# QaSAL: QoS-aware State-Augmented Learnable Algorithms for Coexistence of 5G NR-U/Wi-Fi

Mohammad Reza Fasihi and Brian L. Mark
Dept. of Electrical and Computer Engineering and Wireless Cyber Center
George Mason University, Fairfax, Virginia, United States
Email: mfasihi4@gmu.edu, bmark@gmu.edu

***Abstract*—With the increasing demand for wireless connectivity, ensuring the efficient coexistence of multiple radio access technologies in shared unlicensed spectrum has become an important issue. This paper focuses on optimizing Medium Access Control (MAC) parameters to enhance the coexistence of 5G New Radio in Unlicensed Spectrum (NR-U) and Wi-Fi networks operating in unlicensed spectrum with multiple priority classes of traffic that may have varying quality-of-service (QoS) requirements. In this context, we tackle the coexistence parameter management problem by introducing a QoS-aware State-Augmented Learnable (QaSAL) framework, designed to improve network performance under various traffic conditions. Our approach augments the state representation with constraint information, enabling dynamic policy adjustments to enforce QoS requirements effectively. Simulation results validate the effectiveness of QaSAL in managing NR-U and Wi-Fi coexistence, demonstrating improved channel access fairness while satisfying a latency constraint for high-priority traffic.**

***Index Terms*—5G NR-U, Wi-Fi, QoS, coexistence, constrained reinforcment learning, Lagrangian duality, primal-dual, state augmentation.**

## I. Introduction

The rapid growth in wireless connectivity demands, driven by diverse applications ranging from mobile communications to IoT networks, has led to an increasing reliance on unlicensed spectrum. These bands are shared by multiple Radio Access Technologies (RATs), such as 5G New Radio in Unlicensed Spectrum (NR-U) and Wi-Fi, introducing complex interference dynamics that can degrade network performance if not managed effectively. The primary challenge lies in achieving high network performance while maintaining fairness among different technologies sharing the same spectrum [1], [2], [3].

In this context, the coexistence of 5G NR-U and Wi-Fi networks introduces significant complexities. The two technologies use distinct channel access mechanisms: 5G NR-U relies on Listen Before Talk (LBT) procedures, while Wi-Fi employs Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA). These mechanisms must operate harmoniously to minimize inter-network collisions and ensure equitable resource allocation. However, this is non-trivial due to the diverse Quality of Service (QoS) requirements, traffic patterns, and contention behaviors of the two networks. For instance, 5G NR-U often has stringent latency and reliability requirements, particularly for high-priority traffic, which can conflict with the opportunistic nature of Wi-Fi's channel access mechanism.

Efficient coexistence is further complicated by the dynamic nature of unlicensed spectrum usage, where network configurations, traffic loads, and environmental factors fluctuate over time. These fluctuations demand adaptive and robust strategies to ensure that coexistence mechanisms can respond effectively to varying conditions [4]. Without such strategies, the performance of both networks can degrade due to the increased collisions, higher latencies, and unfair resource allocation. For the coexistence of 5G NR-U and Wi-Fi networks on unlicensed spectrum, the problem managing contention and enhancing network performance can be formulated in terms of QoS-aware network utility maximization with respect to the Medium Access Control (MAC) parameters of both networks. One approach to this problem is based on regularized multi-objective reinforcement learning (RL), whereby the agent receives a reward representing a weighted combination of individual task-specific rewards [5], [6]. Although this method is widely used and can be effective, a significant drawback is the need for problem-specific selection of the weighting coefficients, which requires manual tuning and results in substantial computational overhead due to the extensive calibration process. Moreover, QoS constraints may be violated since they are not explicitly enforced in this approach.

Alternatively, QoS requirements can be explicitly defined within a constrained reinforcement learning (CRL) framework [7]. In this setting, a single objective function referred to as the *Lagrangian function* is maximized with respect to primal variables and minimized with respect to dual variables, each dual variable corresponding to a constraint in the original problem. A key advantage of this method is its ability to automatically adjust multiplier values, removing the necessity for manual tuning and thus reducing design complexity. By enforcing constraints in conjunction with RL algorithm, this approach can ensure that QoS constraints are met with high probability. Nonetheless, implementing this approach in dynamic environments, especially under varying network conditions, is challenging. Specifically, the penalty terms require extensive calibration and lack sufficient adaptability, ultimately leading to suboptimal performance in certain scenarios.

This work was supported in part by the U.S. National Science Foundation under Grant 2034616.

In this paper, we optimize the coexistence parameters of 5G NR-U and Wi-Fi networks to meet the QoS requirements of high-priority transmitters. We utilize dual variables in the CRL framework to track constraint violations over time [8], [9], [10]. The wireless network state is *augmented* with these dual variables at each time step, serving as dynamic inputs to the learning algorithm. This augmentation improves the algorithm's understanding of constraints and their environmental relationships, allowing the agent to adjust policies dynamically while minimizing reliance on indirect penalty mechanisms. A *SimPy*-based simulation environment is implemented to model the MAC layer of 5G NR-U and Wi-Fi, supporting dynamic configurations for detailed performance evaluation. Transmitters from both networks operate in a saturated mode, representing high traffic conditions. This study builds upon our previous work [11], where we proposed a traffic priority-aware deep reinforcement learning (DRL) framework for dynamically adjusting contention window sizes to balance network performance and fairness. The current work expands on this by introducing a state-augmented learnable algorithm that directly integrates constraint variables into the state space.

The remainder of the paper is organized as follows. In Section II, we define our problem and study the DRL for CPM problem. Then, the primal-dual approach for our problem introduced in Section III. The QaSAL algorithm for solving the CPM problem is proposed in Section IV. Application of QaSAL framework to the coexistence of 5G NR-U and Wi-Fi is developed in V. In Section VI, we present simulation results, which demonstrate that our proposed algorithm is able to protect the delay performance of high-priority traffic when contending with varying number of lower-priority traffic nodes for channel access. The paper is concluded in Section VII.

## II. Problem Formulation

Consider the coexistence of 5G NR-U and Wi-Fi in an unlicensed spectrum band. Let $\mathcal{S} \subset \mathbb{R}^n$ represent a compact set of *coexistence environment states*, which captures the status of both the 5G NR-U and Wi-Fi networks. Given the state $\boldsymbol{S}_t \in \mathcal{S}$, let $\mathbf{a}(\boldsymbol{S}_t)$ denote the vector of CPM decisions across both networks, where $\mathbf{a}: \mathcal{S} \to \mathbb{R}^a$ is the CPM function. The possible states and CPM decisions are described by a Markov Decision Process (MDP). The agent sequentially makes CPM decisions over discrete time steps $t \in \mathbb{N} \cup \{0\}$ based on a policy $\pi$, leading to the performance vector $\boldsymbol{O}_t = \mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t)) \in \mathbb{R}^m$ which captures various performance metrics of both networks, such as transmission delay, collision percentage, airtime efficiency, fairness, number of successful transmissions, channel utilization ratio, etc. Typically, in any MDP, the focus is on the accumulated performance over time, represented by the value function

$$\mathbf{V}_i(\pi) \triangleq \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{(\boldsymbol{S},\mathbf{a}(\boldsymbol{S}))\sim\pi} \left[ \sum_{t=0}^{T} f_i(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t)) \right], \tag{1}$$

where $i = 1, \ldots, m$. In practice, because the probability distribution of the environment is unknown, the objective is derived based on the *ergodic average* network performance

$$\tilde{\mathbf{V}}_i(\pi) = \frac{1}{T} \sum_{t=0}^{T-1} f_i(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t)). \tag{2}$$

Note that the value functions $\tilde{\mathbf{V}}_i(\pi)$, $i = 1, ..., m$, might be in conflict with each other, and a policy $\pi$ that is optimal for some $\tilde{\mathbf{V}}_i$ may not be good for some other $\tilde{\mathbf{V}}_j$. To include the QoS requirements in our CPM problem, let us define a concave utility function $\mathcal{U}: \mathbb{R}^m \to \mathbb{R}$ and a set of $c$ concave constraints defined in terms of a mapping $\boldsymbol{C}: \mathbb{R}^m \to \mathbb{R}^c$. The goal of the CPM problem is to specify the optimal vector of CPM decisions $\mathbf{a}(\boldsymbol{S}_t)$ for any given state $\boldsymbol{S}_t \in \mathcal{S}$ that optimizes the utility function $\mathcal{U}$ while ensuring the QoS requirements are satisfied.

The dynamic and complex nature of the coexistence environment as well as the different channel access mechanisms and the lack of coordination between two networks makes finding a proper solution for the CPM problem challenging. A learning-based optimization technique can be leveraged to maintain a fair and efficient coexistence. Therefore, we introduce a *parameterized* CPM policy by replacing $\mathbf{a}(\boldsymbol{S})$ with $\mathbf{a}(\boldsymbol{S};\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\Theta}$ denotes a finite-dimensional set of parameters (see Fig. 1). Hence, the generic CPM problem can be defined as

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \mathcal{U}\left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta})) \right), \tag{3a}$$

$$\text{s.t.} \quad \boldsymbol{C}\left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta})) \right) \geq \mathbf{0} \tag{3b}$$

where the maximization is performed over the set of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Note that the objective and constraints are derived based on the ergodic averages of the corresponding performance vectors. The goal of this paper is to develop a learning algorithm to solve (3) for any given coexistence environment state $\boldsymbol{S}_t \in \mathcal{S}$.

## III. Gradient-Based CPM Algorithm In Dual Domain

A customary approach to solve problem (3) is to consider a penalized version in the *Lagrangian dual* domain. Formally, we introduce dual variables $\boldsymbol{\lambda} \in \mathbb{R}_+^c$ associated with the constraints in (3b) and define the Lagrangian

$$\mathcal{L}_\pi(\boldsymbol{\lambda};\boldsymbol{\theta}) = \mathcal{U}\left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta})) \right) + \boldsymbol{\lambda}^T \boldsymbol{C}\left( \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta})) \right). \tag{4}$$

The Lagrangian in (4) should be maximized over $\boldsymbol{\theta}$, while subsequently minimizing over the dual variables $\boldsymbol{\lambda}$, i.e.,

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}_+^c} \quad \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \mathcal{L}_\pi(\boldsymbol{\lambda};\boldsymbol{\theta}). \tag{5}$$

The advantage of replacing the objective in (3) with Lagrangian in (4) is that the latter can be optimized using any parameterized learning framework, such as standard

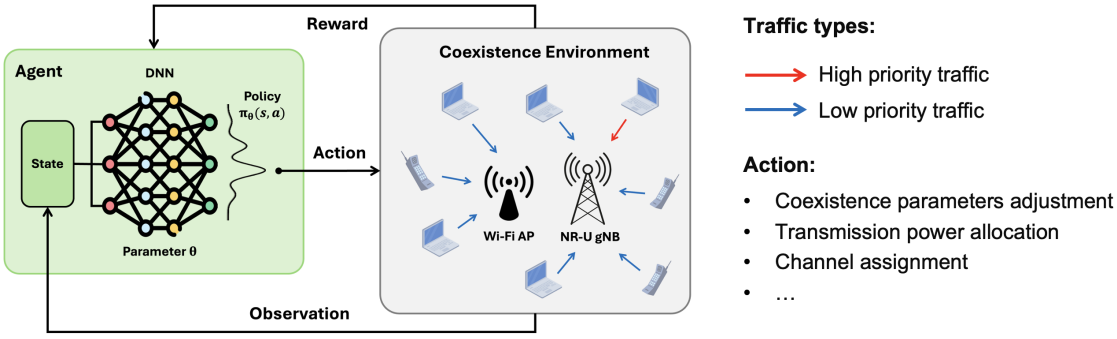


Fig. 1. Parameterized optimization of CPM policy.

reinforcement learning algorithms. One limitation of (5) is the ambiguity in determining suitable values for the dual variables. The optimal choice for $\boldsymbol{\lambda}$ depends on the transition probability $p(\boldsymbol{S}_{t+1}|\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t))$, which is typically unknown. This challenge can be addressed by *dynamically adjusting* the $\boldsymbol{\lambda}$. To achieve this, we introduce an iteration index $k$, a step size $\boldsymbol{\eta}_k \in \mathbb{R}_+^c$, and an epoch duration $T_0$ which is defined as the number of time steps between consecutive model parameter updates. Thus, the model parameters and dual variables are updated iteratively as

$$\boldsymbol{\theta}_k = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \quad \mathcal{U}\left(\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta}))\right) + \boldsymbol{\lambda}_k^T \boldsymbol{C}\left(\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta}))\right), \quad (6)$$

$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \boldsymbol{\eta}_{\boldsymbol{\lambda}}\boldsymbol{C}\left(\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\mathbf{f}(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t;\boldsymbol{\theta}_k))\right)\right]^+, \quad (7)$$

where $[x]^+ := \max(x, 0)$. Note that the step size can be different for each QoS constraint. Dual variables $\boldsymbol{\lambda}$ at iteration $k$ are updated based on whether the constraints are violated or not. Dual variables are increased if the constraints are met and decreased if not, and the scale of update depends on the amount of violation at each time step.

The dual variable update algorithm in (6) and (7) presents a set of challenges that make it difficult to use in practice. Maximizing the Lagrangian in (6) requires knowledge of future network states, which is unattainable during execution, although it may be feasible during the training phase. Furthermore, achieving convergence to a feasible and near-optimal solution is only possible as the operation time $T$ approaches infinity. Finally, the optimal set of model parameters needs to be memorized for any given set of dual variables, which can be computationally expensive, especially during the execution phase. The *state-augmentation* approach overcomes these limitations by embedding dual variables into the state space [8], [9].

## IV. Proposed QoS-aware State-Augmented Learnable Algorithm for CPM

In this section, we propose the QoS-aware State-Augmented Learnable Algorithm (QaSAL) for CPM problem. The key idea of state-augmentation is treating constraint satisfaction as a dynamic component of the agent's environment, which evolves in response to constraint violations through dual dynamics. By augmenting the state with dual variables, the agent learns policies that are constraint-aware and adaptable, leading to feasible, near-optimal solutions that traditional methods cannot guarantee.

We consider state $\boldsymbol{S}_t$ 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 $\tilde{\boldsymbol{S}}_t = (\boldsymbol{S}_t, \boldsymbol{\lambda}_k)$. We introduce an alternative parameterization for the CPM policy, in which the CPM decisions $\boldsymbol{a}(\boldsymbol{S}, \boldsymbol{\theta})$ are represented via the parameterization $\boldsymbol{a}(\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{\theta}})$, where $\tilde{\boldsymbol{\theta}} \in \tilde{\boldsymbol{\Theta}}$ denotes the set of parameters of the state-augmented CPM policy. Then, we define the augmented version of Lagrangian in (4) as

$$\mathcal{L}_\pi(\boldsymbol{\lambda}; \tilde{\boldsymbol{\theta}}) = \mathcal{U}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{f}(\tilde{\boldsymbol{S}}_t, \mathbf{a}(\tilde{\boldsymbol{S}}_t;\tilde{\boldsymbol{\theta}}))\right) + \boldsymbol{\lambda}^T\boldsymbol{C}\left(\frac{1}{T}\sum_{t=0}^{T-1}\mathbf{f}(\tilde{\boldsymbol{S}}_t, \mathbf{a}(\tilde{\boldsymbol{S}}_t;\tilde{\boldsymbol{\theta}}))\right), \quad (8)$$

and formulate the augmented CPM policy optimization problem for any $\boldsymbol{\lambda} \sim p_{\boldsymbol{\lambda}}$ in (5) as

$$\tilde{\boldsymbol{\theta}}^* = \arg\max_{\tilde{\theta}\in\tilde{\Theta}} \mathbb{E}_{\boldsymbol{\lambda}\sim p_{\boldsymbol{\lambda}}}\mathcal{L}_\pi(\boldsymbol{\lambda}; \tilde{\boldsymbol{\theta}}). \quad (9)$$

Utilizing the augmented policy parameterized by (9), we substitute the dual variable update equation in (7) with the augmented version:

$$\boldsymbol{\lambda}_{k+1} = \left[\boldsymbol{\lambda}_k - \boldsymbol{\eta}_{\boldsymbol{\lambda}}\boldsymbol{C}\left(\frac{1}{T_0}\sum_{t=kT_0}^{(k+1)T_0-1}\mathbf{f}(\tilde{\boldsymbol{S}}_t, \mathbf{a}(\tilde{\boldsymbol{S}}_t;\tilde{\boldsymbol{\theta}}^*))\right)\right]^+. \quad (10)$$

Note that in the multiplier update equation (10), the optimal parameters $\tilde{\boldsymbol{\theta}}^*$ are utilized, which mitigates the challenge of storing the model parameters for any given set of dual variables in (6). The training and execution procedures are summarized in Algorithms 1 and 2, respectively. Fig. 2 depicts the proposed QaSAL algorithm for CPM.

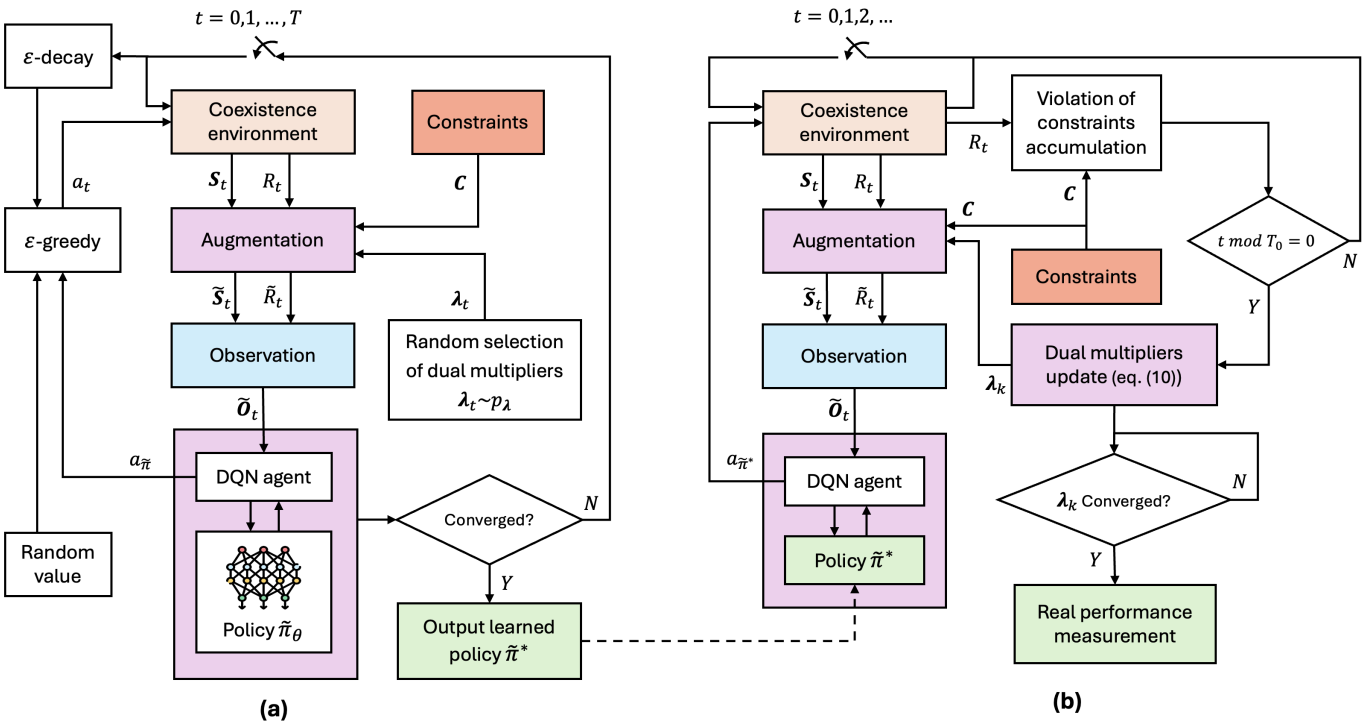


Fig. 2. The flowchart of QaSAL algorithm for CPM: a) Training phase b) Execution phase.

## V. QaSAL-CPM Algorithm for Coexistence of 5G NR-U and Wi-Fi on an Unlicensed Spectrum

We consider the coexistence of 5G NR-U gNBs and Wi-Fi APs on an unlicensed spectrum. Each network handles two types of transmitters: High-priority traffic transmitters with strict delay requirements, and low-priority traffic transmitters with less stringent delay constraints. NR-U gNB transmitters employ the listen-before-talk (LBT) mechanism, which is based on CSMA/CA with a binary exponential backoff procedure and supports four priority classes (PCs) that correspond to four access classes (ACs) in Wi-Fi's enhanced distributed channel access (EDCA) [12]. Performing the LBT protocol can harm the latency performance of high-priority traffic such as Ultra-Reliable Low Latency Communication (URLLC) packets, especially when contending for channel access with low-priority traffic [13]. Additionally, the LBT protocol may cause additional transmission delay compared to the licensed spectrum due to the unpredictability of transmission opportunities. Therefore, the delay performance of high-priority traffic may not meet the requirements.

In this section, we propose a QaSAL-CPM algorithm to protect the low-latency performance of high-priority transmitters while maintaining high fairness among two networks. We denote the high-priority and low-priority transmitters by *PC1* (priority class 1) and *PC3* (priority class 3), respectively. The 5G NR-U network's transmissions align with slot boundaries determined by the licensed spectrum numerology. If the channel is found idle, gNBs send reservation signals (RS) until the slot boundary to prevent other transmitters from occupying the channel.

**Algorithm 1** QaSAL algorithm's training phase for CPM.

1: Sample $\tilde{\boldsymbol{S}} = (\boldsymbol{S}, \boldsymbol{\lambda})$ from augmented space $\mathcal{S} \times \Lambda$.
2: Calculate the augmented Lagrangian according to (8).
3: Use the parameterized learning algorithm to obtain policy according to (9): $\tilde{\boldsymbol{\theta}}^* = \arg\max_{\tilde{\theta} \in \tilde{\Theta}} \mathcal{L}_\pi(\boldsymbol{\lambda}; \tilde{\boldsymbol{\theta}})$.

**Return:** Optimal model parameters $\tilde{\boldsymbol{\theta}}^*$.

**Algorithm 2** QaSAL algorithm's execution phase for CPM.

**Input:** Optimal model parameters $\tilde{\boldsymbol{\theta}}^*$, step $\boldsymbol{\eta}_k$, QoS constraints $\boldsymbol{C}$, epoch $T_0$
**Initialize:** $\tilde{\boldsymbol{S}}_0 \leftarrow (\boldsymbol{S}_0, \boldsymbol{\lambda}_0 = \boldsymbol{0})$

1: **for** $k = 0, 1, \ldots$ **do**
2: Rollout $T_0$ steps with CPM decisions $\boldsymbol{a}(\tilde{\boldsymbol{S}}, \tilde{\boldsymbol{\theta}}^*)$.
3: Update dual dynamics according to (10).
4: **end for**

**Output:** Sequence of CPM decisions $\{\boldsymbol{a}_k; k = 0, 1, \ldots\}$

The coexistence environment state at time step $t$, i.e., $\boldsymbol{S}_t$ is designed to comprehensively capture the dynamic environment of 5G NR-U and Wi-Fi coexistence on unlicensed spectrum. It encodes critical metrics that influence decision-making, including network performance and

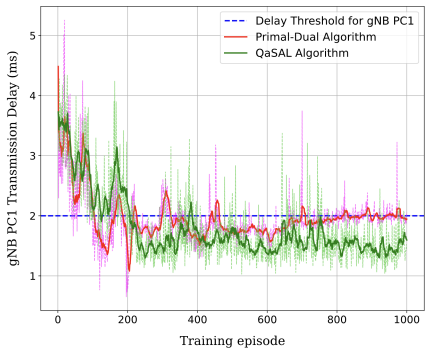


Fig. 3. Training evolution of transmission delay of gNB PC1 transmitter ($D_{\text{th,PC1}} = 2$ ms).

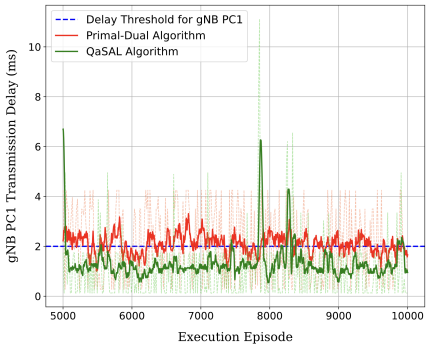


Fig. 4. Execution evolution of transmission delay of gNB PC1 transmitter ($D_{\text{th,PC1}} = 2$ ms).

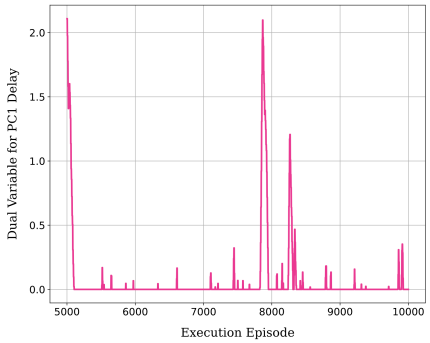


Fig. 5. Dual variable evolution for delay constraint of gNB PC1 transmitter ($D_{\text{th,PC1}} = 2$ ms).

resource utilization such as the average and step-wise transmission delay of high-priority traffic, the percentage of collisions, channel airtime utilization, and Jain's Fairness Index (JFI). Additionally, it tracks trends in delay variation and short-term collision statistics to provide insights into ongoing network conditions.

The CPM decisions $\boldsymbol{a}(\boldsymbol{S}_t)$ is represented as a discrete selection from the set $\{0, 1, \ldots, 6\}$, where the maximum contention window (CW) size is calculated as $\text{CW}_{\max,\,\text{PC}_i} = 2^{\boldsymbol{a}_i + c} - 1$, where $\boldsymbol{a}_i$ is the CPM decision for priority class $i$ and $c = 0$ and 4 for PC1 and PC3, respectively. This design directly influence the backoff timing, which affects each transmitter's access to the shared spectrum. Moreover, the performance function is designed as $f(\boldsymbol{S}_t, \mathbf{a}(\boldsymbol{S}_t)) = (\text{JFI})_t$, where $(\text{JFI})_t$ denotes the JFI among both networks at time $t$. We define a constraint of the form $C(x) \le 0$, where $C(x) = x - x_{\text{th}}$ with a threshold $x_{\text{th}}$. Consequently, the optimization problem in (3) can be formulated as

$$\max_{\pi} \quad \frac{1}{T} \sum_{t=0}^{T-1} (\text{JFI})_t, \tag{11a}$$

$$\text{s.t.} \quad \frac{1}{T} \sum_{t=0}^{T-1} (\text{D}_{\text{PC}_1})_t \le D_{\text{th}} \tag{11b}$$

We implement the QaSAL algorithm to solve the above CPM problem. The training phase involves the agent interacting with a simulated MAC layer environment, which models the behavior of 5G NR-U and Wi-Fi transmitters under realistic coexistence conditions. At each time step, the agent observes the current augmented state, selects an action which is adjusting contention window sizes, and observes performance functions. The constraint is carefully designed to penalize violations of QoS requirements. Experience replay is used to store and sample past transitions, and a target network is updated periodically to stabilize learning. Training is conducted across varying number of transmitters to ensure the generalization.

In the execution phase, the trained policy is deployed in the simulation environment to evaluate its effectiveness. The augmented state representation ensures that the agent adapts dynamically to changing network conditions, making real-time adjustments to contention window sizes to optimize system performance. The QaSAL algorithm effectively balances competing objectives, ensuring that high-priority traffic meets its delay requirements while promoting fairness between 5G NR-U and Wi-Fi networks. The hyperparameters of QaSAL algorithm are summarized in Table I. The step duration is selected to be large enough to include several transmission attempts to enable the accurate calculation of the transmission delay.

TABLE I
HYPER-PARAMETERS OF QASAL ALGORITHM

| Parameter | Value |
|---|---|
| Interaction time $T$ | 20 s |
| Step duration | 2.5 ms |
| Discount factor | 0.99 |
| Replay buffer size | 100,000 |
| Range of $\epsilon$ | 1 to 0.1 |
| DQN learning rate | $10^{-4}$ |
| Batch size | 16 |
| Hidden layers | $32 \times 32 \times 32$ |
| $\eta_\lambda, \lambda_{\max}, T_0$ in (10) | 0.1, 10, 5 |

## VI. SIMULATION RESULTS

In this section, we highlight the performance of the proposed QaSAL algorithm. For the simulation scenario, we consider one gNB PC1 transmitter sharing the channel with varying number of AP PC3 transmitters, and evaluate the performance of primal-dual method (Section III) and QaSAL algorithm (Section IV). The simulations were conducted to analyze the algorithm's ability to balance fairness and delay metrics across different numbers of Wi-Fi's low-priority transmitters.

Figs. 3 and 4 illustrate the transmission delay dynamics for gNB PC1 traffic coexisting with 25 AP PC3 transmitters during both the training and execution phases of the primal-dual and QaSAL algorithms, respectively. While the

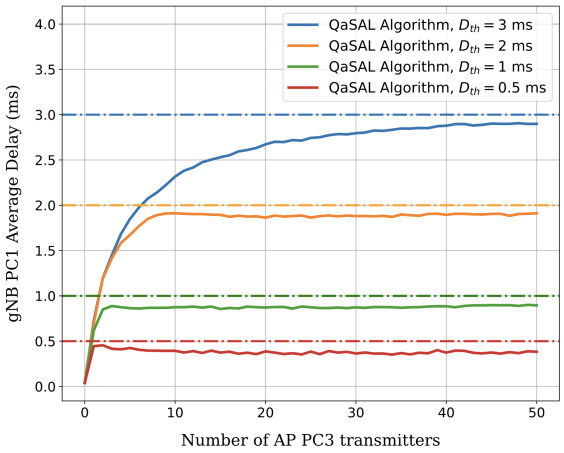


Fig. 6. Average Transmission delay of PC1 transmitter with QaSAL algorithm for various delay thresholds and varying number of PC3 transmitters.

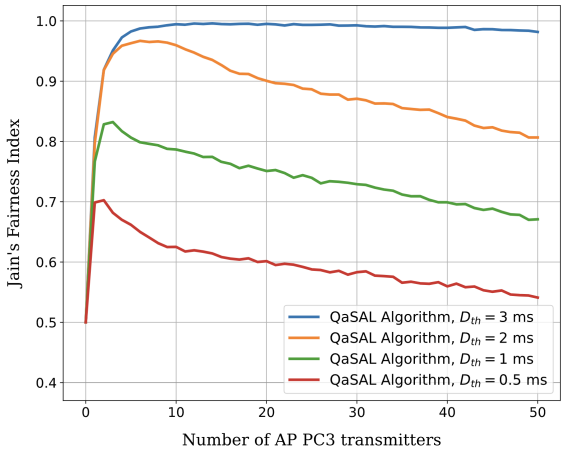


Fig. 7. JFI among Wi-Fi and NR-U networks with QaSAL algorithm for various delay thresholds and varying number of PC3 transmitters.

policy derived from the primal-dual algorithm can solve problem (11) on average, it does not guarantee near-optimality at a specific episode $k$. In contrast, the QaSAL algorithm effectively enforces the constraint by incorporating dual variables into the state space, leading to improved constraint satisfaction. Fig. 5 illustrates the evolution of the dual variable associated with the delay constraint of gNB PC1, as defined in (11b), during the execution phase. The dual variable adjusts dynamically in response to constraint violations, ensuring that the delay remains within the specified threshold.

Figs. 6 and 7 illustrate the average transmission delay for gNB PC1 traffic and JFI across both networks as the number of AP PC3 transmitters varies. The results demonstrate that the QaSAL algorithm effectively maintains the QoS requirements for high-priority traffic across different delay thresholds and varying number of transmitters while simultaneously optimizing fairness between the two networks.

## VII. Conclusions

In this paper, we introduced the QoS-aware State-Augmented Learnable (QaSAL) algorithm, a reinforcement learning-based approach designed to optimize the coexistence of 5G NR-U and Wi-Fi networks in unlicensed spectrum environments. By augmenting the network state with dual variables, our framework enables dynamic adaptation to QoS constraints, ensuring efficient spectrum sharing while maintaining low-latency transmission for high-priority traffic. Our results demonstrate that the QaSAL algorithm effectively balances network fairness with explicit delay constraints, outperforming traditional primal-dual methods in achieving constraint satisfaction. Unlike conventional approaches that require extensive parameter tuning, QaSAL learns optimal policies dynamically, improving system adaptability under varying traffic loads.

Future work will focus on extending the QaSAL framework to multi-channel coexistence scenarios. Additionally, further improvements in training efficiency and multi-agent learning could enhance the scalability of QaSAL for next-generation wireless networks.

## 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 Communications Society*, 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," *GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Taipei, Taiwan*, pp. 1–6, 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," *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Singapore*, pp. 191–199, 2013.

[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] Y. Uslu, R. Doostnejad, A. Ribeiro, and N. NaderiAlizadeh, "Learning to Slice Wi-Fi Networks: A State-Augmented Primal-Dual Approach," *arXiv preprint arXiv:2405.05748*, 2025.

[11] M. R. Fasihi and B. L. Mark, "Traffic Priority-Aware 5G NR-U/Wi-Fi Coexistence with Deep Reinforcement Learning," *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall), Washington, DC, USA*, pp. 1–6, 2024.

[12] 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.

[13] T. K. Le, U. Salim, and F. Kaltenberger, "An Overview of Physical Layer Design for Ultra-Reliable Low-Latency Communications in 3GPP Releases 15, 16, and 17," *IEEE Access*, vol. 9, pp. 433–444, 2021.