Ribeiro, "State Augmented Constrained Reinforcement Learning: Overcoming the Limitations of Learning With Rewards," *IEEE Trans. Autom. Control*, vol. 69, no. 7, pp. 4275–4290, 2024.

[9] N. NaderiAlizadeh, M. Eisen, and A. Ribeiro, "State-Augmented Learnable Algorithms for Resource Management in Wireless Networks," *IEEE Trans. Signal Process.*, vol. 70, no. 7, pp. 5898–5912, 2022.

[10] M. R. Fasihi and B. L. Mark, "Traffic Priority-Aware 5G NR-U/Wi-Fi Coexistence with Deep Reinforcement Learning," in *IEEE 100th Vehic. Tech. Conf. (VTC2024-Fall)*, Washington, DC, USA, 2024.

[11] ——, "QaSAL: QoS-aware State-Augmented Learnable Algorithms for Coexistence of 5G NR-U/Wi-Fi," in *IEEE 59th Conf. Inf. Sci. Sys. (CISS)*, Baltimore, MD, USA, Apr. 2025.

[12] G. Naik, J. M. Park, J. Ashdown, and W. Lehr, "Next Generation Wi-Fi and 5G NR-U in the 6 GHz Bands: Opportunities and Challenges," *IEEE Access*, vol. 8, pp. 153 027–153 056, 2020.

[13] M. Hirzallah, M. Krunz, and Y. Xiao, "Harmonious Cross-Technology Coexistence With Heterogeneous Traffic in Unlicensed Bands: Analysis and Approximations," *IEEE Trans. on Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 690–701, 2019.

[14] J. Oh, Y. Kim, Y. Li, J. Bang, and J. Lee, "Expanding 5G New Radio Technology to Unlicensed Spectrum," in *IEEE Globecom Workshops (GC Wkshps)*, Waikoloa, HI, USA, 2019.

[15] D. Farahiyah, Iskandar, K. Anwar, and Hendrawan, "Study on Potential Interferences of 5G NR-U and WiFi 6E to Fixed Service in 6 GHz," in *IEEE Int. Conf. on Comp. (ICOCO)*, Kuala Lumpur, Malaysia, 2024, pp. 249–254.
[16] J. Ssimbwa, S. H. Yoon, Y. Lee, and Y. C. Ko, "Towards 5G-advanced NR-unlicensed systems: Physical layer design and performance," *J. Commun. Net.*, vol. 26, no. 2, pp. 207–214, 2024.
[17] Y. Shi, Q. Cui, W. Ni, and Z. Fei, "Proactive Dynamic Channel Selection Based on Multi-Armed Bandit Learning for 5G NR-U," *IEEE Access*, vol. 8, pp. 196 363–196 374, 2020.
[18] M. Hirzallah and M. Krunz, "Sense-Bandits: AI-based Adaptation of Sensing Thresholds for Heterogeneous-technology Coexistence Over Unlicensed Bands," in *Int. Conf. Comp. Commun. and Netw. (ICCCN)*, Athens, Greece, 2021, pp. 1–9.
[19] R. Bajracharya, R. Shrestha, and H. Jung, "Bandit Approach for Fair and Efficient Coexistence of NR-U in Unlicensed Bands," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5208–5223, 2023.
[20] Z. Guo, C. Zhang, M. Li, and M. Krunz, "Fair Coexistence of Heterogeneous Networks: A Novel Probabilistic Multi-Armed Bandit Approach," in *21st Int. Symp. on Model. and Opt. in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, Singapore, Singapore, 2023.
[21] L. Wang, M. Zeng, J. Guo, Q. Cui, and Z. Fei, "Joint Bandwidth and Transmission Opportunity Allocation for the Coexistence Between NR-U and WiFi Systems in the Unlicensed Band," *IEEE Trans. Veh. Technol.*, vol. 70, no. 11, pp. 11 881–11 893, 2021.
[22] F. Zeinali, S. Norouzi, N. Mokari, and E. A. Jorswieck, "AI-based Radio Resource and Transmission Opportunity Allocation for 5G-V2X HetNets: NR and NR-U networks," *Int. Jour. of Elec. and Commun. Eng.*, vol. 17, no. 9, pp. 209–216, 2023.
[23] Y. Liu, H. Zhou, Y. Deng, and A. Nallanathan, "Channel Access Optimization in Unlicensed Spectrum for Downlink URLLC: Centralized and Federated DRL Approaches," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2208–2222, 2023.
[24] H. Zou, Y. Li, X. Chu, C. Xu, and T. Wang, "Improving Fairness in Coexisting 5G and Wi-Fi Network on Unlicensed Band with URLLC," in *IEEE/ACM 31st Int. Symp. on Qual. of Serv. (IWQoS)*, Orlando, FL, USA, 2023.
[25] Q. Liu, N. Choi, and T. Han, "Constraint-Aware Deep Reinforcement Learning for End-to-End Resource Orchestration in Mobile Networks," in *IEEE 29th Int. Conf. on Netw. Prot. (ICNP)*, Dallas, TX, USA, 2021.
[26] Y. B. Uslu, R. Doostnejad, A. Ribeiro, and N. NaderiAlizadeh, "Learning to Slice Wi-Fi Networks: A State-Augmented Primal-Dual Approach," in *IEEE Global Communications Conference*, Cape Town, South Africa, 2024.
[27] S. Das, N. NaderiAlizadeh, and A. Ribeiro, "State-Augmented Opportunistic Routing in Wireless Communication Systems with Graph Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hyderabad, India, 2025.
[28] Y. B. Uslu, N. NaderiAlizadeh, M. Eisen, and A. Ribeiro, "Fast State-Augmented Learning for Wireless Resource Allocation with Dual Variable Regression," *arXiv preprint arXiv:2506.18748*, 2025. [Online]. Available: https://arxiv.org/abs/2506.18748
[29] V. Loginov, A. Troegubov, A. Lyakhov, and E. Khorov, "Enhanced Collision Resolution Methods With Mini-Slot Support for 5G NR-U," *IEEE Access*, vol. 9, pp. 146 137–146 152, 2021.
[30] ETSI, "LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 37.213 version 16.3.0 Release 16)," ETSI, Tech. Rep. ETSI TS 137 213 V16.3.0, July 2020.
[31] H. van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-Learning," in *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2094–2100.

Mohammad Reza Fasihi (S'23) received the B.Sc. degree in Electrical Engineering from Ferdowsi University of Mashhad, Mashhad, Iran, in 2008, and the M.Sc. degree in Electrical Engineering from Iran University of Science and Technology, Tehran, Iran, in 2020. He is currently pursuing the Ph.D. degree in Electrical and Computer Engineering at George Mason University, Fairfax, VA, USA. He has served as a faculty member in the Department of Computer Science at Kateb University, Afghanistan, and currently serves as Chair of the IEEE Student Branch at George Mason University. His research focuses on intelligent resource management for next-generation wireless networks, with an emphasis on QoS-aware NR-U/Wi-Fi coexistence, reinforcement learning-based control, integrated sensing and communication, and digital twin-enabled wireless systems.

Brian L. Mark (S'91–M'95–SM'08) received the B.A.Sc. degree in Computer Engineering with a minor in mathematics from the University of Waterloo in 1991 and the Ph.D. degree in Electrical Engineering from Princeton University in 1995. From 1995 to 1999, he was a Research Staff Member with NEC USA, Princeton, New Jersey. In 1999, he was a Visiting Researcher with Télécom Paris in France. In 2000, he joined George Mason University, where he is currently a Professor of Electrical and Computer Engineering. He served as Interim Chair of the Department of Bioengineering from 2015–2017 and as Interim Chair of the Department of Electrical and Computer Engineering from 2023–2025. Dr. Mark was an Associate Editor of the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2006 to 2009. His research interests lie generally in the design, architecture, and performance of communication networks, including wireless networks, broadband networks and quantum networks, alongside applications of stochastic modeling, signal processing, and machine learning.

Omar Alotaibi (S'11–M'25–SM'25) received the B.S. degree in Electrical Engineering from King Saud University, Riyadh, Saudi Arabia, in 2012 and the M.S. degree in Electrical Engineering from Arizona State University in 2015. He received the Ph.D. in Electrical and Computer Engineering from George Mason University in 2025. He is currently an Assistant Professor in the Department of Computer Engineering at King Saud University, where he also serves as Vice-Chief Artificial Intelligence Officer in the Artificial Intelligence Office. He served as Vice-Chair of the IEEE Northern Virginia Section Control Systems Society Chapter in 2025. His current research interests include Bayesian filtering and robust estimation, nonlinear discrete-time dynamical systems, stochastic modeling, machine learning, tracking and navigation, and sensor networks and data fusion.